
Imperial College London
Faculty of Natural Sciences
Department of Mathematics

Bayesian Uncertainty Quantification for PDE Models: New Perspectives on Inference and Dimension Reduction

Yanni Papandreou

November, 2023

Supervised by Dr Andrew B. Duncan & Dr Jon Cockayne

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in
Statistics of Imperial College London and the Diploma of Imperial College London

Abstract

Mathematical models based on partial differential equations (PDEs) play a central role in science and engineering. This thesis aims to present some new perspectives on two different topics regarding such models.

The first issue is concerned with how mismatch between the model and the real system can often manifest itself. One way of dealing with such mismatch is to introduce stochasticity to model any uncertainty present. Concurrently, the increasing prevalence of instrumentation within modern engineering systems has driven the need for tools to combine sensor data with model predictions to help mitigate any model misspecification. This has led to an increased interest in uncertainty quantification (UQ). In the first part of this thesis we provide a review of the statistical finite element method (StatFEM) which is an emerging probabilistic method that allows observations of a physical system to be synthesised with the numerical solution of a PDE. We present a new theoretical analysis of StatFEM demonstrating that it has similar convergence properties to the finite element method on which it is based. Numerical examples are presented to demonstrate our theory, including an example which tests the robustness of StatFEM when extended to *nonlinear* quantities of interest.

The second issue is concerned with the fact that PDE models often involve a large number of parameters increasing the computational burden. Thus, the study of methods to perform dimension reduction are of vital importance. In the second part of this thesis we motivate the idea of localised active subspaces for dimension reduction, where the dimension reducing subspace varies smoothly as a function of the input parameters. To estimate localised active subspaces we then present a fully Bayesian model based on Gaussian Processes, called GrassGP, which can be used to interpolate a field of subspace-valued response-variables. Numerical experiments are presented to demonstrate this.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Yanni Papandreou

November, 2023

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

Firstly, I would like to thank my main supervisor, Dr Andrew Duncan, for giving me the opportunity to pursue this research, and for providing me with valuable advice, guidance, and support over the last 4 years. His subject knowledge, expertise, and enthusiasm for this research have been both inspirational and crucial to the completion of this thesis.

Secondly, I would like to thank my co-supervisor, Dr Jon Cockayne, for his abundant and generous support and for always being there with friendly advice. Our supervision meetings have always felt more collaborative rather than supervisory and I really appreciate all the times you helped refrain me from continuing to hit my head on the wall - I do not think I could have finished this thesis without all your support!

I would also like to thank my thesis examiners, Dr Nikolas Kantas, and Professor Simon Cotter, for their thorough feedback and discussion. I am also thankful for the financial support received from the Department of Mathematics at Imperial College in the form of a Roth Scholarship. Without this funding this PhD would simply not have been possible.

A special thanks goes to the friends I have met over the years at Imperial College. They helped carry me through the many highs and lows of the PhD. I am especially thankful to my office mate, Francesco Sanna Passino, who was always there with useful advice over many lunches and coffee breaks. I also want to thank my academic “brother” Enrico Crovini who has become one of my closest friends - here is to many more years of friendship! A big thanks also goes to all of my friends and support system outside academia, with a special mention to Zoe, Elias, Francesco, Shura, and Manos, for all their patience and support throughout this long journey. I am so grateful to have you all in my life.

To my family, with a special mention to my Godparents and my sister, Eleni: thank you for always looking out for me and for providing me with love and support, especially during the difficult moments of the PhD. Mom and Dad, thanks for always believing in me and for all the sacrifices you guys made to allow me to follow this path.

Finally, to my partner, Meli. I cannot imagine getting through this PhD without all your love, endless patience, and loyal support. You have been my anchor throughout this journey, and I am eternally grateful to you.

Table of Contents

Abstract	3
Statement of Originality	4
Copyright	5
Acknowledgements	6
Table of Contents	7
List of Figures	11
List of Tables	14
1 Introduction	15
1.1 Theoretical Guarantees for the Statistical Finite Element Method	15
1.2 Grassmann-valued Gaussian Processes for Probabilistic Localised Active Subspaces	17
1.3 Contributions and Structure of the Thesis	18
2 Background for StatFEM	21
2.1 Functional Analysis	21
2.1.1 Basics of Functional Analysis	21
2.1.2 Sobolev Spaces	24
2.2 Partial Differential Equations and the Finite Element Method	27
2.2.1 Partial Differential Equations	27
2.2.2 The Finite Element Method	33
2.3 Probability Theory and Inference	36
2.3.1 Basics of Probability Theory	36
2.3.2 The Multivariate Normal Distribution	37
2.3.3 Bayesian Inference	38
2.3.4 Gaussian Measures on Hilbert Spaces	40

2.3.5	Gaussian Processes	42
2.3.6	Divergences on Spaces of Probability Measures	46
3	Theoretical Guarantees for the Statistical Finite Element Method	51
3.1	Introduction	51
3.2	Related Work	51
3.3	The Statistical Finite Element Method	53
3.3.1	The StatFEM Model	53
3.3.2	Posterior Updates of the StatFEM approximation	56
3.4	Discretisation Error Analysis	59
3.5	Extension to time-dependent problems	68
3.5.1	StatFEM for time-dependent problems	68
3.5.2	Discretisation Error Analysis	75
3.5.3	Future work for the time-dependent case	76
3.6	Discussion	76
4	StatFEM Error Analysis Numerical Experiments	78
4.1	Introduction	78
4.2	One-dimensional Poisson Equation	79
4.2.1	Prior results for 1-D example	79
4.2.2	Posterior results for 1-D example	80
4.3	Two-dimensional Poisson Equation	82
4.3.1	Prior results for 2-D example	82
4.3.2	Posterior results for 2-D example	84
4.4	Distribution of the maximum	85
4.4.1	Prior results for maximum example	86
4.4.2	Posterior results for maximum example	86
4.5	Discussion	87
5	Background for Probabilistic Modelling of Localised Active Subspaces	88
5.1	The Matrix Normal Distribution	88
5.2	Matrix-Valued Gaussian Processes	90
5.3	Hamiltonian Monte Carlo	93
5.4	Differential Geometry and Manifolds	99
5.4.1	Basics of Differential Geometry	99
5.4.2	Statistics on Riemannian Manifolds	105
5.4.3	The Stiefel and Grassmann Manifolds	108

6 Grassmann-valued Gaussian Processes for Probabilistic Localised Active Sub-	114
spaces	
6.1 Introduction	114
6.2 Related Work	115
6.3 Ridge Approximation and Active Subspaces	116
6.4 Localised Active Subspaces	122
6.5 Grassmann Valued Gaussian Processes	124
6.5.1 Problem Formulation	125
6.5.2 Lifting of the Dataset to the Tangent Space	125
6.5.3 Gaussian Process Model on the Tangent Space	126
6.5.4 GrassGP Model	127
6.6 Inference using GrassGP model	128
6.6.1 Choice of Likelihood	128
6.6.2 The Full Bayesian Model	129
6.6.3 Prediction	130
6.7 Numerical Experiments	132
6.7.1 Visualisation of GrassGP Model	132
6.7.2 Localised active subspaces for a parametrised ODE	144
6.7.3 Localised active subspaces for a parametrised PDE	149
6.8 Discussion	153
7 Conclusion and Future Work	155
7.1 Contributions	155
7.2 Potential Future Research Directions	156
7.3 Closing Remarks	157
Bibliography	158
List of Notation	173
A Additional Background Material	177
A.1 Analysis	177
A.1.1 Basics of Functional Analysis	177
A.1.2 Fractional Sobolev Spaces	179
A.2 PDEs and FEM	181
A.2.1 Additional Concepts from PDE Theory	181
A.2.2 Additional Concepts from FEM Theory	183

A.3	Probability Theory	184
A.3.1	Additional Background Material	184
A.3.2	Additional Concepts from Probability Theory and Inference	193
A.4	Differential Geometry and Manifolds	196
A.4.1	Additional Background Material	196
B	Proofs and Additional Details	199
B.1	Additional Proofs for Chapter 2	199
B.2	Additional Proof Details for Chapter 3	200
B.3	Additional Proofs for Chapter 5	207
C	Additional Implementation Details for Experiments	209
C.1	Additional Details for Experiments in Chapter 4	209

List of Figures

2.1	Illustration of the FEM mesh and nodal basis functions.	34
2.2	Coupling between two distributions.	48
4.1	1-D prior and posterior results. Fig. 4.1a shows a log-log plot of Wasserstein distance against h for the priors. The slope of the fitted line is estimated to be $p = 1.99$, in agreement visually with the reference line. Fig. 4.1b shows a log-log plot of Wasserstein distance against h for the posteriors for varying levels of sensor noise. In both sub-figures, the reference line has a slope of 2. Notice how the best fit lines are all parallel to the reference line.	81
4.2	2-D prior results. Figure 4.2a shows a plot of $\text{LR}(h)$ against h . Figure 4.2b show a plot of smoothed $\text{LR}(h)$ for $h \leq 0.15$. The horizontal axis indicates the smallest h value taken in the cumulative averages.	83
4.3	2-D posterior results. Figure 4.3a shows a plot of $\text{LR}(h)$ against h . Figure 4.3b show a plot of smoothed $\text{LR}(h)$ for $h \leq 0.15$. The horizontal axis indicates the smallest h value taken in the cumulative averages.	85
4.4	1-D prior and posterior results for the maximum. Figure 4.4a shows a log-log plot of Wasserstein distance against h for the priors for the maximum. The reference line has a slope of 0.415. The slope of the fitted line is estimated to be $p = 0.415$, in agreement visually with the reference line. Figure 4.4b shows a log-log plot of Wasserstein distance against h for the posteriors for the maximum for different levels of sensor noise.	86
5.1	Visualization of the tangent space $T_p\mathcal{M}$ of a manifold (shown here as a curved surface) at a point $p \in \mathcal{M}$. Note that the tangent vector $v = \dot{c}(0) \in T_p\mathcal{M}$ is the velocity vector of a curve on the manifold $c : I \rightarrow \mathcal{M}$. This image is taken from [165].	101

6.1	Illustration of equivalence of $\text{Gr}(2, 1)$ with the unit circle in the plane S^1 . The red ray denotes an element, $\mathcal{X}$, of $\text{Gr}(2, 1)$, i.e. a 1 dimensional subspace of the plane. This subspace can be uniquely identified with two antipodal points on the unit circle S^1 . By treating these antipodal points as equivalent we get a unique representation of $\mathcal{X}$ via an angle $\theta \in [0, \pi)$	133
6.2	Plots of train and test datasets in various representations. <i>Top row</i> : plots of the components of $P(s)$. <i>Bottom row</i> : (<i>left</i>) plot of $\alpha(s)$, where $\alpha(s)$ is the angle the subspace $P(s)$ makes with positive x -axis, (<i>right</i>) plot of datasets on the unit circle representation of $\text{Gr}(2, 1)$ together with the anchor point.	134
6.3	Plot of θ vs $S(\theta)$ showing the location of the minimum, and hence the location of the anchor point, in red.	137
6.4	Plot of observed points with the Karcher mean on the circle representation of $\text{Gr}(2, 1)$	137
6.5	Diagnostic plots for HMC sampler (Section 6.7.1). <i>Left column</i> : Traceplots for the sampled hyper-parameters. <i>Right column</i> : Autocorrelation plots for the sampled hyper-parameters.	139
6.6	Plot of s vs $\alpha(s)$ showing the in-sample results from the HMC sampler (Section 6.7.1). The training points are plotted in red, while the test data is plotted as the dashed black line.	140
6.7	Predictions for tangent vectors $\mathbf{U}(s)$ for the dataset Eq. (6.59). The solid red curves show the averaged mean prediction, i.e. the average of the mean GP predictions for each sampled posterior hyper-parameter. The shaded blue region gives a 95% credible region over the GP predictions for each sampled posterior hyper-parameter.	142
6.8	Plot of s vs $\alpha(s)$ showing the resulting predictions for the test dataset Eq. (6.59). The training points are plotted in red, while the test data is plotted as the dashed black line.	143
6.9	Plot of errors against s using both the GP predictive mean and predictions (Section 6.7.1). The shaded regions depict a measure of uncertainty given by the error plus 1 standard deviation computed via Eq. (6.69) for both the predictive mean and predictions respectively. The vertical dashed green lines indicate the locations of the training data.	143
6.10	Plot illustrating the evolving localised active subspace by plotting the Riemannian Grassmann distance between $P(x_0^*)$ and the train and test datasets Eqs. (6.75) and (6.76) respectively against x	145

6.11	Diagnostic plots for HMC sampler (Section 6.7.2). <i>Left column:</i> Traceplots for the sampled hyper-parameters. <i>Right column:</i> Autocorrelation plots for the sampled hyper-parameters.	148
6.12	Plot illustrating the train and test locations for the numerical experiment in Section 6.7.3.	150
6.13	Plot illustrating the evolving localised active subspace by plotting the Riemannian Grassmann distance between $P(\mathbf{x}_0^*)$ and the test dataset.	151
6.14	Diagnostic plots for HMC sampler (Section 6.7.3). <i>Left column:</i> Traceplots for the sampled hyper-parameters. <i>Right column:</i> Autocorrelation plots for the sampled hyper-parameters.	154

List of Tables

4.1	Estimate of slopes and intercepts for the different levels of sensor noise ϵ . All of the slopes are around 2.	82
4.2	Estimate of slopes and intercepts for the different levels of sensor noise ϵ	87
6.1	Statistics for in-sample errors (Section 6.7.1).	140
6.2	Statistics for predictive errors (Section 6.7.1).	141
6.3	Statistics for in-sample errors (Section 6.7.2).	147
6.4	Statistics for predictive errors (Section 6.7.2).	149
6.5	Statistics for in-sample errors (Section 6.7.3).	152
6.6	Statistics for predictive errors (Section 6.7.3).	152

Chapter 1

Introduction

The work contained in this thesis is separated into two different, albeit related, projects. We now introduce these two projects separately.

1.1 Theoretical Guarantees for the Statistical Finite Element Method

Mathematical models based on partial differential equations (PDEs) play a central role in science and engineering. Since PDEs do not generally yield closed form solutions, approximate solutions are typically obtained via numerical methods. The Finite Element Method (FEM) [145] is one such scheme which approximates the solution by solving a discretised problem over a finite mesh of the PDE domain. FEM is used for a variety of classes of PDEs, however it is most classically and commonly applied to elliptic and parabolic PDEs. FEM-based methods have been widely studied, and as such, rigorous estimates of the error introduced through discretisation have been obtained for various classes of PDEs. In particular, these results characterise the convergence of the error to zero as a function of the size of the mesh.

Mismatch between the model and the real system can occur for a variety of reasons, for example, if the geometry of the domain is not fully resolved, if underlying processes are not completely understood, or if approximations must be introduced to reduce the model to a closed system of PDEs. Stochasticity is often introduced in models to reflect the uncertainty arising from this lack of knowledge. Model mismatch is not the only reason one introduces stochasticity into models. In particular, stochasticity can be introduced for a myriad of reasons, such as capturing intrinsically noisy processes, handling incomplete knowledge of model parameters, modelling errors made in measurements, and exploring sensitivity and robustness of a model to perturbations in the inputs. In the context of PDEs, stochasticity may appear in various forms, in the initial and/or boundary

conditions, in the forcing terms, in the coefficients of the PDE or even in the geometry of the domain on which the PDE is defined. The resulting governing equations, known as Stochastic Partial Differential Equations (SPDEs) are solved by probability distributions over the space of admissible solutions to the underlying deterministic PDE. The development and analysis of numerical methods for SPDEs is a very active area of research, see [164] for a review.

The increasing prevalence of instrumentation within modern engineering systems has driven the need for tools able to combine sensor data with predictions from PDE models in a manner which can mitigate any model misspecification. This technology forms an integral part of the modern notion of a digital twin [111], as a means of providing effective continuous monitoring of complex engineering assets. In particular, PDEs are often used to model the physical systems which digital twins aim to virtually represent, and it is thus vital to be able to synthesis observational data with predictions made from PDE based models.

The recently proposed Statistical Finite Element Method (StatFEM) framework [65] seeks to address this need through a statistical formulation of FEM, which permits numerical approximations of a linear PDE to be combined with sensor observations in a principled manner through the Bayesian formalism. Due to its general applicability and ability to provide quantification of uncertainty without the necessity of expensive Monte Carlo simulations, it has established itself as an effective means of assimilating data into the large-scale PDEs which arise in civil, marine and aerospace engineering. For example, see [56] for an application to digital railway monitoring, [51] for an application to shallow water equations, and [3] for a fusion of StatFEM with Langevin samplers.

StatFEM is a method which extends FEM enabling an approximation of the solution of SPDEs. Any numerical method which is used to approximate solutions to SPDEs will have approximation errors manifested both in the mean prediction and the quantified uncertainty. The novel contributions of this thesis in relation to StatFEM are two fold:

- We study how errors in the underlying FEM scheme propagate to errors in the approximate predictive distribution obtained by StatFEM. In particular, we derive error estimates for a StatFEM approximate solution of a stochastically-forced linear elliptic PDE. This is a critical gap in the literature that the work of this thesis will seek to address.
- We then apply this error result to obtain error bounds on a posterior predictive distribution which incorporates observational data.

This error analysis is an example of understanding how errors in the prior propagate to errors in the posterior. Standard approaches to such stability analysis [142] based on metrics such as Kullback-Leibler (KL) divergence, Hellinger, and Total Variation (TV) distances are not applicable

in the context of StatFEM since the underlying priors are mutually singular (as will be justified in Chapter 3). This leads us to instead consider the Wasserstein-2 distance as a metric to compare both the priors and posteriors. After analysing the case of spatial elliptic PDEs we then present some work aiming to analyse StatFEM for spatio-temporal problems.

1.2 Grassmann-valued Gaussian Processes for Probabilistic Localised Active Subspaces

Of equally vital importance to the modern day fields of data analysis and scientific computing is the task of dimension reduction. Simulations and mathematical models are increasingly important tools in the study of complex physical processes in many scientific and engineering disciplines. Unfortunately, complex mathematical models can often be expensive to simulate, and practitioners often turn to fitting response surfaces as surrogates for such models. In this context, a response surface is a simpler and cheaper model which aims to approximate the true complex model. Fitting such a response surface often entails tuning several parameters of these simpler models. Examples of such surrogate models include linear regression, random forests, support vector machines [70], and Gaussian process (GP) regression [125] (often known as kriging). As noted in [36], when the models involve a large number of parameters the cost of fitting accurate response surfaces via standard methods increases exponentially, a phenomenon infamously referred to as *the curse of dimensionality*. For example, in the context of GP regression a large number of input parameters is typically associated with requiring more data points (larger N) to achieve suitable accuracy; for naive implementations of GP regression this leads to an exponential growth in the computation time, $\mathcal{O}(N^3)$. This renders the construction of response surfaces intractable for complex high dimensional models. This naturally leads to the study of methods to perform dimension reduction for such models.

In particular, many statistical/machine learning tasks boil down to approximating some unknown function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ based on exact or noisy evaluations. When f represents a large computer model such a task is often referred to as *emulation* [132]. In many applications the dimension d of the domain is very large and many popular approaches for estimating f , such as GP regression, suffer from the aforementioned curse of dimensionality. However, suppose that the function f is such that not all of the d input variables are significant, and only a subspace $\mathcal{M}$ of the input space, $\mathbb{R}^d$, matters. Learning such a subspace $\mathcal{M}$ allows us to perform *model order reduction*. This has many practical uses, for instance if the true function f is very expensive to compute, knowledge of such a dimension-reducing subspace can enable us to cheaply emulate it by approximating it with a lower-dimensional version. This idea originates in the study of *Active*

subspaces (*AS*) [33], where the active subspace can be thought of as the collection of directions along which f is most sensitive to perturbations (on average). This concept will be expanded on in more detail in Section 6.3.

In the original formulation of active subspaces one makes the assumption that there exists a **single global** dimension reducing subspace which is assumed to be valid throughout the entire parameter space and any emulators based on AS approximations are constructed within this single subspace. In practice, such an assumption does not always hold, especially when dealing with highly complex models which can exhibit diverse behaviour across different phase transitions. This observation naturally motivates one to instead consider finding **different** low dimensional subspaces which are only valid locally in some sub-region of the original input space. We call such spaces *localised active subspaces*, and this will be expanded on in more detail in Section 6.4. In this thesis we investigate this problem through the lens of differential geometry, by viewing the active subspaces as elements of a particular Riemannian manifold, the Grassmann manifold. In particular, taking motivation from the more general field of research which deals with the statistical analysis of data lying on a Riemannian manifold, we are lead to consider the formal problem of learning a subspace-valued mapping $P : \mathbb{R}^d \rightarrow \text{Gr}(d, n)$. Here $\text{Gr}(d, n)$ is the Grassmann manifold of all n -dimensional (linear) subspaces of $\mathbb{R}^d$. This leads us to the second main contribution of this thesis: we will investigate a probabilistic model based on GPs which can be used to interpolate a field of Grassmann-valued response variables, i.e. given a set of N spatially indexed points in $\text{Gr}(d, n)$, $\{(\mathbf{s}_i, P(\mathbf{s}_i))\}_{i=1}^N$, we will investigate a probabilistic model capable of making predictions for $P(\tilde{\mathbf{s}})$ at unobserved test points $\tilde{\mathbf{s}} \in \mathbb{R}^d$. In particular, we will apply this model to illustrate how one can consider localised active subspaces.

1.3 Contributions and Structure of the Thesis

For clarity we now outline the novel contributions of this thesis. These are separated according to the two different projects introduced above:

Theoretical Guarantees for the Statistical Finite Element Method

- We provide a detailed review of the recent StatFEM model, deriving it from first principles.
- We study the bias introduced by discretisation error in the StatFEM model, for the case of spatial elliptic boundary value problems (BVPs), and derive rigorous upper bounds for the bias introduced in the prior and posterior in terms of the properties of the underlying finite element mesh.

- We conduct a thorough simulation study confirming our theoretical results on a number of illustrative FEM test problems.
- We empirically explore the limits of our theory by studying the bias induced by StatFEM in a scenario where the assumptions underpinning our theoretical results are no longer valid.
- We present our work towards deriving discretisation error bounds for spatio-temporal problems.

The theoretical contributions for this project are presented in Chapter 3, while the numerical experiments are presented in Chapter 4. The necessary technical background for this project is contained in Chapter 2.

Grassmann-valued Gaussian Processes for Probabilistic Localised Active Subspaces

- We provide a review of the concepts of optimal ridge approximation and active subspaces. We then motivate the idea of localised active subspaces where the dimension reducing subspace smoothly varies as a function of the input parameters.
- We present a fully Bayesian probabilistic model based on GPs, called GrassGP, which can be used to interpolate a field of Grassmann-valued response variables.
- We present a simple, low dimensional, numerical example allowing us to visualise both the GrassGP model and its performance.
- We present two higher dimensional numerical examples based on a parameterised ordinary differential equation (ODE) and PDE which allows us to demonstrate the use of our GrassGP model to estimate localised active subspaces at novel locations.

The contributions for this project are presented in Chapter 6, while the necessary technical background for this work is contained in Chapter 5.

Finally, Chapter 7 summarises the contributions and outlines potential future research avenues arising from the work contained in this thesis. We also include additional supplementary material for both projects in Appendices A to C.

Remark 1.1. The novel contributions in this thesis have resulted from collaborative work with various co-authors, including my supervisors and Professor Mark Girolami. In particular, the bulk of the StatFEM project has been accepted for publication in the SIAM/ASA Journal on Uncertainty Quantification under the title “Theoretical Guarantees for the Statistical Finite Element Method”. A preprint of this paper can be found online on [arXiv](https://arxiv.org/abs/2111.07691)¹. The dimension reduction

¹<https://arxiv.org/abs/2111.07691>

project is collaborative work with my supervisors, and is currently being prepared to be made into another paper. ♦

Chapter 2

Background for StatFEM

In this chapter we will introduce relevant and required background knowledge for the first project this thesis will present.

The sets of integers, natural numbers, and real numbers will be denoted by $\mathbb{Z}, \mathbb{N}, \mathbb{R}$, respectively. We follow the convention that the natural numbers exclude 0. The set of non-negative integers will be denoted by $\mathbb{N}_0$. Unless otherwise stated, we will use boldfaced lower-case Latin letters, e.g. $\mathbf{x}$, to denote vectors in $\mathbb{R}^d$, and boldfaced upper-case Latin letters, e.g. $\mathbf{M}$, to denote matrices of arbitrary, but finite, dimension. The symbol $\otimes$ will denote the outer product for vectors and the Kroenecker product for matrices.

2.1 Functional Analysis

2.1.1 Basics of Functional Analysis

Given a normed space $\mathcal{X}$ denote the associated norm (see Definition A.2) by $\|\cdot\|_{\mathcal{X}}$. Given two normed spaces $\mathcal{X}, \mathcal{Y}$ we denote by $L(\mathcal{X}, \mathcal{Y})$ the set of all bounded linear operators (see Definition A.3) from $\mathcal{X}$ to $\mathcal{Y}$. We denote the dual of $\mathcal{X}$ by $\mathcal{X}' = L(\mathcal{X}, \mathbb{R})$. When $\mathcal{X}$ is a Hilbert space (see Definition A.6) we denote the associated inner product (see Definition A.4) by $\langle \cdot, \cdot \rangle_{\mathcal{X}}$. We will sometimes drop the subscript for the norm and inner product when there is no risk of confusion over which spaces we are referring to.

Let $\mathcal{X}, \mathcal{Y}$ be Hilbert spaces. Given an operator $A \in L(\mathcal{X}, \mathcal{Y})$ we denote the adjoint of A by $A^\dagger \in L(\mathcal{Y}, \mathcal{X})$. Given $A \in L(\mathcal{X}, \mathcal{X})$, we say that A is symmetric if $A = A^\dagger$ and positive definite if it is symmetric and $\langle x, Ax \rangle > 0$ for all nonzero $x \in \mathcal{X}$. Similarly A is called positive semi-definite if it is symmetric and $\langle x, Ax \rangle \geq 0$ for all $x \in \mathcal{X}$. Recall that the trace of A can be defined as

$$\text{tr}(A) = \sum_{i=1}^{\infty} \langle Ae_i, e_i \rangle, \quad (2.1)$$

where $\{e_i\}_{i=1}^\infty$ is any orthonormal basis of $\mathcal{X}$. We will say that $A \in L(\mathcal{X}, \mathcal{X})$ is trace-class if $\text{tr}(A) < \infty$.

For a linear operator $A : \mathcal{X} \rightarrow \mathcal{Y}$ recall that we can define various norms. The operator norm is defined to be:

$$\|A\|_{\mathcal{X} \rightarrow \mathcal{Y}} = \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{\|Ax\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}}. \quad (2.2)$$

The trace norm is defined to be:

$$\|A\|_{\text{tr}} = \text{tr}([A^\dagger A]^{1/2}), \quad (2.3)$$

while the Hilbert-Schmidt norm is defined to be:

$$\|A\|_{\text{HS}} = [\text{tr}(A^\dagger A)]^{1/2}. \quad (2.4)$$

Recall the well known inequality [105] linking all three:

$$\|A\|_{\mathcal{X} \rightarrow \mathcal{Y}} \leq \|A\|_{\text{HS}} \leq \|A\|_{\text{tr}}. \quad (2.5)$$

For an operator $A : \mathcal{X} \rightarrow \mathcal{Y}$ between two vectors spaces $\mathcal{X}, \mathcal{Y}$ and for a subset $U \subset \mathcal{X}$ we will denote the image of U under A via the notation $A(U)$, i.e.

$$A(U) := \{y \in \mathcal{Y} \mid \exists u \in U \text{ such that } A(u) = y\}. \quad (2.6)$$

We will denote the *image* of A via $\text{im}(A)$, i.e. $\text{im}(A) = A(\mathcal{X})$, and we will denote the *kernel/null-space* of A via:

$$\ker(A) := \{x \in \mathcal{X} \mid A(x) = 0\}. \quad (2.7)$$

L^p spaces

Let $D \subset \mathbb{R}^d$ be a connected, bounded, and open domain with boundary ∂D and let $\bar{D} = D \cup \partial D$ be its closure. Denote by $L^p(D)$ the spaces of p -integrable functions on D , i.e. the following spaces:

$$L^p(D) := \{u : D \rightarrow \mathbb{R} \mid \|u\|_{L^p(D)} < \infty\}, \quad (2.8)$$

where $\|\cdot\|_{L^p(D)}$ is the norm on $L^p(D)$ defined via:

$$\|u\|_{L^p(D)} := \left(\int_D |u(\mathbf{x})|^p d\mathbf{x} \right)^{\frac{1}{p}}. \quad (2.9)$$

In the case $p = \infty$, we define,

$$\|u\|_{L^\infty(D)} := \operatorname{ess\,sup}_{\mathbf{x} \in D} |u(\mathbf{x})|, \quad (2.10)$$

and thus we can identify $L^\infty(D)$ as the space of all essentially bounded functions on D . Note that here $\operatorname{ess\,sup}$ denotes the essential supremum and so we can omit sets of measure zero.

Remark 2.1. Occasionally it is useful to consider a version of Eq. (2.8) with respect to a general measure, μ , on D , i.e. we will consider the following space, with associated norm:

$$L^p(D, \mu) := \{u : D \rightarrow \mathbb{R} \mid \|u\|_{L^p(D, \mu)} < \infty\}, \quad (2.11)$$

$$\|u\|_{L^p(D, \mu)} := \left(\int_D |u(\mathbf{x})|^p d\mu(\mathbf{x}) \right)^{\frac{1}{p}}. \quad (2.12)$$

◆

Note that in the case $p = 2$ we have that $L^2(D)$ is a *Hilbert space* with associated inner product:

$$\langle u, v \rangle_{L^2(D)} := \int_D u(\mathbf{x})v(\mathbf{x})d\mathbf{x}, \quad \forall u, v \in L^2(D). \quad (2.13)$$

We have, for $u \in L^2(D)$, the relation between the norm and inner product, $\|u\|_{L^2(D)}^2 = \langle u, u \rangle_{L^2(D)}$.

Spaces of continuous and differentiable functions

We denote by $C(D), C(\overline{D})$ the space of continuous functions on D and $\overline{D}$, respectively. Note that the sup-norm Eq. (2.10) also defines a norm on both $C(D), C(\overline{D})$ **provided** that D is bounded.

Remark 2.2. When D is bounded, we have that $\overline{D}$ is compact. Thus, any function $u \in C(\overline{D})$ is bounded and obtains its maximum at some point in $\overline{D}$ via the Extreme Value Theorem [130, Theorem 4.16]. Thus, one can replace the supremum in Eq. (2.10) with the maximum, i.e.

$$\|u\|_{L^\infty(D)} = \max_{\mathbf{x} \in \overline{D}} |u(\mathbf{x})|, \quad \forall u \in C(\overline{D}). \quad (2.14)$$

◆

In a similar manner $C^n(D)$ denotes the set of all continuous functions which have n continuous derivatives on D . We will denote by $C^\infty(D)$ the set of infinitely differentiable functions on D , i.e.

$$C^\infty(D) := \bigcap_{n=1}^{\infty} C^n(D). \quad (2.15)$$

The set of functions $u \in C^\infty(D)$ with compact support will be denoted by $C_c^\infty(D)$, i.e.

$$C_c^\infty(D) := \{u \in C^\infty(D) \mid \operatorname{supp}(u) \text{ is a compact subset of } D\}, \quad (2.16)$$

where here $\text{supp}(u)$ is the support of u and is defined as the closure of the set $\{\mathbf{x} \in D \mid u(\mathbf{x}) \neq 0\}$.

We will denote by $C_b^n(D)$ the subset of $C^n(D)$ of functions whose partial derivatives of order up to n are bounded and uniformly continuous on D . The spaces of Hölder continuous-functions, $C_b^{n,\lambda}(D)$, are important subspaces of $C_b^n(D)$ and can be defined (adopted from Definition 1.49 in [43]) as the space of functions $u \in C_b^n(D)$ such that there exists $C > 0$ for which the following holds:

$$|\partial^\alpha u(\mathbf{x}) - \partial^\alpha u(\mathbf{y})| \leq C \|\mathbf{x} - \mathbf{y}\|^\lambda, \quad \forall \alpha, |\alpha| = n, \forall \mathbf{x}, \mathbf{y} \in D, \quad (2.17)$$

where here $\alpha \in \mathbb{N}_0^d$ is a multi-index, $|\alpha| = \sum_{i=1}^d \alpha_i$, and the symbol ∂^α denotes the following partial derivative:

$$\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}. \quad (2.18)$$

Note that in the case $n = 0$, $\lambda = 1$ this is just the space of Lipschitz continuous functions. The spaces, $C_b^n(D)$, when equipped with the norm,

$$\|u\|_{C_b^n(D)} := \sup_{|\alpha| \leq n} \left(\sup_{\mathbf{x} \in D} |\partial^\alpha u(\mathbf{x})| \right), \quad (2.19)$$

can be shown [43, see Definition 1.47] to be Banach spaces (see Definition A.5) and we have the following continuous embeddings which will be useful when discussing the Sobolev Embedding Theorem (see Theorem 2.1):

$$C_b^{n,\lambda}(D) \hookrightarrow C_b^{n,\nu}(D) \hookrightarrow C_b^n(D), \quad \forall \nu, \lambda \text{ such that } 0 < \nu < \lambda < 1, \quad (2.20)$$

where the inclusions are strict.

Remark 2.3. When considering PDEs one typically starts out by interpreting the solutions as functions in the spaces $C^n(D)$ for appropriate values of $n \in \mathbb{N}$. This leads to a limited notion of solutions, referred to as classical solutions, to PDEs which motivates one to instead consider solutions belonging in *Sobolev spaces* where we interpret the derivatives in a distributional manner as discussed in the following section. ♦

2.1.2 Sobolev Spaces

Central to the analysis of PDEs is the theory of Sobolev spaces. The concept of *weak derivatives* is central to discussing such spaces. A weak derivative of a function $u : D \rightarrow \mathbb{R}$ can be defined as follows (adopted from Definition 1.42 in [101]):

Definition 2.1 (Weak derivative). Let $D \subset \mathbb{R}^d$ be open, and let $\alpha \in \mathbb{N}_0^d$ be a multi-index. Let $u : D \rightarrow \mathbb{R}$ be a given function. A function $v : D \rightarrow \mathbb{R}$ is called the (α) -weak-derivative of u if it

satisfies

$$\int_D \partial^\alpha \varphi(\mathbf{x}) u(\mathbf{x}) d\mathbf{x} = (-1)^{|\alpha|} \int_D \varphi(\mathbf{x}) v(\mathbf{x}) d\mathbf{x}, \quad \forall \varphi \in C_c^\infty(D). \quad (2.21)$$

We will denote the weak-derivative with the notation $v = \mathcal{D}^\alpha u$. $\blacklozenge$

We can now introduce Sobolev spaces which are defined as follows (adopted from Definition 1.44 in [101]):

Definition 2.2 (Sobolev spaces). For $p \in [1, \infty)$, $k \in \mathbb{N}$ the *Sobolev space*, $W^{k,p}(D)$, is the space of functions whose weak derivatives lie in $L^p(D)$ for all orders up to and including order k , i.e.

$$W^{k,p}(D) := \{u : D \rightarrow \mathbb{R} \mid \mathcal{D}^\alpha u \in L^p(D) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ such that } |\alpha| \leq k\}. \quad (2.22)$$

$\blacklozenge$

Sobolev spaces are Banach spaces with the following norms [101, see Proposition 1.45]:

Proposition 2.1. $W^{k,p}(D)$ is a Banach space with the norm

$$\|u\|_{W^{k,p}(D)} := \left(\sum_{0 \leq |\alpha| \leq k} \|\mathcal{D}^\alpha u\|_{L^p(D)}^p \right)^{\frac{1}{p}}. \quad (2.23)$$

In the special case $p = 2$ we use the notation $H^k(D)$ to denote the space $W^{k,2}(D)$. In this case $H^k(D)$ is a Hilbert space with the inner product

$$\langle u, v \rangle_{H^k(D)} := \sum_{0 \leq |\alpha| \leq k} \langle \mathcal{D}^\alpha u, \mathcal{D}^\alpha v \rangle_{L^2(D)}. \quad (2.24)$$

It is also convenient to define the Sobolev semi-norm:

$$|u|_{H^k(D)} := \left(\sum_{|\alpha|=k} \|\mathcal{D}^\alpha u\|_{L^2(D)}^2 \right)^{\frac{1}{2}}. \quad (2.25)$$

For brevity, we will denote by $\|\cdot\|_k$, $\langle \cdot, \cdot \rangle_k$, $|\cdot|_k$ the norm, inner product, and semi-norm, respectively, of $H^k(D)$. $\blacklozenge$

Remark 2.4. Some of the results in this thesis will require analysis of functions of continuously varying smoothness as opposed to those with an integral order of weak derivatives introduced thus far in Definition 2.2. This consideration leads one to consider the notion of *fractional Sobolev spaces*. A full discussion of such spaces is beyond the scope of this thesis. We instead include, for convenience, a brief discussion of fractional Sobolev spaces, including a definition, in Appendix A.1.2. For the main purposes of the this thesis it suffices to know that we can consider fractional Sobolev

spaces, $W^{s,p}(D)$, for $s \in (0, \infty)$, and for non-integer s one can think of such spaces as sitting somewhere in-between the integral Sobolev spaces as will be made more precise in Proposition 2.2 below. In what follows in this thesis when we state “Sobolev spaces” we refer to the integral order and fractional Sobolev spaces as appropriate based on the exponent. $\blacklozenge$

We now introduce an important assumption regarding the regularity of the domain D . This assumption will be useful for the error analysis of StatFEM in Chapter 3.

Assumption 2.1 (domain regularity). The dimension of the domain is restricted to $d \in \{1, 2, 3\}$. The domain D is convex polygonal, i.e. it is convex and its boundary ∂D is a polygon. Note that this implies that the domain is bounded. $\blacklozenge$

Under Assumption 2.1 on the domain D we have that the Sobolev spaces form a continuous hierarchy of functions of increasing smoothness as the following result adopted from Propositions 2.1, 2.2, and Corollary 2.3 in [45] shows:

Proposition 2.2. Let $p \in [1, \infty)$ and let $s, s' \in (0, \infty)$. Suppose D satisfies Assumption 2.1. Then, if $s' \geq s$, we have:

$$W^{s',p}(D) \subset W^{s,p}(D). \quad (2.26)$$

$\blacklozenge$

Remark 2.5. Note that Proposition 2.2 holds under more general assumptions on the domain D , as stated in the results in Section 2 of [45]. For the purposes of this thesis we will always be assuming that Assumption 2.1 holds and so we state the result for this case only. $\blacklozenge$

An important result in the theory of Sobolev spaces is the *Sobolev Embedding Theorem*. This result provides various embeddings of Sobolev spaces into L^q and Hölder spaces and is presented below. This result is adopted from Theorem 4.57 in [43]

Theorem 2.1 (Sobolev Embedding). Let D be a Lipschitz open set. The following statements hold:

- If $sp < d$, then $W^{s,p}(D) \hookrightarrow L^q(D)$ for every $q \leq dp/(d - sp)$.
- If $sp = d$, then $W^{s,p}(D) \hookrightarrow L^q(D)$ for every $q < \infty$.
- If $sp > d$, then we have:
 - If $s - d/p \notin \mathbb{N}$, then $W^{s,p}(D) \hookrightarrow C_b^{[s-d/p], s-d/p-[s-d/p]}(D)$.
 - If $s - d/p \in \mathbb{N}$, then $W^{s,p}(D) \hookrightarrow C_b^{s-d/p-1, \lambda}(D)$ for every $\lambda < 1$.

where $[\cdot]$ denotes the integral part of a real number. $\blacklozenge$

Note that the domains we consider satisfy Assumption 2.1 and so are Lipschitz open. A corollary of this theorem which will be useful in Theorem 3.2 is the following which states that provided $sp > d$ the Sobolev space $W^{s,p}(D)$ contains functions which are continuous:

Corollary 2.1. Let D be a Lipschitz open set. If $sp > d$, then $W^{s,p}(D)$ is contained in $C(D)$. ♦

Proof. We split this into the two cases $sp - d \notin \mathbb{N}$ and $sp - d \in \mathbb{N}$.

In the first case, we have via Theorem 2.1 that $W^{s,p}(D)$ is contained in $C_b^{[s-d/p], s-d/p-[s-d/p]}(D)$. Since $s - d/p \notin \mathbb{N}$ and $s - d/p > 0$ we have that $0 < s - d/p - [s - d/p] < 1$ and so by Eq. (2.20) we have $W^{s,p}(D)$ is contained in $C_b^{[s-d/p]}$. On noting that $s - d/p > 0$ we must have that $[s - d/p]$ is a non-negative integer and so we get that $W^{s,p}(D)$ is contained in $C(D)$ as required.

In the second case, we have again via Theorem 2.1 that $W^{s,p}(D)$ is contained in $C_b^{s-d/p-1, \lambda}$ for every $\lambda < 1$. In particular, we have via Eq. (2.20) that $W^{s,p}(D)$ is contained in $C_b^{s-d/p-1} \subset C(D)$ since $s - d/p - 1$ must be a non-negative integer. ■

For further details on Sobolev spaces the reader is directed to [93, 43], while for more details on fractional Sobolev spaces one can consult [45, 161, 43].

2.2 Partial Differential Equations and the Finite Element Method

In this section we will introduce and discuss the essential concepts from PDE theory required for this thesis in Section 2.2.1. Then in Section 2.2.2 we will review the necessary concepts from FEM theory which will be needed for this thesis.

2.2.1 Partial Differential Equations

In Chapter 3 we provide detailed error analysis for StatFEM, an emerging Bayesian statistical methodology to perform inference over numerical solutions of physical models based on PDEs. PDEs model a plethora of processes across several fields in science and engineering; as such providing an exhaustive list is difficult. For a canonical text on the theoretical analysis of PDEs the reader is directed to [55]. For a brief history of the development and analysis of PDEs the interested reader is directed to [20] where the authors Brezis and Browder note that PDEs are used to model a myriad of physical phenomena from “vibrating strings, elasticity, the Newtonian gravitational field of extended matter, electrostatics,...electricity and magnetism.” This emphasises their importance and leads to many computer models having PDEs at their core.

In order to gain understanding into the physical processes modelled by PDEs it is necessary to be able to solve the relevant equations. Only a small subset of PDEs admit closed-form solutions,

and instead one has to resort to numerical methods to approximate the solution. StatFEM is based on FEM which is a popular method for approximating solutions of PDEs. Before we can understand FEM it is important to pin down precisely what one means by a *solution* of a PDE. In the remainder of this section we follow the presentations given in [101, 55, 89]. There are three notions of solutions to PDEs: *classical*, *strong*, and *weak* solutions. We now present these below.

For concreteness we will consider the following linear elliptic BVP:

$$\mathcal{L}u(\mathbf{x}) := - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(\kappa_{ij}(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial x_j} \right) = f(\mathbf{x}), \quad \mathbf{x} \in D, \quad (2.27)$$

$$u(\mathbf{x}) = g(\mathbf{x}), \quad \mathbf{x} \in \partial D, \quad (2.28)$$

for some symmetric positive definite, spatially-dependent conductivity $\boldsymbol{\kappa} : D \rightarrow \mathbb{R}^{d \times d}$ with components $\{\kappa_{ij}\}_{i,j=1}^d$, forcing $f : D \rightarrow \mathbb{R}$, and boundary term $g : \partial D \rightarrow \mathbb{R}$. Note that $\boldsymbol{\kappa}$ is not always a conductivity, and can be used to represent permeability instead. For brevity we will also express the differential operator, $\mathcal{L}$, in Eq. (2.27) in vector calculus notation as follows:

$$\mathcal{L}u(\mathbf{x}) = -\nabla \cdot (\boldsymbol{\kappa}(\mathbf{x}) \nabla u(\mathbf{x})), \quad (2.29)$$

where ∇ is the *gradient operator*, i.e. for a given scalar field, $v : \mathbb{R}^d \rightarrow \mathbb{R}$, we have:

$$\nabla v(\mathbf{x}) := \begin{pmatrix} \frac{\partial v(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial v(\mathbf{x})}{\partial x_d} \end{pmatrix}, \quad (2.30)$$

and where for a given vector field, $\mathbf{v} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, the notation $\nabla \cdot \mathbf{v}$ denotes the *divergence* of the vector field, i.e.

$$(\nabla \cdot \mathbf{v})(\mathbf{x}) := \sum_{i=1}^d \frac{\partial v(\mathbf{x})}{\partial x_i}. \quad (2.31)$$

We assume that the domain D satisfies Assumption 2.1. This is assumed for simplicity; we note that it should be possible to relax some of these assumptions though this is not investigated in this thesis. *Note: For some of our results we will require that the dimension of the domain is further restricted to $d \in \{1, 2\}$. This will be explicitly stated when required.*

We will consider different further assumptions on the data for the three types of solutions.

Classical solutions

Perhaps the most familiar notion of solution is that of a *classical* solution where one interprets the derivatives in Eq. (2.27) in the classical sense and seeks a solution $u : D \rightarrow \mathbb{R}$ which satisfies

Eqs. (2.27) and (2.28) at every point $\mathbf{x} \in D$ as defined below:

Definition 2.3 (Classical solution). We call a function $u \in C^2(D) \cap C(\overline{D})$ satisfying Eqs. (2.27) and (2.28) for all $\mathbf{x} \in D$ a *classical solution* of the elliptic BVP from Eqs. (2.27) and (2.28). ♦

In order for this definition to make sense we require further assumptions on the data of the elliptic BVP. For simplicity, we will assume that the data (D, κ, f, g) satisfies Assumption 2.1 as well as the following assumption:

Assumption 2.2 (classical conductivity and forcing regularity). The conductivity κ is symmetric, i.e. $\kappa(\mathbf{x})^T = \kappa(\mathbf{x})$ for all $\mathbf{x} \in D$ and its components are continuous with at least one continuous derivative, i.e. $\kappa_{ij} \in C^1(\overline{D})$ for all $i, j \in \{1, \dots, d\}$. The forcing f is continuous on $\overline{D}$, i.e. $f \in C(\overline{D})$. ♦

Under Assumptions 2.1 and 2.2 one can show that there exists a unique classical solution to the elliptic BVP, which will be a smooth function with 2 continuous derivatives. Note that this notion of solution coincides with what is first typically studied when introducing PDEs as is discussed in Remark 2.3. However, in applications we often need to solve problems where the forcing term, f , is less smooth, for example we often consider $f \in L^2(D)$. The notion of a classical solution is too restrictive in this case and we instead seek a solution which solves Eq. (2.27) *almost-everywhere* with respect to the Lebesgue measure on D with the derivatives interpreted in the weak sense. This leads us to consider *strong* solutions; before we can present such solutions we must first introduce the concept of *trace spaces*.

Trace Spaces

In order to deal with the boundary condition Eq. (2.28) in the elliptic BVP we must introduce Sobolev spaces which incorporate boundary conditions. This leads us to consider the spaces $H_0^k(D)$ for $k \in \mathbb{N}$ which are defined (see Section 5.2 in [55]) to be the closure of $C_c^\infty(D)$ in $H^k(D)$. These spaces are interpreted as the space of functions $u \in H^k(D)$ such that,

$$\mathcal{D}^\alpha u = 0 \text{ almost everywhere on } \partial D \text{ for all } |\alpha| \leq k - 1. \quad (2.32)$$

In particular, the space $H_0^1(D)$, which can be interpreted as

$$H_0^1(D) := \{u \in H^1(D) \mid u(\mathbf{x}) = 0 \text{ for almost all } \mathbf{x} \in \partial D\}, \quad (2.33)$$

will be crucial for the notion of strong and weak solutions. This space of functions consists of those in $H^1(D)$ which are zero on the boundary ∂D . In order to incorporate non-zero g in Eq. (2.28) we

must consider the following space, $H_g^1(D)$, which is defined as follows (see (2.52) in [101]):

$$H_g^1(D) := \{w \in H^1(D) \mid \gamma w = g\}, \quad (2.34)$$

where $\gamma : H^1(D) \rightarrow L^2(\partial D)$ is the so-called *trace operator* which maps functions on D to square-integrable functions on the boundary ∂D . The trace operator has been studied extensively [55, 101] and for our purposes it suffices to know that such an operator exists and that its image is not the whole of $L^2(\partial D)$ but instead a certain subspace which is called the *trace space* (definition adopted from Definition 2.38 in [101]):

Definition 2.4 (*Trace space*). Let $D \subset \mathbb{R}^d$ be a bounded domain. The *trace space*, denoted by $H^{1/2}(\partial D)$ is defined as the image of the trace operator γ on $H^1(D)$ i.e.

$$H^{1/2}(\partial D) := \gamma(H^1(D)) = \{\gamma w \mid w \in H^1(D)\}. \quad (2.35)$$

◆

Since the trace-space is not all of $L^2(\partial D)$ we must have that $g \in H^{1/2}(\partial D)$ in Eq. (2.28). For the work in this thesis we will also require the dual space of $H_0^1(D)$ which we will denote by $H^{-1}(D)$.

Strong solutions

Having introduced weak derivatives and Sobolev spaces we are now in a position to introduce the concept of strong solutions of Eqs. (2.27) and (2.28).

Definition 2.5 (*Strong solution*). We call a function $u \in H^2(D) \cap H_g^1(D)$ satisfying the governing equation Eq. (2.27) for almost all $\mathbf{x} \in D$ with derivatives interpreted in the weak sense (see Definition 2.1) a *strong solution* of the elliptic BVP from Eqs. (2.27) and (2.28). ◆

In order for this definition to make sense we again require further assumptions on the data of the elliptic BVP. For simplicity, we will make the following alternative assumptions on the data of the elliptic BVP:

Assumption 2.3 (*strong conductivity and forcing regularity*). The conductivity κ is symmetric, i.e. $\kappa(\mathbf{x})^T = \kappa(\mathbf{x})$ for all $\mathbf{x} \in D$ and its components are almost surely bounded, i.e. $\kappa_{ij} \in L^\infty(D)$ for all $i, j \in \{1, \dots, d\}$. The forcing f is square integrable, i.e. $f \in L^2(D)$. ◆

Under Assumptions 2.1 and 2.3, and assuming that $g \in H^{1/2}(\partial D)$, one can show that there exists a unique strong solution, which will have 2 weak, square integrable, derivatives, and which satisfies Eq. (2.27) almost everywhere with respect to the Lebesgue measure on D .

Even this more relaxed setting of strong solutions is often too restrictive for applications where we might be interested in rougher forcing terms $f \notin L^2(D)$. For an example of this scenario the reader is directed to Example 2.6 in [101]. This leads us to consider a further generalisation of the meaning of a solution to Eqs. (2.27) and (2.28), the so called *weak* or *variational solutions* which we now present.

Weak solutions

We now consider rougher forcing terms $f \in H^{-1}(D)$. Instead of considering Eqs. (2.27) and (2.28) directly we will instead consider its *variational formulation* which informally can be obtained by post-multiplying Eq. (2.27) with a test function $\varphi \in C_c^\infty(D)$ and then using integration by parts to obtain:

$$\int_D \nabla \varphi(\mathbf{x}) \cdot (\boldsymbol{\kappa}(\mathbf{x}) \nabla u(\mathbf{x})) \, d\mathbf{x} = \int_D f(\mathbf{x}) \varphi(\mathbf{x}) \, d\mathbf{x}, \quad \forall \varphi \in C_c^\infty(D), \quad (2.36)$$

where the additional boundary term arising from the integration by parts vanishes due to φ being zero on ∂D . Note that by utilising integration by parts we are able to shift one derivative from u to the test function φ which allows us to consider less smooth functions u as solutions in the case where we have rougher forcing terms. Further, since by definition $C_c^\infty(D)$ is dense in $H_0^1(D)$ we can replace the $\varphi \in C_c^\infty(D)$ in Eq. (2.36) with $v \in H_0^1(D)$. Thus, any u which satisfies Eq. (2.27) also satisfies the variational formulation:

$$\int_D \nabla v(\mathbf{x}) \cdot (\boldsymbol{\kappa}(\mathbf{x}) \nabla u(\mathbf{x})) \, d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}, \quad \forall v \in H_0^1(D). \quad (2.37)$$

We can reformulate the variational formulation Eq. (2.37) as

$$a(u, v) = \langle f, v \rangle, \quad \forall v \in H_0^1(D), \quad (2.38)$$

where the bilinear form $a : H_0^1(D) \times H_0^1(D) \rightarrow \mathbb{R}$ is defined by

$$a(u, v) := \int_D \nabla v(\mathbf{x}) \cdot (\boldsymbol{\kappa}(\mathbf{x}) \nabla u(\mathbf{x})) \, d\mathbf{x}, \quad (2.39)$$

and where $\langle f, v \rangle$ denotes the integral on the right hand side of Eq. (2.37).

Remark 2.6. Technically $f \in H^{-1}(D)$ is not a function but instead is a bounded linear functional on $H_0^1(D)$. Thus, to be more precise $\langle f, v \rangle$ in Eq. (2.38) actually denotes the application of the linear functional f to $v \in H_0^1(D)$, i.e. $\langle \cdot, \cdot \rangle$ is the $H^{-1}(D)$, $H_0^1(D)$ duality pairing. For more details on making this precise the interested reader is directed to [55]. We will abuse notation in this thesis and write $f \in H^{-1}(D)$ as if it were a function and be careful to only interpret it inside integrals against test functions in $H_0^1(D)$ where this makes sense. ♦

We are now ready to define a weak solution as a function solving the variational formulation Eq. (2.37) of the elliptic BVP, i.e.

Definition 2.6 (Weak solution). We call a function $u \in H_g^1(D)$ that satisfies Eq. (2.38) a *weak solution* of the elliptic BVP from Eqs. (2.27) and (2.28). ♦

In order for this to be well-defined we now make the following assumption on the conductivity of the elliptic BVP:

Assumption 2.4 (conductivity regularity). The conductivity $\kappa \in L_\infty(D; \mathbb{R}^{d \times d})$ is symmetric and uniformly positive definite, i.e. there exists a constant $a_0 > 0$ such that:

$$\operatorname{ess\,inf}_{x \in D} \xi^T \kappa(x) \xi \geq a_0 |\xi|^2, \quad \forall \xi \in \mathbb{R}^d, \quad (2.40)$$

Furthermore, the components of κ are Lipschitz continuous on the closure $\overline{D}$, i.e. there exists $a_{\text{Lip}} > 0$ such that:

$$|\kappa_{ij}(\mathbf{x}) - \kappa_{ij}(\mathbf{y})| \leq a_{\text{Lip}} |\mathbf{x} - \mathbf{y}|, \quad \forall \mathbf{x}, \mathbf{y} \in \overline{D}, \forall i, j \in \{1, \dots, d\}. \quad (2.41)$$

Inequality Eq. (2.40) is often referred to as (*uniform*) *ellipticity* of the conductivity. ♦

Under Assumptions 2.1 and 2.4, and assuming $f \in H^{-1}(D)$, $g \in H^{1/2}(\partial D)$, it is possible to show existence and uniqueness of a weak solution of the elliptic BVP. In particular, we can exploit the (bi-)linearity of Eq. (2.38) and consider the following equivalent formulation of Eq. (2.38): find $u_0 \in H_0^1(D)$ such that,

$$a(u_0, v) = \langle f, v \rangle - a(u_g, v), \quad \forall v \in H_0^1(D), \quad (2.42)$$

where $u_g \in H_g^1(D)$ is some fixed function satisfying Eq. (2.28). Assumption 2.4 on κ ensures that the bilinear form a is symmetric, bounded and coercive (see Definitions A.9 and A.10) [55]. This, together with the boundedness of the map $v \mapsto \langle f, v \rangle$, allows one to use the Lax-Milgram Lemma (see Lemma A.1) to establish existence and uniqueness of a weak solution. For details on this the interested reader is directed to [55].

Setting for discussing the Finite Element Method

FEM seeks an approximation of the weak solution of a PDE. As such, we will be concerned exclusively with weak solutions in this thesis. For clarity, we now state precisely the assumptions we will adopt for the remainder of the thesis (unless otherwise stated). Noting the equivalence of Eq. (2.38) and Eq. (2.42) we will, for simplicity, only consider zero Dirichlet boundary conditions, i.e. we will take $g = 0$ in Eq. (2.28). We will also assume that Assumptions 2.1 and 2.4 hold unless

otherwise stated. Further, we will consider $f \in L^2(D)$ in what follows unless we explicitly state otherwise. Note that in some of the results in Section 3.4 we will consider random forcing terms which are rougher and are not in $L^2(D)$. This will be made explicit where appropriate.

Under Assumptions 2.1 and 2.4, we have the following standard regularity results. The first theorem is adapted from results in Section 6.2.2 of [55] and provides us with an existence and uniqueness result:

Theorem 2.2. *Under Assumptions 2.1 and 2.4 we have, for each $f \in L^2(D)$, that there exists a unique weak solution $u \in H_0^1(D)$ satisfying Eq. (2.37). ♦*

Further, under our assumptions we have the following stronger regularity result [68, Theorem 3.2.1.2]:

Theorem 2.3. *Under Assumptions 2.1 and 2.4, and assuming that $f \in L^2(D)$, we have that the unique weak solution, u , given by Theorem 2.2, lies in the Sobolev space $H^2(D)$, i.e. $u \in H^2(D)$, and there exists a constant $\tilde{C} > 0$, independent of f (but depending on D), such that*

$$\|u\|_2 \leq \tilde{C} \|f\|_{L^2(D)}. \quad (2.43)$$

Note that this extra regularity can be generalised to elliptic problems in higher dimensions [68]. ♦

2.2.2 The Finite Element Method

Having defined the notion of a weak solution of a PDE in Section 2.2.1 we can now discuss how one can approximate the solution of the variational formulation Eq. (2.38). By moving to the variational formulation we have weakened the smoothness requirements on our solution. In spite of this the variational formulation is still an equation which has to be solved in an infinite dimensional function space. As such we must turn to a discretisation into a finite dimensional space if we are to have any hope of approximating our solution in a computationally tractable manner. FEM utilises the so-called *Galerkin approximation* of the variational problem (2.38) and seeks an approximate solution $u_h \in V_h$ within a finite dimensional subspace V_h of $H_0^1(D)$ which satisfies

$$a(u_h, v) = \langle f, v \rangle, \quad \forall v \in V_h. \quad (2.44)$$

The construction of the finite element method arises from a specific Galerkin approximation, where we introduce a *triangulation* of the domain, $\mathcal{T}_h = \{\omega_e\}_{e=1}^{n_e}$, where the ω_e are pairwise disjoint triangular elements with $\overline{D} = \cup_{e=1}^{n_e} \overline{\omega_e}$, and h is the diameter of the largest element. This triangulation is usually referred to as the finite element mesh and we will refer to h as the *mesh width*. The discrete function space V_h is then chosen as a subspace of the space of piecewise polynomial functions

defined on this mesh. The approximate weak solution u_h can be expressed as a linear combination of basis functions $\{\varphi_i\}_{i=1}^{n_u}$. In this thesis, we assume that V_h is the space of continuous piecewise linear basis functions, and that the $\{\varphi_i\}_{i=1}^{n_u}$ are chosen to be a *nodal basis* which satisfies $\varphi_i(\mathbf{x}_j) = \delta_{ij}$, where $\mathbf{x}_1, \dots, \mathbf{x}_{n_u}$ are the internal vertices of the mesh. Both the FEM mesh and the nodal basis functions are illustrated in Fig. 2.1. Note that the nodal basis function shown in Fig. 2.1b is not necessarily associated with the mesh shown in Fig. 2.1a.

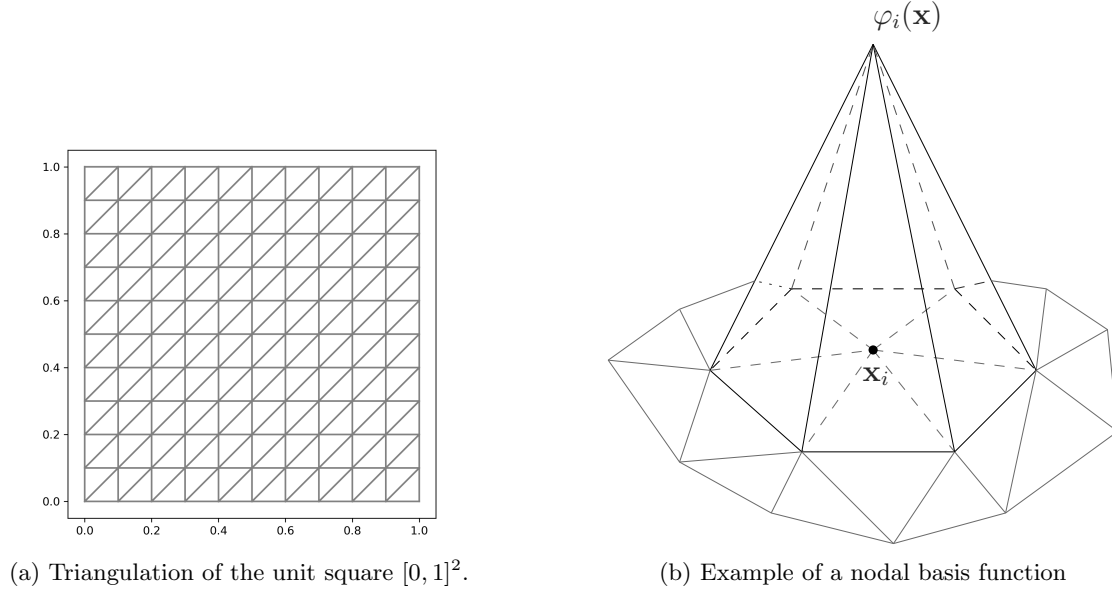


Figure 2.1: Illustration of the FEM mesh and nodal basis functions.

Remark 2.7. Note that our choice for V_h means that the finite element (FE) method we consider in this thesis is what is called a *conforming* method [89, see Section 5.6] in the sense that V_h is a subset of the true solution space $H_0^1(D)$ and the bilinear and linear forms in the variational formulation of our elliptic BVP remain the same in the discretised version Eq. (2.44). ♦

By exploiting the linearity of both the bilinear form and the inner product in Eq. (2.44) the FEM discretisation of the variational formulation of our BVP yields the following discrete system of linear equations:

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \quad (2.45)$$

where $\mathbf{A} \in \mathbb{R}^{n_u \times n_u}$ is the *stiffness matrix* whose ij -th element is given by:

$$A_{ij} = a(\varphi_j, \varphi_i) = \int_D \nabla \varphi_i \cdot (\boldsymbol{\kappa} \nabla \varphi_j) \, d\mathbf{x}. \quad (2.46)$$

The vector $\mathbf{u} \in \mathbb{R}^{n_u}$ is the vector of degrees of freedom (DOFs) and $\mathbf{f} \in \mathbb{R}^{n_u}$ is the *load vector* whose i -th entry is given by:

$$f_i = \langle f, \varphi_i \rangle = \int_D f \varphi_i \, d\mathbf{x}. \quad (2.47)$$

The stiffness matrix, $\mathbf{A}$, is a real symmetric matrix which is invertible by Assumption 2.4 [101, see Theorem 2.16]. Thus, the system of linear equations Eq. (2.45) can be solved for $\mathbf{u}$, yielding the finite element approximation $u_h = \Phi^T \mathbf{u}$ where $\Phi := (\varphi_1, \dots, \varphi_{n_u})^T$ is the vector of the FE basis functions. The *mass matrix*, $\mathbf{M} \in \mathbb{R}^{n_u \times n_u}$ which is defined by

$$M_{ij} = \langle \varphi_i, \varphi_j \rangle = \int_D \varphi_i \varphi_j \, d\mathbf{x}, \quad (2.48)$$

will also play an important role in the error analysis of Section 3.4.

It will be important for the analysis in Chapter 3 to have access to a bound on the error between the true solution and the FEM approximation in terms of h . To this end consider a family of triangulations $\{\mathcal{T}_h\}_{h>0}$. We will require a further technical assumption on the family of the finite element meshes:

Assumption 2.5. The meshes under consideration are quasi-uniform, i.e. there exists a constant, $\alpha > 0$, independent of h such that for any $h > 0$, and for all triangles $\omega \in \mathcal{T}_h$ we have:

$$\text{diam}(\omega) \geq \alpha h, \quad (2.49)$$

where $\text{diam}(\omega)$ is the diameter of the triangle ω and we recall that h is the diameter of the largest element. $\blacklozenge$

Under the assumptions above we have the following classical H^1 error bound [26, see Theorem 3.2.2]:

Theorem 2.4. *Let u be the unique weak solution to the BVP from Eqs. (2.27) and (2.28) and let u_h by the FEM approximation. Assume that $f \in L^2(D)$ and that the Assumptions 2.1, 2.4, and 2.5 hold. We then have the following H^1 error bound, for sufficiently small h :*

$$|u - u_h|_1 \leq Ch \|f\|_{L^2(D)}, \quad (2.50)$$

where $C > 0$ is a positive constant independent of h, f , and u . $\blacklozenge$

From this H^1 error bound a standard duality argument yields the following L^2 error bound [26, see Theorem 3.2.5], which will play an important role in the error analysis of Section 3.4.

Theorem 2.5. *Let u be the unique weak solution to the BVP from Eqs. (2.27) and (2.28) and let u_h by the FEM approximation. Assume that $f \in L^2(D)$ and that the Assumptions 2.1, 2.4, and 2.5 hold. We then have the following error bound, for sufficiently small h :*

$$\|u - u_h\|_{L^2(D)} \leq Ch^2 \|f\|_{L^2(D)}, \quad (2.51)$$

where $C > 0$ is a positive constant independent of h . $\blacklozenge$

2.3 Probability Theory and Inference

2.3.1 Basics of Probability Theory

In this section we introduce and discuss the essential probability-theoretic concepts we will require for this thesis. Further information can be found in Appendix A.3 and for a more in-depth treatment of probability (and measure) theory the reader is referred to [30, 11, 49].

For a measurable space $\mathcal{E}$, let $\mathcal{B}(\mathcal{E})$ denote the Borel sigma-algebra generated by the open sets of $\mathcal{E}$. We will denote by $\mathcal{P}(\mathcal{E})$ the set of all Borel probability measures on the measurable space $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$. Note that we will typically use lower-case Greek letters to denote probability measures, i.e. $\mu, \nu \in \mathcal{P}(\mathcal{E})$.

Given two probability measures μ, ν on the same measurable space $(\Omega, \mathcal{F})$, we say that μ is *absolutely continuous* with respect to ν if $\mu(A) = 0$ for all events $A \in \mathcal{F}$ such that $\nu(A) = 0$. We denote this by $\mu \ll \nu$. The probability measures are said to be *equivalent* if $\mu \ll \nu$ **and** $\nu \ll \mu$. When the probability measures μ and ν are instead concentrated on disjoint sets then we say they are *singular* and we will denote this by $\mu \perp \nu$.

When $\mu \ll \nu$ then the Radon-Nikodym Theorem [146, Theorem 6.1] gives us that there exists a ν -measurable function $f : \Omega \rightarrow [0, \infty)$ such that, for all events $A \in \mathcal{F}$ we have:

$$\mu(A) = \int_A f(x) d\nu(x). \quad (2.52)$$

The function f is called the *Radon-Nikodym derivative* of μ with respect to ν and is also denoted as $f = \frac{d\mu}{d\nu}$. We will often refer to f as simply the *density* of μ with respect to ν .

Random variables

An $\mathcal{E}$ -valued random variable is a measurable function $X : \Omega \rightarrow \mathcal{E}$. A measurable function $f : \Omega \rightarrow \mathcal{E}$ is one such that the pre-image of any Borel set $E \in \mathcal{B}(\mathcal{E})$ belongs to the underlying sigma algebra $\mathcal{F}$, i.e. $f^{-1}(E) \in \mathcal{F}$ for all $E \in \mathcal{B}(\mathcal{E})$. Note that we will typically use upper-case Latin letters to denote random variables.

Remark 2.8. A random variable $X : \Omega \rightarrow \mathcal{E}$ induces a probability measure μ_X on the measurable space $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ via:

$$\mu_X(E) = \mu(X^{-1}(E)) = \mu(\{\omega \in \Omega | X(\omega) \in E\}), \quad \forall E \in \mathcal{B}(\mathcal{E}), \quad (2.53)$$

yielding a probability space $(\mathcal{E}, \mathcal{B}(\mathcal{E}), \mu_X)$. The induced probability measure, μ_X , is referred to as

the *law* or *distribution* of the random variable X and can be expressed as the *pushforward* of μ by X , $\mu_X := X_{\#}\mu$. More generally we will use the notation $Y \sim \nu$ to denote that a random variable Y is distributed according to the law ν . ♦

Let X, Y, Z be $\mathcal{E}$ -valued random variables. We use the notation $X \perp\!\!\!\perp Y$ to denote that X and Y are independent. Similarly we write $X \perp\!\!\!\perp Y|Z$ to denote that X and Y are conditionally independent given Z .

For a random variable $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$ the *probability density function* (p.d.f.), $p_{\mathbf{X}} : \mathbb{R}^d \rightarrow [0, \infty)$ is simply the Radon-Nikodym derivative of the law of $\mathbf{X}$ with respect to the Lebesgue measure on $\mathbb{R}^d$.

2.3.2 The Multivariate Normal Distribution

In this subsection we will introduce and discuss the *multivariate normal distribution* which will be useful throughout this thesis.

Definition 2.7 (Multivariate Normal Distribution). A random vector $\mathbf{X} \in \mathbb{R}^d$ is said to follow a *multivariate normal distribution* with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ written

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2.54)$$

if and only if for any arbitrary fixed (deterministic) vector $\mathbf{a} \in \mathbb{R}^d$ we have that the inner product of $\mathbf{a}$ with $\mathbf{X}$ follows a normal distribution, i.e.

$$\langle \mathbf{a}, \mathbf{X} \rangle = \mathbf{a}^T \mathbf{X} \sim \mathcal{N}(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}). \quad (2.55)$$

In the case where $\boldsymbol{\Sigma}$ is invertible $\mathbf{X}$ has the following p.d.f.:

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right), \quad (2.56)$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$, $|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma})$. ♦

Remark 2.9. We also allow the case of singular covariance matrices, $\boldsymbol{\Sigma}$, in Definition 2.7. Such cases correspond to the centred random vector, $\mathbf{X} - \boldsymbol{\mu}$, being constrained to lie in the lower-dimensional subspace corresponding to the orthogonal complement of the eigenspace of $\boldsymbol{\Sigma}$ with associated zero eigenvalue(s). Further note that the fully degenerate case of $\boldsymbol{\Sigma} = \mathbf{0}$ corresponds to a point-mass located at the mean, $\boldsymbol{\mu}$, i.e. $\mathcal{N}(\boldsymbol{\mu}, \mathbf{0})$ is considered to be the same as the probability measure given by a Dirac-delta measure located at $\boldsymbol{\mu}$, $\delta_{\boldsymbol{\mu}} : \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$, defined via $\delta_{\boldsymbol{\mu}}(A) =$

$\mathbb{1}_A(\boldsymbol{\mu})$, for all $A \in \mathcal{B}(\mathbb{R}^d)$, where $\mathbb{1}_A$ is the indicator function of the set A defined by:

$$\mathbb{1}_A(\mathbf{x}) := \begin{cases} 1, & \mathbf{x} \in A \\ 0, & \mathbf{x} \notin A. \end{cases} \quad (2.57)$$

In this degenerate case the p.d.f. of $\mathbf{X}$ is undefined. However, it is possible to interpret the p.d.f. here as the Dirac-delta ‘*function*’, $\delta_{\boldsymbol{\mu}}$, though this does not technically fit with our definition of p.d.f.. ♦

Remark 2.10. The relevant dimension, d , should be clear from the sizes of the random variables and their respective means and covariances. We will occasionally subscript the symbol, $\mathcal{N}$, with an integer to make explicit the corresponding dimension of the random vector, i.e. we will write $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that $\mathbf{X}$ is a random vector in $\mathbb{R}^d$ following a multivariate normal distribution. ♦

For completeness, we include in Appendix A.3 several well-known properties of the multivariate normal distribution which will be useful throughout the thesis.

2.3.3 Bayesian Inference

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Recall that given events $A, B \in \mathcal{F}$ with $\mathbb{P}(A), \mathbb{P}(B) > 0$ we can define (see additional background material in Appendix A.3.1) the conditional probability of A given B and of B given A via:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad (2.58)$$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}. \quad (2.59)$$

These two equations can be combined to yield the famous Bayes’ rule:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \quad (2.60)$$

This important update rule was derived by Thomas Bayes [7] and is of fundamental significance to the field of statistical inference. In the Bayesian setting, one often interprets the various components of Eq. (2.60) in the following manner:

- **Prior:** $\mathbb{P}(A)$ can be viewed as quantifying the *prior* probability of some event of interest, A , before any further information has been collected.
- **Likelihood:** $\mathbb{P}(B|A)$ can be viewed as describing the probability of observing some information represented by the event B given that A holds/is true. Often the event B has a

causal, albeit inaccurate or noisy relationship with the event A ; the *likelihood* quantifies this relationship.

- **Marginal Likelihood:** $\mathbb{P}(B)$ is often referred to as the *marginal likelihood* of the event B and this term quantifies how likely it is to observe the event B irregardless if A holds/is true.
- **Posterior:** $\mathbb{P}(A|B)$ represents the *posterior* probability of A holding/being true given that we have observed some information represented by the event B .

Thus, the Bayesian interpretation of probability can be viewed as quantifying one's belief in whether some event holds/is true. Bayes' rule is the mechanism which allows one to combine observations/data to update one's *a priori* belief to yield an *a posteriori* belief of an event.

Of particular relevance to this thesis is the Bayesian formulation of statistical inference: any unknown parameter of interest, θ , is treated as **random** and has an associated *prior* distribution, $p(\theta)$, placed on it representing our prior beliefs about the parameter **before** observing any data. This prior is then combined with a *likelihood*, $p(\mathbf{y}|\theta)$, which captures the often noisy relationship between the unknown parameter and the observed data $\mathbf{y}$. Given this observed data one can then uses Bayes' rule to update our belief in the parameter, obtaining the *posterior* distribution $p(\theta|\mathbf{y})$ which in this scenario can be expressed as:

$$p(\theta|\mathbf{y}) = \frac{p(\theta)p(\mathbf{y}|\theta)}{p(\mathbf{y})}, \quad (2.61)$$

where $p(\mathbf{y})$ is formally given via:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta. \quad (2.62)$$

Remark 2.11. One key challenge to adopting Bayesian methods is the difficulty in computing the posterior distribution Eq. (2.61). This difficulty stems from the underlying difficulty in computing the integral required to obtain the marginal likelihood Eq. (2.62). There are only a handful of scenarios where it is possible to compute $p(\mathbf{y})$ analytically in closed form, for example when dealing with multivariate normal distributions (see Proposition A.4). This barrier prevented Bayesian models from being widely adopted until the mid-1990s [127]¹ which saw the advent of Markov Chain Monte Carlo (MCMC) algorithms. These algorithms generate a sequence of numbers which are approximately distributed according to the posterior distribution by constructing a Markov Chain with a stationary distribution which can be shown to coincide with the required posterior distribution [22, 24, 151]. More recent developments involve the advent of Hamiltonian Monte

¹↑See this reference for a short history of MCMC.

Carlo (HMC) [9, 110] which has risen in popularity and can be used to estimate high-dimensional posterior distributions with complex geometries when gradients are available. We will utilize an implementation of the popular No-U-Turn Sampler (NUTS) algorithm [73], which is an extension of HMC, found in the Python package NumPyro [118, 12] in Chapter 6 for the purposes of estimating the parameters of a Bayesian model. In Section 5.3 we present a brief overview of the HMC algorithm. For a more detailed overview of MCMC methods the interested reader is directed to [128]. ♦

2.3.4 Gaussian Measures on Hilbert Spaces

We now present some results regarding the construction of Gaussian measures on infinite dimensional Hilbert spaces. We start by introducing the concept of a *stochastic process*, which can be viewed as the generalisation of random variables to infinite dimensional spaces, adopting the definition from [112]:

Definition 2.8 (*Stochastic process*). A *stochastic process* is an indexed collection of random variables

$$\{X(t, \omega) \mid t \in T, \omega \in \Omega\}, \quad (2.63)$$

defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where T is the index set. When the context is clear it is common to suppress the ω argument, shortening $X(t, \omega)$ to $X(t)$. For a fixed $\omega \in \Omega$ we call the function $X(\cdot, \omega)$ a *sample path* of the process. ♦

From a statistical inference perspective it can be useful to view $X(\cdot, \omega)$ as a random variable in some appropriate space of functions. The choice of such a function space is a fundamental modelling assumption and characterizes the relation between $X(t, \omega)$ and $X(t', \omega)$ for different $t, t' \in T$. In this thesis we will exclusively be concerned with the case of stochastic processes where the random functions $X(\cdot, \omega)$ lie in a separable Hilbert space $\mathcal{E}$, which we will, in an abuse of notation, denote by $X \in \mathcal{E}$. Typically we will take $\mathcal{E} = L^2(D)$ for some domain $D \subset \mathbb{R}^d$ which is taken to be the index set of the stochastic process.

Two important objects related to stochastic processes are their expectation/mean function and covariance operator. These are generalisations of the usual expectation and covariance matrix in finite dimensions to infinite dimensional Hilbert spaces. Due to the infinite dimensional nature some care is needed in defining these generalisations. There are two common notions of expectation for stochastic processes given by the following two types of integrals:

- **Bochner integral:** Similar to the construction of Lebesgue integrals one can start by defining an appropriate class of “simple functions” consisting of linear combinations of indicator functions, whose expectations are given by the usual Lebesgue integral, and then

define the expectation of our stochastic process X via the limit of a sequence of “simple functions” which approximates our stochastic process. The resulting expectation is known as the *Bochner integral*. A stochastic process is *Bochner integrable* if and only if the univariate random variable $\|X\|(\omega) := \|X(\cdot, \omega)\|$ (where $\|\cdot\|$ is the norm on $\mathcal{E}$) satisfies $\mathbb{E}(\|X\|) < \infty$ [16].

- **Pettis integral:** Since for each $t \in T$ we have that $X(t, \cdot)$ is a (multivariate) random variable we know how to compute the expectation $\mathbb{E}(X(t, \cdot))$. This idea leads to the important notion of a stochastic process being *Pettis integrable* or *weakly integrable* if there exists some function $m \in \mathcal{E}$ such that $\mathbb{E}[\ell(X)] = \ell(m)$ for all bounded linear functionals $\ell \in \mathcal{E}'$. It is possible to show that if a function is Bochner integrable that the Pettis integral coincides with the Bochner integral [14].

For more details on these two notions of expectation the reader is directed to [14]. In this thesis we will henceforth take the expectation of a stochastic process to be the Pettis integral. We will say that a stochastic process is *centred* if $\mathbb{E}[X] = 0$.

The covariance operator of a stochastic process on a Hilbert space $\mathcal{E}$ can be defined as follows [146, Section 6.3]:

Definition 2.9 (Covariance operator). Let X be a stochastic process in the Hilbert space $\mathcal{E}$. Let $m = \mathbb{E}(X) \in \mathcal{E}$ denote its mean function. A linear operator $K : \mathcal{E} \rightarrow \mathcal{E}$ is called the *covariance operator* if, for all $g, g' \in \mathcal{E}$ one has,

$$\langle g', Kg \rangle = \mathbb{E}(\langle g', (X - m) \rangle \langle (X - m), g \rangle), \quad (2.64)$$

where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathcal{E}$. Note the above Eq. (2.64) can be written compactly as $K = \mathbb{E}[(X - m) \otimes (X - m)]$ and this is analogous to the finite-dimensional case Eq. (A.26) presented in Appendix A.3.1. ♦

Having defined both the mean function and the covariance operator for stochastic processes we can now introduce the concept of a Gaussian measure on $\mathcal{E}$, which can be viewed as a generalisation of their finite dimensional counterparts on $\mathbb{R}^d$, the multivariate normal distribution (Definition 2.7) as follows:

Definition 2.10 (Gaussian measure). A measure μ on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$, where $\mathcal{E}$ here is an infinite-dimensional Hilbert space, is a *Gaussian measure* if, for any linear functional $\ell \in \mathcal{E}'$ the pushforward measure $\ell_{\#}\mu$ is a univariate normal distribution. Due to the Riesz Representation Theorem (see Theorem A.1), one has $\mathcal{E}' = \mathcal{E}$, and we can characterise the above as follows: a stochastic process $X \in \mathcal{E}$ follows a Gaussian measure, $X \sim \mu$, if and only if for any fixed (deterministic) $g \in \mathcal{E}$ one

has $\langle g, X \rangle \sim \mathcal{N}(m_g, \sigma_g^2)$ for some $m_g \in \mathbb{R}, \sigma_g^2 \geq 0$. Note that $m_g = \langle g, \mathbb{E}[X] \rangle$ and $\sigma_g^2 = \langle g, Kg \rangle$ where $\mathbb{E}[X], K$ are the mean and covariance operator of X respectively, and are well-defined in this definition. $\blacklozenge$

Note that this closely matches the relationship between the multivariate and univariate normal distributions since for $\mathcal{E} = \mathbb{R}^d$ one has that any continuous linear functional on $\mathbb{R}^d$ can precisely be written as $\ell(\mathbf{v}) = \mathbf{a}^T \mathbf{v}$ for some fixed $\mathbf{a} \in \mathbb{R}^d$. We also note that in Definition 2.10 it is not strictly necessary to restrict $\mathcal{E}$ to a Hilbert space only; the definition can be generalised to Banach spaces [146, Section 6.3]. We will only consider Gaussian measures on Hilbert spaces in this thesis so we exclude this generalisation.

The following result, adapted from [41] shows that the mean function and covariance operator completely characterise a Gaussian measure on a Hilbert space:

Theorem 2.6. *A Gaussian measure μ on a Hilbert space $\mathcal{E}$ with mean function m and covariance operator K has the characteristic function:*

$$\varphi(g) = \exp \left(i \langle g, m \rangle - \frac{1}{2} \langle g, Kg \rangle \right), \quad \forall g \in \mathcal{E}. \quad (2.65)$$

$\blacklozenge$

This theorem justifies the use of the notation $\mathcal{N}(m, K)$ to denote a Gaussian measure with mean function m and covariance operator K . Further, note that given any function $m \in \mathcal{E}$ and any positive semi-definite, trace-class operator $K : \mathcal{E} \rightarrow \mathcal{E}$ there exists a Gaussian measure $\mathcal{N}(m, K)$ with mean m and covariance operator K [41, Theorem 1.12].

2.3.5 Gaussian Processes

Throughout this thesis we will rely on the concept of GPs which will be related to Gaussian measures below. Let $D \subset \mathbb{R}^d$ be a bounded and connected domain. In order to define a GP it will be helpful to first introduce the notion of a positive-definite kernel function:

Definition 2.11 (Positive definite kernel function). A function $c : D \times D \rightarrow \mathbb{R}$ is said to be a *positive (semi-)definite kernel function* if it is symmetric, i.e. $c(\mathbf{x}, \mathbf{y}) = c(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in D$, and if for any finite number of arbitrary points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset D$ the symmetric matrix $\mathbf{C} := \{c(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n \in \mathbb{R}^{n \times n}$ is positive (semi-)definite, i.e.

$$\mathbf{v}^T \mathbf{C} \mathbf{v} > 0, \quad \forall \mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}, \quad (2.66)$$

where in the positive semi-definite case we replace the strict inequality in Eq. (2.66) by $\geq$. $\blacklozenge$

We can now define a GP as follows:

Definition 2.12 (Gaussian Process). Given a positive definite kernel function $k : D \times D \rightarrow \mathbb{R}$ and a function $m : D \rightarrow \mathbb{R}$ we say that a stochastic process $X : D \rightarrow \mathbb{R}$ is a *Gaussian Process (GP)* with mean function m and covariance function k if for any arbitrary finite collection of distinct points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset D$ the resulting random vector $\mathbf{X} := (X(\mathbf{x}_1), \dots, X(\mathbf{x}_n))^T$ follows a multivariate normal distribution with mean vector $\mathbf{m} := (m(\mathbf{x}_1), \dots, m(\mathbf{x}_n))^T$ and covariance matrix $\mathbf{K} := \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n \in \mathbb{R}^{n \times n}$:

$$\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}). \quad (2.67)$$

The mean function and kernel/covariance function completely determine the GP. We write:

$$X \sim \mathcal{GP}(m, k), \quad (2.68)$$

to denote the Gaussian process with mean function m and covariance function k . ♦

Remark 2.12. Note that it is possible to define GPs on unbounded domains $D \subset \mathbb{R}^d$ including on the entirety of $\mathbb{R}^d$; the definition is exactly the same as in Definition 2.12. For the purposes of this thesis we will be relying solely on GPs defined on bounded domains and so we restrict ourselves to this setting as this simplifies the presentation of the link between Gaussian measures and GPs presented below. ♦

Linking the two viewpoints

We now discuss an important connection between Gaussian measures and Gaussian processes. Let $f \sim \mathcal{GP}(m, k)$ be a Gaussian process which takes values in $\mathcal{E} = L^2(D)$, i.e. the stochastic process given by f lies in $\mathcal{E}$ (so in particular $m \in \mathcal{E}$). Under the assumption that the covariance function, k , is continuous we have that the GP can be associated with the Gaussian measure $\mathcal{N}(m, K)$ on $\mathcal{E}$ with mean m and covariance operator K , which is related to the covariance function of the GP via the equation:

$$Kg(\cdot) = \int_D k(\cdot, \mathbf{y})g(\mathbf{y})d\mathbf{y}, \quad g \in \mathcal{E}, \quad (2.69)$$

i.e. the covariance operator associated with the GP, $\mathcal{GP}(m, k)$, is an integral operator with kernel given by the corresponding covariance function k . Continuity of k guarantees that the resulting covariance operator K is trace-class.

Conversely, given a Gaussian measure $\mathcal{N}(m, K)$ on $\mathcal{E} = L^2(D)$ there exists [124, Theorem 2 (b)] an associated Gaussian process with paths in $L^2(D)$ which induces the Gaussian measure

on $L^2(D)$. In particular, note that since the covariance operator K is Hilbert-Schmidt it can be written as an integral operator with some kernel $k \in L^2(D \times D)$. Since it can then be shown that the covariance operator K is also trace-class [14, see for example Theorem 2.3.1] it follows that the kernel k is in fact continuous and can be taken to be the covariance function of the corresponding GP associated to $\mathcal{N}(m, K)$.

Properties of GPs/Gaussian measures

We now briefly discuss some well known results regarding GPs and Gaussian measures. One very useful property of GPs is the transformation property given in the following proposition

Proposition 2.3 (Transformation property of GPs). *Let $f \sim \mathcal{GP}(m, k)$ be a GP associated with the Gaussian measure $\mathcal{N}(m, K)$ on $L^2(D)$. Let $\mathcal{L} : L^2(D) \rightarrow \mathcal{X}$ be a bounded linear operator to some other finite/infinite-dimensional Hilbert space $\mathcal{X}$. Assume that m lies in the domain of $\mathcal{L}$ and k lies in the domain of $\mathcal{L}\mathcal{L}^\dagger$. It follows that the pushforward of f by $\mathcal{L}$ is also a GP:*

$$\mathcal{L}f \sim \mathcal{GP}(\mathcal{L}m, \mathcal{L}\mathcal{L}^\dagger k), \quad (2.70)$$

where the action of $\mathcal{L}\mathcal{L}^\dagger$ on k is equivalent to taking the following:

$$(\mathcal{L}\mathcal{L}^\dagger k)(\mathbf{x}, \mathbf{x}') = \mathcal{L}_{\mathbf{x}} \mathcal{L}_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}'), \quad (2.71)$$

where $\mathcal{L}_{\mathbf{x}}$ denotes the action of $\mathcal{L}$ on the first argument of k and $\mathcal{L}_{\mathbf{x}'}$ its action on the second argument of k . Eq. (2.70) is equivalent to stating that $\mathcal{L}f$ is distributed according to the following Gaussian measure on $\mathcal{X}$:

$$\mathcal{L}f \sim \mathcal{N}(\mathcal{L}m, \mathcal{L}K\mathcal{L}^\dagger). \quad (2.72)$$

◆

Throughout this thesis we will often be interested in modelling functions in a Bayesian manner (see Section 2.3.3) and we will typically assign GP priors to model our functions. In order to make predictions with our GP priors we will need to be able to condition them on observed data points and obtain the relevant posterior distributions.

In the remainder of this section we briefly outline how we can obtain the posterior distributions for GPs. To be more precise we are interested in the following scenario: we place a GP prior on a function of interest,

$$f \sim \mathcal{GP}(m, k). \quad (2.73)$$

Here we assume f is a function on $D \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$. We assume that we obtain $n \in \mathbb{N}$

noisy observations of f on some grid of n locations $\mathcal{D}_n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset D$, i.e. we observe $\mathbf{y} := (y_1, \dots, y_n)^T$ where:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad \sigma > 0, \quad (2.74)$$

where the noise term $\boldsymbol{\epsilon} := (\epsilon_1, \dots, \epsilon_n)^T$ is assumed to be independent from the GP f . We wish to condition our GP prior on the noisy observation $\mathbf{y}$ to obtain the posterior of $f|\mathbf{y}$. We can then use this posterior to make predictions for f at new locations. In the case of GPs the resulting posterior will also define a GP whose mean and covariance can be computed in closed form which the following well-known result, adopted from Theorem 3.1 in [78] shows:

Proposition 2.4. *Let f have the GP prior specified by Eq. (2.73). Suppose $n \in \mathbb{N}$ noisy observations, $\mathbf{y}$, of f are made according to Eq. (2.74). It follows that the posterior of f given $\mathbf{y}$ is also a GP, with the following form:*

$$f|\mathbf{y} \sim \mathcal{GP}(\tilde{m}, \tilde{k}), \quad (2.75)$$

where the posterior mean and covariance functions have the following forms:

$$\tilde{m}(\mathbf{x}) := m(\mathbf{x}) + \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} (\mathbf{y} - \mathbf{m}), \quad \mathbf{x} \in D, \quad (2.76)$$

$$\tilde{k}(\mathbf{x}, \mathbf{x}') := k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^T (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k}(\mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in D, \quad (2.77)$$

where here $\mathbf{k}(\mathbf{x})$ is the (column) vector in $\mathbb{R}^n$ with i -th entry given by

$$(\mathbf{k}(\mathbf{x}))_i := k(\mathbf{x}_i, \mathbf{x}), \quad (2.78)$$

$\mathbf{m}$ is the vector in $\mathbb{R}^n$ with i -th entry given by $m(\mathbf{x}_i)$, and $\mathbf{K} := \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n \in \mathbb{R}^{n \times n}$, similar to the notation used in Section 2.3.5. ♦

Thus, the posterior distribution has a form where the mean and covariance functions can be obtained using simple linear algebra via Eqs. (2.76) and (2.77). The proof of this result is given in Appendix B.1.

Remark 2.13. In light of Proposition 2.4, we can easily make predictions from GP posteriors by evaluating the posterior at the required test locations $\mathcal{D}_m^* := \{\mathbf{x}_1^*, \dots, \mathbf{x}_m^*\} \subset D$ for any $m \in \mathbb{N}$ and simulating a multivariate normal random vector with mean and covariance given by evaluating Eqs. (2.76) and (2.77) on the grid of test locations. That is to say, we can make predictions for $(f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_m^*))^T$ from the posterior Eq. (2.75) by drawing a random vector $\mathbf{w}^* \sim \mathcal{N}_m(\mathbf{m}^*, \boldsymbol{\Sigma}^*)$ where $\mathbf{m}^* \in \mathbb{R}^m$ has i -th entry given by $\tilde{m}(\mathbf{x}_i^*)$ and $\boldsymbol{\Sigma}^* := \{\tilde{k}(\mathbf{x}_i^*, \mathbf{x}_j^*)\}_{i,j=1}^m$. ♦

It is also possible to obtain the conditional distribution of a Gaussian measure on a Hilbert

space in a similar manner to Proposition 2.4 for GPs. This is presented in Theorem A.2 in Appendix A.3.1.

For a more comprehensive overview of Gaussian measures/processes we refer the reader to the works [14, 123, 125, 97]. [125] deals with GPs in a more practical manner than [14, 123, 97] where the treatments are more measure-theoretic in flavour. The Gaussian measure/process view is also closely connected to the *random field perspective* which is discussed in great detail in the works [146, 42].

2.3.6 Divergences on Spaces of Probability Measures

Central to the analysis in this thesis is a means of measuring the divergence between probability distributions, i.e. being able to quantify how different two probability distributions are. Perhaps one of the most popular discrepancies between probability measures is the KL divergence which we define below:

Definition 2.13 (*Kullback-Leibler divergence*). Let μ, ν be two probability measures where $\mu \ll \nu$. The *Kullback-Leibler (KL) divergence* from μ to ν is given as follows:

$$\text{KL}(\mu \parallel \nu) := \mathbb{E}_\mu \left[\log \left(\frac{d\mu}{d\nu} \right) \right], \quad (2.79)$$

where $\frac{d\mu}{d\nu}$ is the Radon-Nikodym derivative of μ with respect to ν . In the common case where μ, ν are probability measures on $\mathbb{R}^d$ with densities, p, q , respectively we can express the KL divergence as:

$$\text{KL}(\mu \parallel \nu) := \int_{\mathbb{R}^d} \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) p(\mathbf{x}) d\mathbf{x}, \quad (2.80)$$

and in this scenario we often refer to Eq. (2.80) in terms of the densities as the KL divergence from p to q . ♦

Remark 2.14. Note that the KL divergence is not a metric (see Definition A.1) on the space of probability measures due to the lack of symmetry of Definition 2.13. ♦

The KL divergence is a key ingredient in Stochastic Variational Inference (SVI), a popular approach to performing Bayesian inference.

We now introduce the *Procrustes distance*, as a divergence between centred Gaussians, which will be useful when discussing the Wasserstein distance below. The definition below is adopted from [119].

Definition 2.14 (*Procrustes Distance*). Let $C_i : \mathcal{E} \rightarrow \mathcal{E}$ be covariance operators for $i = 1, 2$. Let S_i each be (not necessarily symmetric) square roots of C_i , i.e. $C_i = S_i^\dagger S_i$ for $i = 1, 2$. The

Procrustes distance between C_1, C_2 , denoted by $d_P(C_1, C_2)$, is defined as:

$$d_P(C_1, C_2) := \inf_{R: R^\dagger R = I} \|S_1 - S_2 R\|_{\text{HS}}, \quad (2.81)$$

where I is the identity operator on $\mathcal{E}$ and $R: \mathcal{E} \rightarrow \mathcal{E}$ is thus a unitary operator. $\blacklozenge$

As noted in Definition 2.13 a key requirement is that the two probability measures being compared are absolutely continuous; the KL divergence cannot be used to compare two mutually singular distributions. For the purposes of studying the error in StatFEM in Chapter 3 we will need to introduce a metric on the space of probability distributions which is able to handle the mutually singular distributions involved in this setting as will be justified in Section 3.4. The probability metric we will work with in Chapter 3 is the Wasserstein distance. Here we introduce the Wasserstein distance and outline some of its key properties which we will require.

Definition 2.15 (*Wasserstein-2 distance*). The *Wasserstein-2 distance* between two probability measures μ, ν on a normed space $\mathcal{E}$, having the first two moments finite, denoted by $W_2(\mu, \nu)$, is defined as:

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Gamma(\mu, \nu)} \int_{\mathcal{E} \times \mathcal{E}} \|x - y\|_{\mathcal{E}}^2 d\pi(x, y), \quad (2.82)$$

where $\Gamma(\mu, \nu)$ is the set of *couplings* of μ and ν , i.e.

$$\Gamma(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{E} \times \mathcal{E}) \mid \pi(A \times \mathcal{E}) = \mu(A) \text{ and } \pi(\mathcal{E} \times B) = \nu(B) \text{ for all } A, B \in \mathcal{B}(\mathcal{E})\}. \quad (2.83)$$

$\blacklozenge$

Throughout the thesis we will abbreviate $W_2(\mu, \nu) = W(\mu, \nu)$.

Remark 2.15. Note that the integral in Eq. (2.82) can be written as $\mathbb{E}_\pi(\|X - Y\|_{\mathcal{E}}^2)$ where the random variables X and Y have joint distribution $(X, Y) \sim \pi$. Equation (2.83) says that a coupling $\pi \in \Gamma(\mu, \nu)$ is a probability measure on the product space $\mathcal{E} \times \mathcal{E}$ whose marginals in the first and second arguments are μ and ν respectively. This is illustrated in Fig. 2.2. Just as with L^p norms one can consider a Wasserstein- p distance where the squares in Eq. (2.82) are replaced by powers of p for $p \in [1, \infty)$. Note that in the literature the Wasserstein-1 distance is often referred to as the Earth-Mover's distance, especially in applied fields, since if one intuitively views one of the probability measures, say μ , as a pile of sand and the other, say ν , as a hole in the ground then the Wasserstein-1 distance can be thought of as the minimum “cost”, when measured according to $\|x - y\|_{\mathcal{E}}$, required to transport the pile of sand to the hole. $\blacklozenge$

Henceforth, when we refer to the Wasserstein distance we mean the Wasserstein-2 distance unless otherwise stated.

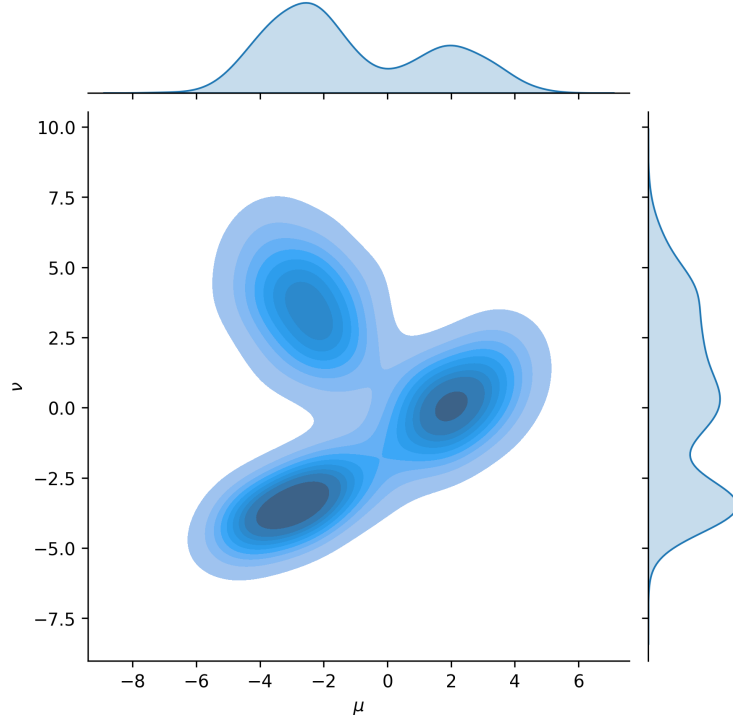


Figure 2.2: Coupling between two distributions.

It will be useful in this thesis to note that the Wasserstein distance between two probability measures μ, ν can be linked to the Wasserstein distance between the centered measures. Let m denote the expectation of μ ; then the centered measure μ^* is defined by,

$$\mu^*(E) = \mu(\{x - m | x \in E\}), \quad \forall E \in \mathcal{B}(\mathcal{E}). \quad (2.84)$$

Denoting the means of μ, ν by m_1, m_2 respectively we have from [40],

$$W^2(\mu, \nu) = \|m_1 - m_2\|_{\mathcal{E}}^2 + W^2(\mu^*, \nu^*). \quad (2.85)$$

We will require an explicit expression for the Wasserstein distance between two Gaussian. Let $\mu_i = \mathcal{N}(m_i, \Sigma_i), i = 1, 2$ be two Gaussian measures on a separable Hilbert space $\mathcal{E}$. Then from [105], we have

$$W^2(\mu_1, \mu_2) = \|m_1 - m_2\|_{\mathcal{E}}^2 + \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2 \text{tr} \sqrt{\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}}. \quad (2.86)$$

Remark 2.16. Note that combining Eq. (2.85) and Eq. (2.86) one obtains:

$$W^2(\mathcal{N}(0, \Sigma_1), \mathcal{N}(0, \Sigma_2)) = \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2 \text{tr} \sqrt{\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}}, \quad (2.87)$$

as the Wasserstein distance between two mean-zero Gaussian measures. $\blacklozenge$

The following bound found in [27, see Lemma S2] will be very important in the error analysis in Chapter 3,

Lemma 2.1. Let $\mu_i = \mathcal{N}(a_i, C_i)$, $i = 1, 2$ each be Gaussian measures on a Hilbert space $\mathcal{E}$. Let S_i each be square roots of C_i such that $C_i = S_i^\dagger S_i$. Then we have that,

$$W^2(\mu_1, \mu_2) \leq \|a_1 - a_2\|_{\mathcal{E}}^2 + \|S_1 - S_2\|_{\text{HS}}^2. \quad (2.88)$$

$\blacklozenge$

For completeness we include a proof of this key result, adapted from the proof of Lemma S2 in [27]:

Proof of Lemma 2.1.

We have, using Eq. (2.85), that the Wasserstein distance between μ_1 and μ_2 can be written as:

$$W^2(\mu_1, \mu_2) = \|a_1 - a_2\|_{\mathcal{E}}^2 + W^2(\mathcal{N}(0, C_1), \mathcal{N}(0, C_2)), \quad (2.89)$$

transferring our problem to the Wasserstein-2 distance between two centred Gaussians. We can now utilise an important connection between the Wasserstein distance between centred Gaussian measures and the Procrustes metric between the respective covariance operators found in [105, see Proposition 3], namely that:

$$W(\mathcal{N}(0, C_1), \mathcal{N}(0, C_2)) = \inf_{R: R^\dagger R = I} \|S_1 - S_2 R\|_{\text{HS}} = d_P(C_1, C_2), \quad (2.90)$$

i.e. the Wasserstein-2 distance between the centered Gaussians is equal to the Procrustes distance between their respective covariance operators. We note here that in Proposition 3 of [105] the authors consider **symmetric** square roots of the covariance operators in their definition of the Procrustes distance. In the proof here we note that, as is discussed in the proof of Lemma S2 in [27], we need not express the Procrustes distance in terms of symmetric square roots [119]. This justifies Eq. (2.90) above. We can now finally take R in this equation to be the identity operator, $R = I$, which yields the following inequality:

$$W^2(\mathcal{N}(0, C_1), \mathcal{N}(0, C_2)) \leq \|S_1 - S_2\|_{\text{HS}}^2, \quad (2.91)$$

which yields Eq. (2.88) as required. $\blacksquare$

For a more detailed treatment of Wasserstein distances the reader is directed to [156, 134, 117].

[156, 134] are more theoretical in nature, while [117] focuses more on numerical methods for solving optimal transport problems (*the solutions being related to Wasserstein distances*).

Chapter 3

Theoretical Guarantees for the Statistical Finite Element Method

3.1 Introduction

This chapter contains the theoretical contributions for the first project of this thesis which were outlined in Chapter 1. In particular, this chapter is structured as follows. In Section 3.2 we provide a literature review of related work to that presented in this chapter. In Section 3.3 we present StatFEM, deriving it from first principles. Section 3.4 contains our main theoretical contributions, error bounds for the StatFEM prior and the posterior when this prior is conditioned on sensor observations. Our preliminary attempts to adapt these results to time-dependent problems are described in Section 3.5. Any additional details required for the proofs of the main results of this chapter are included in Appendix B.2.

3.2 Related Work

This chapter is focused on providing detailed error analysis for the StatFEM method first introduced in [65]. For clarity of ideas we start by limiting our attention to linear elliptic problems. More recently [81] have derived error rates, in terms of the expected L^2 norm, for the StatFEM posterior distribution both with and without discretisation when used to model some deterministic true system response function.

The error analysis for the StatFEM method is an example of analysing the effects of perturbations of the posterior distribution resulting from a perturbation of the underlying prior distribution. In the literature surrounding *well-posedness theory of Bayesian inverse problems* such perturbation results are commonplace. These however tend to focus on analysing posterior perturbations

resulting from likelihood perturbations [42, 146, 75, 76, 90, 149, 115, 142] as opposed to prior perturbations which have been studied more recently, see for instance [142, 61]. An important reason for fewer results being available for prior perturbations is the choice of metric/divergence for the distance used to compare the posteriors; typically the Hellinger distance or Total Variation distance are used which require the posteriors and hence the priors to be absolutely continuous with respect to a common reference probability measure (i.e. one of the measures has to be absolutely continuous with respect to the other). As will be discussed in Section 3.4 the priors and posteriors considered in the analysis of StatFEM are mutually singular which prohibits the use of these metrics.

There has been extensive research on deriving numerical methods for solving various classes of SPDEs. For detailed reviews of such methods the reader is referred to [101, 106, 147, 143, 160, 4]. Some of these methods, for example the Stochastic Galerkin Finite Element method (SG-FEM), introduced in Section 3.1, and its many variants, also find their roots in the classical finite element method, but differ fundamentally from StatFEM. More specifically, SG-FEM methods reformulate the SPDE as a deterministic PDE on a higher dimensional domain, where additional variables are introduced to characterise the underlying noise process. For example, if the underlying noise process is given by a truncated Karhunen-Lo  ve expansion (see Theorem A.4) consisting of m random basis functions then SG-FEM methods reformulate the problem to a weak formulation on a space of functions with $d + m$ variables. On the other hand, in the StatFEM approach, the approximation is itself intrinsically stochastic as will be expanded on in Section 3.3. While SG-FEM methods are applicable to a wider class of SPDEs, their scalability is limited by the curse of dimensionality resulting from the augmented noise dimensions. Additionally, while the StatFEM formulation permits fast assimilation of measurement data, this is not straightforward for SG-FEM and related methods.

In the area of spatial statistics, Gaussian random fields are constructed as the solution of appropriately chosen SPDEs [98]. The solution is approximated through a finite element discretisation of the weak form of the underlying SPDE. The resulting discretised solution is subsequently employed as a prior within a Bayesian inference problem for a spatial or spatio-temporal model [140]. For non-Gaussian likelihoods, fast approximate Bayesian Inference for the resulting latent Gaussian model is possible using an Integrated Nested Laplace Approximation (INLA) [104]. The construction of the StatFEM model is strongly connected to this statistical viewpoint of SPDEs. Indeed, in both cases, the spatial priors are Gaussian Markov Random Fields (GMRFs). In the case of StatFEM the mean and covariance of the GMRF reflect the behaviour of the underlying deterministic PDE, while in the SPDE case the covariance is tailored to ensure a specific degree of spatial decorrelation of the underlying latent Gaussian random field.

Conceptually, StatFEM also bears some similarity to Probabilistic Numerical Methods [72], both being statistical reformulations of numerical methods. Indeed, such methods have also been developed for approximating solutions of ODEs [152] as well as PDEs [31, 25, 28]. The two methodologies share common foundations, particularly Bayesian Probabilistic Numerical Methods [29] which are formulated as conditioning a prior distribution of potential solutions on partial evaluations of the underlying equations. StatFEM also builds on the Bayesian formalism, employing conditioning to incorporate new information, specifically noisy sensor observations, starting from a prior which characterises the distribution of approximate PDE solutions. However, these methods fundamentally differ in what they seek to achieve: Bayesian Probabilistic Numerical methods seek to recast numerical discretisation error as uncertainty, whereas the uncertainty arising in StatFEM primarily seeks to account for model error.

3.3 The Statistical Finite Element Method

In this section we will present the StatFEM model. We reviewed FEM in Section 2.2.2. In Section 3.3.1 we will derive the prior distribution over the solution to the PDE that is used in StatFEM. Finally, in Section 3.3.2 we will condition this prior distribution on sensor observations to obtain the StatFEM posterior distribution.

3.3.1 The StatFEM Model

In this section we present the StatFEM approximation for the solution of a stochastically-forced elliptic PDE. The resulting probability distribution will subsequently be employed as a prior to be combined with sensor data through Bayesian updates. To this end, we consider the BVP from Eqs. (2.27) and (2.28) where f is chosen to be a GP, $f \sim \mathcal{GP}(\bar{f}, k_f)$, where $\bar{f} \in L^2(D)$ and where we make the following assumption on the covariance function k_f and its associated covariance operator K .

Assumption 3.1. The eigenvalues, $\{\mu_j\}_{j=1}^\infty$, of K behave asymptotically like $j^{-\lambda}$ for some $\lambda \geq 0$, i.e. there exists constants $C_1, C_2 > 0$ such that $C_1 j^{-\lambda} \leq \mu_j \leq C_2 j^{-\lambda}$ for all $j \in \mathbb{N}$. Further, in the case of $\lambda \in [0, 1]$, we will assume K shares the same eigenfunctions as $\mathcal{L}$. ♦

Note also that we take the boundary value to be zero, i.e. $g = 0$, as discussed in Section 2.2.

Remark 3.1. As part of our error analysis we will consider the two distinct cases: $\lambda \in [0, 1]$ and $\lambda \in (1, \infty)$. Note that in the case $\lambda = 0$ we instead consider f to be white noise, i.e. K is the identity on $L^2(D)$. The convergence of FEM for white-noise forcing has been established in the literature, e.g. [39, 135]. In the case $\lambda > 1$ we have that the covariance operator K is trace-class on $L^2(D)$ which is equivalent to the covariance k_f being regular enough so that realisations

of f lie almost surely in $L^2(D)$. To guarantee this it is sufficient to take $k_f(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$ where $\psi : D \rightarrow \mathbb{R}$ is continuous [136, see Corollary 1]. This is a mild requirement satisfied by many commonly used kernels, for example by the Squared Exponential kernel or any kernel in the Matérn family. Another alternative sufficient condition would be to choose the kernel so that the associated covariance operator is an appropriate negative power of the Laplacian. This is outlined in detail in [146, see section 6]. ♦

In our analysis we shall assume that the conductivity coefficient κ is deterministic, unlike [65] where this is assumed to be stochastic. Since Eqs. (2.27) and (2.28) are well-posed, the solution can be expressed as $u = \mathcal{L}^{-1}f$, where $\mathcal{L}^{-1}$ is the inverse of the differential operator $\mathcal{L}$. The pushforward of the probability measure for f through $\mathcal{L}^{-1}$ induces a probability distribution of solutions supported on $H_0^1(D)$. Since the inverse operator is linear, the resulting distribution for u will also be a Gaussian measure supported on $L^2(D)$, with mean $\mathcal{L}^{-1}\bar{f}$ and covariance operator $\mathcal{L}^{-1}K\mathcal{L}^{-\dagger}$. Indeed, u can be expressed as a Gaussian process [114, Proposition 3.1] (see Proposition 2.3),

$$u \sim \mathcal{GP}(\mathcal{L}^{-1}\bar{f}, k_{\mathcal{L}}), \quad (3.1)$$

where

$$k_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \mathcal{L}_{\mathbf{x}}^{-1}\mathcal{L}_{\mathbf{y}}^{-1}k_f(\mathbf{x}, \mathbf{y}). \quad (3.2)$$

Here $\mathcal{L}_{\mathbf{x}}^{-1}$ denotes the action of the operator $\mathcal{L}^{-1}$ on the first argument of the kernel and $\mathcal{L}_{\mathbf{y}}^{-1}$ its action on the second argument.

We intend to use Eq. (3.1) as a prior, which will later be combined with sensor observations through sequential Bayesian updates. An important observation is that the mean and covariance of the GP defined in Eq. (3.1) cannot be computed exactly for general problems, which precludes the calculation of marginal distributions or realisations of the GP. To overcome this issue, another GP, which employs a finite element approximation of the mean and covariance functions will be introduced.

As discussed in Section 2.2.2 FEM approximates the solution to our elliptic BVP as a linear combination of the FE basis functions. In particular, the FEM approximation to the solution of our elliptic BVP can be expressed as follows:

$$u_h = \Phi^T \mathbf{A}^{-1} \mathbf{f} =: \hat{L}_h^{-1} f, \quad (3.3)$$

where the operator $\hat{L}_h^{-1} : L^2(D) \rightarrow L^2(D)$ is defined as $\hat{L}_h^{-1} := \Phi^T \mathbf{A}^{-1} \mathcal{F}_h$. Here the operator $\mathcal{F}_h : L^2(D) \rightarrow \mathbb{R}^{n_u}$ takes a function f to the load vector $\mathbf{f}$, defined in Eq. (2.47), and $\mathbf{A}$ is the stiffness matrix defined in Eq. (2.46). Note that we have denoted by $\Phi := (\varphi_1, \dots, \varphi_{n_u})^T$ the vector

of FE basis functions. The StatFEM approximation to (3.1) is obtained by formally replacing the solution operator $\mathcal{L}^{-1}$ with $\hat{\mathcal{L}}_h^{-1}$ in the mean and covariance terms. Due to the linearity of the approximation solution map, the resulting process is still a Gaussian process, with transformed mean and covariance.

Proposition 3.1. *The StatFEM model from [65] is a GP of the form*

$$u_h \sim \mathcal{GP}(\bar{u}_h, k_h), \quad (3.4)$$

where the mean and covariance function are defined by:

$$\bar{u}_h(\mathbf{x}) := \Phi(\mathbf{x})^T \mathbf{A}^{-1} \bar{\mathbf{f}}, \quad (3.5)$$

$$k_h(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x})^T \mathbf{A}^{-1} \mathbf{K}_{\mathcal{F}_h} \mathbf{A}^{-1} \Phi(\mathbf{y}), \quad (3.6)$$

where $\mathbf{K}_{\mathcal{F}_h} := \mathcal{F}_h K \mathcal{F}_h^\dagger \in \mathbb{R}^{n_u \times n_u}$ and $\bar{\mathbf{f}} := \mathcal{F}_h \bar{f} \in \mathbb{R}^{n_u}$. $\blacklozenge$

Proof of Proposition 3.1.

This follows trivially from the fact that the approximate solution operator is a bounded linear operator together with the well known transformation results of a Gaussian measure by linear maps, see Proposition 2.3. $\blacksquare$

Remark 3.2. Note that both the true and approximate FEM solution operators can be uniquely extend to operators on $H^{-1}(D)$. This fact will be useful when considering rougher forcing terms.

$\blacklozenge$

Next we quantify how much error is introduced by replacing u with u_h by bounding the distance between their laws. To this end, we define $\nu_\star$ and ν_h to be the Gaussian probability measures supported on $L^2(D)$ which are associated with the GPs (3.1) and (3.4), respectively. We can write $\nu_\star = \mathcal{N}(m_\star, \Sigma_\star)$ and $\nu_h = \mathcal{N}(m_h, \Sigma_h)$ where

$$\begin{aligned} m_\star &= \mathcal{L}^{-1} \bar{f}, \\ \Sigma_\star &= \mathcal{L}^{-1} K \mathcal{L}^{-\dagger}, \end{aligned}$$

and

$$\begin{aligned} m_h &= \hat{\mathcal{L}}_h^{-1} \bar{f}, \\ \Sigma_h &= \hat{\mathcal{L}}_h^{-1} K \hat{\mathcal{L}}_h^{-\dagger}. \end{aligned}$$

The next theorem shows that $\nu_\star, \nu_h$ are mutually singular, justifying our use of Wasserstein as

a distance measure.

Theorem 3.1. *The measures $\nu_\star$ and ν_h are mutually singular. $\blacklozenge$*

Proof. Let $V_\star, V_h$ be the supports of $\nu_\star, \nu_h$ respectively. Since the image of the approximate solution operator $\hat{L}_h^{-1}$ is the span of the FE basis functions $\{\varphi_i\}_{i=1}^{n_u}$ it follows that, for all $h > 0$, V_h is a finite dimensional subspace of $H_0^1(D) \subseteq L^2(D)$. Further, under Assumptions 2.1 and 2.4 we have that $V_\star$ is an infinite dimensional subspace of $H_0^1(D) \subseteq L^2(D)$. In the case that V_h and $V_\star$ are disjoint we clearly have that $\nu_\star$ and ν_h are mutually singular. Consider now the case that V_h and $V_\star$ are not disjoint and $V_h \not\subseteq V_\star$. This means that V_h and $V_\star$ have a non-empty intersection, and that there exists a subset $\tilde{V} \subseteq V_\star$ not in the intersection. Further, ν_h gives 0 measure to $\tilde{V}$, while $\tilde{V}$ has positive measure under $\nu_\star$. This implies that $\nu_\star$ and ν_h are mutually singular. Finally, in the case that $V_h \subset V_\star$ we can write $V_\star$ as the orthogonal direct sum of V_h and its complement. Since ν_h gives full measure to V_h it must give measure 0 to the orthogonal complement of V_h , which has positive measure under $\nu_\star$. Hence, we again have that $\nu_\star$ and ν_h are mutually singular. In all cases, the mutual singularity follows from the Feldman-Hájek theorem (see Theorem A.3) [123, Theorem 2.25]. $\blacksquare$

3.3.2 Posterior Updates of the StatFEM approximation

We now consider the problem of inferring the solution u of the BVP from Eqs. (2.27) and (2.28) for an unknown forcing term f based on noisy observations of the form

$$\mathbf{v} = Su + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(0, \epsilon^2 \mathbf{I}), \quad (3.7)$$

where S is a linear observation map. Following the Bayesian formalism [146], the solution of this inverse problem is characterised by the posterior distribution $\tilde{\nu}_\star(du)$ of u given the observations $\mathbf{v}$,

$$\tilde{\nu}_\star(du) \propto e^{-L(\mathbf{v}; u)} \nu_\star(du), \quad (3.8)$$

where $L(\mathbf{v}; u) = \frac{1}{2\epsilon^2} \|Su - \mathbf{v}\|^2$, is the negative log-likelihood, often referred to as the *potential* in the literature on Bayesian inverse problems, and ϵ is the observational noise standard deviation. In practice, we would work with an alternative posterior distribution, where the prior distribution $\nu_\star$ is replaced with a prior obtained from a finite element approximation. To understand the consequences of making this approximation, we must quantify the distance (in an appropriate sense) between the posterior probability measures (3.8) and

$$\tilde{\nu}_h(du) \propto e^{-L(\mathbf{v}; u)} \nu_h(du). \quad (3.9)$$

We shall make the following assumption, on the observation map S .

Assumption 3.2. The observation map $S : C(D) \rightarrow \mathbb{R}^s$ is a bounded linear operator with full range. ♦

The choice of $C(D)$ as the domain of the observation map S in Assumption 3.2 is justified by the following theorem:

Theorem 3.2. *The supports of $\nu_\star$ and ν_h are subsets of $C(D)$. Note that for this to be true in the case of $\lambda \in [0, 1]$ in Assumption 3.1 we must restrict further to $\lambda \in (0, 1]$ when $d = 2$ and $\lambda \in (2/3, 1]$ when $d = 3$. ♦*

Proof. Note that restricting $\lambda \in (1, \infty)$ in Assumption 3.1 guarantees that solutions to our noisy elliptic BVP lie in $H^2(D)$ almost surely. By the Sobolev Embedding Theorem (see Theorem 2.1) for $d = 1, 2, 3$ one has that $H^2(D) \subset C(D)$ and thus draws from the true prior are almost surely continuous. In the case $\lambda \in [0, 1]$ it can be shown (see discussion following the proof of Lemma B.1 in Appendix B.2) that solutions from the true prior lie almost surely in the fractional Sobolev Space $H^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ for all $\epsilon' > 0$. Using the Embedding Theorem for fractional Sobolev spaces (see Theorem 2.1) one can again show that draws from the true prior are almost surely continuous. Note that in the fractional case when $d = 2$ we must restrict $\lambda \in (0, 1]$ and when $d = 3$ we must restrict $\lambda \in (2/3, 1]$. For further details please see Appendix B.2. Draws from the StatFEM prior are trivially in $C(D)$. ■

Thus, S is well-defined for functions drawn from the true and StatFEM priors. Note that the condition on the range of S will always be true for some $s > 0$.

A property which will be very important for the subsequent error analysis of Section 3.4 is that the posterior distributions $\tilde{\nu}_\star$ and $\tilde{\nu}_h$ remain Gaussian measures. This can be viewed as the infinite dimensional version of the analogous result for multivariate Gaussians based on “completing the square”.

Proposition 3.2. *Suppose that the observation map satisfies Assumption 3.2, k_f satisfies Assumption 3.1, and let $\mathbf{v} \in \mathbb{R}^s$. Then $\tilde{\nu}_\star$ and $\tilde{\nu}_h$ are Gaussian measures with mean and covariance given by*

$$m_{u|\mathbf{v}}^{(a)} := m_a + \Sigma_a S^\dagger (\epsilon^2 \mathbf{I} + S \Sigma_a S^\dagger)^{-1} (\mathbf{v} - S m_a), \quad (3.10)$$

$$\Sigma_{u|\mathbf{v}}^{(a)} := \Sigma_a - \Sigma_a S^\dagger (\epsilon^2 \mathbf{I} + S \Sigma_a S^\dagger)^{-1} S \Sigma_a, \quad (3.11)$$

respectively, for $a \in \{\star, h\}$. ♦

Proof of Proposition 3.2.

We first note that so far we have assumed that the priors $u \sim \nu_i = \mathcal{N}(m_i, \Sigma_i)$ are Gaussian

measures on $L^2(D)$. However, both priors are actually supported on subspaces $V_i \subseteq L^2(D)$. For the purposes of the proof we will be viewing the priors as measures on the respective subspaces V_i . We will then trivially extend the resulting posteriors onto all of $L^2(D)$. When viewed as Gaussian measures on V_i the covariance operators Σ_i are both positive definite.

In order to obtain the posterior $\tilde{\nu}_i$ we will first derive the joint distribution of $(u, \mathbf{v})^T$ when the prior on u is ν_i . Since the observational noise $\boldsymbol{\xi}$ is independent of $\mathbf{u}$ we have that the joint distribution of $(u, \boldsymbol{\xi})^T$ is:

$$\begin{pmatrix} u \\ \boldsymbol{\xi} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} m_i \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_i & 0 \\ 0 & \epsilon^2 \mathbf{I} \end{pmatrix} \right). \quad (3.12)$$

Using the definition Eq. (3.7) of $\mathbf{v}$ we now note that $(u, \mathbf{v})^T$ can be expressed as:

$$\begin{pmatrix} u \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} I & 0 \\ S & \mathbf{I} \end{pmatrix} \begin{pmatrix} u \\ \boldsymbol{\xi} \end{pmatrix}. \quad (3.13)$$

Since this is a linear transformation of $(u, \boldsymbol{\xi})^T$ we have that the joint distribution of $(u, \mathbf{v})^T$ is given by:

$$\begin{pmatrix} u \\ \mathbf{v} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} m_i \\ S m_i \end{pmatrix}, \begin{pmatrix} \Sigma_i & \Sigma_i S^\dagger \\ S \Sigma_i & \epsilon^2 \mathbf{I} + S \Sigma_i S^\dagger \end{pmatrix} \right). \quad (3.14)$$

From this it is simply a matter of conditioning to obtain the posterior distribution of $u|\mathbf{v}$ specified by Eqs. (3.10) and (3.11). This conditioning step is justified under Theorem A.2, in particular, we apply this theorem with $\mathcal{H}_1 = V_i$ and $\mathcal{H}_2 = \mathbb{R}^s$. The corresponding covariance operator $\mathcal{C}$ is given by the covariance operator in Eq. (3.14). The only thing which remains to show is that this covariance operator is positive definite. To that end, note that we can rewrite this operator as:

$$\begin{aligned} \mathcal{C} &= \begin{pmatrix} \Sigma_i & \Sigma_i S^\dagger \\ S \Sigma_i & \epsilon^2 \mathbf{I} + S \Sigma_i S^\dagger \end{pmatrix} = \begin{pmatrix} \Sigma_i & \Sigma_i S^\dagger \\ S \Sigma_i & S \Sigma_i S^\dagger \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \epsilon^2 \mathbf{I} \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} I \\ S \end{pmatrix} \Sigma_i \begin{pmatrix} I & S^\dagger \end{pmatrix}}_{(a)} + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & \epsilon^2 \mathbf{I} \end{pmatrix}}_{(b)}. \end{aligned} \quad (3.15)$$

Written in this form it is clear that $\mathcal{C}$ is positive definite. This is because operator (a) above is positive definite while operator (b) is positive semi-definite. Positive definiteness of (a) follows from the positive definiteness of Σ_i together with the fact that S satisfies Assumption 3.2. ■

3.4 Discretisation Error Analysis

Intuitively we expect that $\nu_h \rightarrow \nu_\star$ as $h \rightarrow 0$ in an appropriate sense because of the convergence properties of the finite element discretisation. Similarly, we expect that the approximate posterior, $\tilde{\nu}_h$, will converge to the true posterior, $\tilde{\nu}_\star$, in an appropriate sense, as $h \rightarrow 0$.

To establish these properties we must select an appropriate metric for the distance between the priors and posteriors. A challenge with selecting this metric is that, as explained in Theorem 3.1, the measures $\{\nu_h, \nu_\star\}$ and $\{\tilde{\nu}_h, \tilde{\nu}_\star\}$ are generally mutually singular. As noted in [142, 61] this complicates the typical stability analysis that arises when studying Bayesian inverse problems, based on Kullback-Leibler divergence or Hellinger distance [38]. In this thesis we shall adopt an alternative approach, employing the Wasserstein-2 metric (see Definition 2.15) to quantify the influence of discretisation error on the posterior distribution. This is advantageous in this context for several reasons: firstly it is not dependent on the use of a common dominating measure. Secondly, unlike other divergences, it is robust under vanishing noise: in this limit, the Wasserstein distance between the two probability measures will converge to the deterministic L^2 discretisation error between the respective two means. Finally, while computing the Wasserstein distance is intractable for general measures, there exists a closed form for the Wasserstein-2 distance between Gaussian measures (see Eq. (2.86)), valid in both finite and infinite dimensions. See [105] for a statement of this and see [47, 113] for proofs in the finite dimensional case, and [40] for a proof in the infinite dimensional case.

The results in this section will focus on bounding the Wasserstein distance between the priors $\nu_\star$ and ν_h , and the posteriors $\tilde{\nu}_\star$ and $\tilde{\nu}_h$ as a function of h . The strategy employed to obtain these bounds involves exploiting an important connection between Wasserstein distance between Gaussian measures on Hilbert spaces and the Procrustes Metric between the respective covariance operators, as detailed in [105]. We now present these results in the following two theorems. The proofs will be given after stating both theorems.

Theorem 3.3 (Prior Error Analysis). *Assuming that $\bar{f} \in L^2(D)$, k_f satisfies Assumption 3.1, and that the domain, D , and parameter, κ , satisfy Assumptions 2.1, 2.4, and 2.5 there exists a constant $\gamma > 0$, independent of h , such that, for sufficiently small h , and for any $\epsilon' > 0$ we have:*

$$W(\nu_\star, \nu_h) \leq \gamma \begin{cases} h^{2-\frac{d}{2}-\epsilon'}, & \text{if } \lambda \in [0, 1] \text{ and } d \in \{1, 2\}, \\ h^2, & \text{if } \lambda \in (1, \infty). \end{cases} \quad (3.16)$$

◆

The next theorem describes how this error in the prior propagates forward to the error between arbitrary linear functionals of the true and StatFEM posterior, again as a function of the mesh

width h .

Theorem 3.4 (Posterior Error Analysis). *Let $\ell : L^2(D) \rightarrow \mathbb{R}$ be a bounded linear functional. Assuming that $\bar{f} \in L^2(D)$, k_f satisfies Assumption 3.1 with the additional restriction that $\lambda \in (0, 1]$ when $d = 2$ and $\lambda \in (2/3, 1]$ when $d = 3$, and that the domain, D , parameter, κ , and the observation map S , satisfy Assumptions 2.1, 2.4, 2.5, and 3.2 there exists a constant $\gamma'(\ell) > 0$, independent of h , such that, for sufficiently small h we have:*

$$W(\ell_{\#}\tilde{\nu}_{\star}, \ell_{\#}\tilde{\nu}_h) \leq \gamma'(\ell)h^2 + \mathcal{O}(h^4) \text{ as } h \rightarrow 0. \quad (3.17)$$

◆

Remark 3.3. Theorem 3.4 provides a bound on $W(\ell_{\#}\tilde{\nu}_{\star}, \ell_{\#}\tilde{\nu}_h)$, for arbitrary ℓ , rather than on $W(\tilde{\nu}_{\star}, \tilde{\nu}_h)$. Naturally a bound on the latter quantity is stronger and more of interest, but deriving such a bound is technically challenging. Nevertheless the empirical results in Chapter 4 suggest that the stronger result may hold, and as a result this will be the focus of future research. ◆

Theorem 3.3 shows that we have control over the Wasserstein distance between the full priors while Theorem 3.4 shows that, in the posterior case, we have control over linear quantities of interest. Nonlinear quantities of interest are also interesting and we will explore whether they are controlled empirically in the numerical experiments in Chapter 4.

Before we provide proofs for Theorems 3.3 and 3.4 we first present a lemma which will be required in the proofs. This lemma is a simple corollary to Theorem 2.5 which places a bound on the operator norm of the difference of the solution operators $\mathcal{L}^{-1}$ and $\hat{\mathcal{L}}_h^{-1}$.

Lemma 3.1. Under the assumptions of Theorem 2.5 we have the following bound for sufficiently small h :

$$\|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{L^2} \leq Ch^2, \quad (3.18)$$

where we have used the notation $\|\cdot\|_{L^2}$ to denote the operator norm for operators from $L^2(D) \rightarrow L^2(D)$. ◆

Proof of Lemma 3.1.

This bound on the operator norm of $\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}$ follows from noting that the error bound Eq. (2.51) given in Theorem 2.5 can be re-expressed as:

$$\|\mathcal{L}^{-1}f - \hat{\mathcal{L}}_h^{-1}f\|_{L^2(D)} = \|(\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1})f\|_{L^2(D)} \leq Ch^2\|f\|_{L^2(D)}.$$

Since the above inequality holds for all $f \in L^2(D)$ we deduce that Eq. (3.18) holds as required. ■

We are now in a position to prove Theorems 3.3 and 3.4. We start by proving Theorem 3.3.

Proof of Theorem 3.3.

Both of the priors $\nu_\star, \nu_h$ are Gaussian measures on the Hilbert space $L^2(D)$. Further, noting that K is a covariance operator, we have that it possesses a symmetric square root $K^{1/2}$, i.e. we have $K = K^{1/2}K^{1/2}$ and $(K^{1/2})^\dagger = K^{1/2}$. Using this fact, we see that we can express the covariance operators of the priors as follows:

$$\begin{aligned}\Sigma_\star &= \mathcal{L}^{-1}K\mathcal{L}^{-\dagger} \\ &= \mathcal{L}^{-1}K^{1/2}K^{1/2}\mathcal{L}^{-\dagger} \\ &= (K^{1/2}\mathcal{L}^{-\dagger})^\dagger(K^{1/2}\mathcal{L}^{-\dagger}) = S_\star^\dagger S_\star,\end{aligned}\tag{3.19}$$

$$\begin{aligned}\Sigma_h &= \hat{L}_h^{-1}K\hat{L}_h^{-\dagger} \\ &= \hat{L}_h^{-1}K^{1/2}K^{1/2}\hat{L}_h^{-\dagger} \\ &= (K^{1/2}\hat{L}_h^{-\dagger})^\dagger(K^{1/2}\hat{L}_h^{-\dagger}) = S_h^\dagger S_h,\end{aligned}\tag{3.20}$$

where $S_\star := K^{1/2}\mathcal{L}^{-\dagger}$ and $S_h := K^{1/2}\hat{L}_h^{-\dagger}$. Thus, our priors $\nu_\star, \nu_h$ satisfy the assumptions of Lemma 2.1 and so utilising this lemma we obtain the following bound on the Wasserstein distance:

$$W^2(\nu_\star, \nu_h) \leq \|m_\star - m_h\|_{L^2(D)}^2 + \|S_\star - S_h\|_{\text{HS}}^2.\tag{3.21}$$

We now proceed to bound each of the terms in Eq. (3.21). Starting with the first term we have, for sufficiently small h , the following:

$$\|m_\star - m_h\|_{L^2(D)} = \|\mathcal{L}^{-1}\bar{f} - \hat{L}_h^{-1}\bar{f}\|_{L^2(D)} \leq Ch^2\|\bar{f}\|_{L^2(D)},\tag{3.22}$$

where we have utilised the error bound from Theorem 2.5, which holds since $\bar{f} \in L^2(D)$ and since all of the Assumptions 2.1, 2.4, and 2.5 hold. For the second term in Eq. (3.21) we will consider the two cases presented in the statement of Theorem 3.3 separately. We start with the second case where $\lambda \in (1, \infty)$ and d is not restricted (i.e. $d \in \{1, 2, 3\}$). In this case we can compute:

$$\|S_\star - S_h\|_{\text{HS}} = \|K^{1/2}\mathcal{L}^{-\dagger} - K^{1/2}\hat{L}_h^{-\dagger}\|_{\text{HS}}\tag{3.23}$$

$$= \|K^{1/2}(\mathcal{L}^{-\dagger} - \hat{L}_h^{-\dagger})\|_{\text{HS}}\tag{3.24}$$

$$\leq \|K^{1/2}\|_{\text{HS}}\|\mathcal{L}^{-\dagger} - \hat{L}_h^{-\dagger}\|_{L^2}\tag{3.25}$$

$$= \|K^{1/2}\|_{\text{HS}}\|\mathcal{L}^{-1} - \hat{L}_h^{-1}\|_{L^2}\tag{3.26}$$

$$\leq Ch^2\|K^{1/2}\|_{\text{HS}}.\tag{3.27}$$

We have utilised several facts in the above which we now detail. In going from Eq. (3.24) to

Eq. (3.25) we have utilised the facts that $\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}$, and hence its adjoint $\mathcal{L}^{-\dagger} - \hat{\mathcal{L}}_h^{-\dagger}$, is a bounded operator and that $K^{1/2}$ is Hilbert-Schmidt. This second fact follows from the facts that K and $K^{1/2}$ are symmetric and that K is positive and trace-class (which follows since $\lambda > 1$) and computing $\|K^{1/2}\|_{\text{HS}} = [\text{tr}(K^{1/2}K^{1/2})]^{1/2} = [\text{tr}(K)]^{1/2} = [\text{tr}((K^2)^{1/2})]^{1/2} = \|K\|_{\text{tr}}^{1/2} < \infty$. Finally, in going from Eq. (3.26) to Eq. (3.27) we used the bound Eq. (3.18) from Lemma 3.1.

We can now combine the bounds on these two terms to obtain for sufficiently small h , the following:

$$W^2(\nu_\star, \nu_h) \leq C^2 h^4 \|\bar{f}\|_{L^2(D)}^2 + C^2 h^4 \|K^{1/2}\|_{\text{HS}}^2 \quad (3.28)$$

$$= C^2 h^4 \left(\|\bar{f}\|_{L^2(D)}^2 + \|K^{1/2}\|_{\text{HS}}^2 \right). \quad (3.29)$$

which implies,

$$W(\nu_\star, \nu_h) \leq C \sqrt{\|\bar{f}\|_{L^2(D)}^2 + \|K^{1/2}\|_{\text{HS}}^2} h^2 = \gamma h^2, \quad (3.30)$$

where $\gamma := C \sqrt{\|\bar{f}\|_{L^2(D)}^2 + \|K^{1/2}\|_{\text{HS}}^2} > 0$ is a positive constant, independent of h , as required.

We now consider the first case where $\lambda \in [0, 1]$ and where d is restricted to $d \in \{1, 2\}$. In this case, since $\lambda \in [0, 1]$, K is no longer trace-class and so we cannot follow the strategy above. In particular, the step going from (3.24) to (3.25) is no longer valid. We instead need to bound the *HS* norm of the difference in solution operators, i.e.

$$\begin{aligned} \|S_\star - S_h\|_{\text{HS}} &= \|K^{1/2} \mathcal{L}^{-\dagger} - K^{1/2} \hat{\mathcal{L}}_h^{-\dagger}\|_{\text{HS}} \\ &\leq \|K^{1/2}\|_{L^2} \|\mathcal{L}^{-\dagger} - \hat{\mathcal{L}}_h^{-\dagger}\|_{\text{HS}} \\ &= \|K^{1/2}\|_{L^2} \|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{\text{HS}}. \end{aligned}$$

Bounds on the HS norm of the difference in solution operators have already been investigated in [39]. In particular, note that the term $\|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{\text{HS}}$ corresponds with the following term from the proof of Proposition 4 in [39]:

$$A_{\mathcal{Z}} := \left(\mathbb{E} \left[\|\mathcal{Z}^1 - \mathcal{Z}_h^1\|_{L^2(D)}^2 \right] \right)^{1/2}, \quad (3.31)$$

(where we have set $\beta = 1, \sigma = 0, q = 2$ in the original term $A_{\mathcal{Z}}$ from [39]). Here $\mathcal{Z}^1$ denotes a Gaussian random field (GRF) which, in the parlance of [39], is *coloured* by the operator $\mathcal{L}^{-1}$, and $\mathcal{Z}_h^1$ is a GRF coloured by the operator $\hat{\mathcal{L}}_h^{-1}$. Note that the covariance operator of a Gaussian random field coloured by an operator T is $C = TT^\dagger$. Thus, $\mathcal{Z}^1, \mathcal{Z}_h^1$ correspond to the centred true and StatFEM priors respectively in the case of white noise forcing. It can be shown that $\mathcal{Z}^1 - \mathcal{Z}_h^1$ is also a GRF which is coloured by $\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}$, as stated in the proof of Proposition 3 from [39].

To see the correspondence between the term $A_{\mathcal{Z}}$ and $\|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{\text{HS}}$ we can now use equation (36) from Proposition 3 in [39] with the GRF $\mathcal{Z}^1 - \mathcal{Z}_h^1$ to compute:

$$\mathbb{E} \left[\|\mathcal{Z}^1 - \mathcal{Z}_h^1\|_{L^2(D)}^2 \right] = \text{tr}((\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1})(\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1})^\dagger) \quad (3.32)$$

$$= \text{tr}((\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1})^\dagger(\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1})) \quad (3.33)$$

$$= \|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{\text{HS}}^2, \quad (3.34)$$

as required. Owing to the similarity in the analysis with the proofs of Proposition 4 and Section 8 in [39] we only include the resulting bound for $\|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{\text{HS}}$ in the proof here. This is justified since the form of the differential operator $\mathcal{L}$ together with the choice of FE basis function and Assumptions 2.1, 2.4, and 2.5 satisfy all the assumptions required by Proposition 4 in [39]. For completeness we include further details verifying this statement in Appendix B.2. Thus, we can conclude, as in the proof of Proposition 4 in [39] that we can bound $\|\mathcal{L}^{-1} - \hat{\mathcal{L}}_h^{-1}\|_{\text{HS}}$ by $\mathcal{O}(h^{2-d/2-\epsilon'})$ (since in our setting $\alpha = 1$ in Proposition 4 of [39] this is the rate we obtain) for any $\epsilon' > 0$ as required. (*Note: the restriction to $d \in \{1, 2\}$ is necessary since we have $\beta = 1$ for the PDE Eq. (2.27) and we are not in the setting $\beta > 1$ as required in Proposition 4 for the bound on $(A_{\mathcal{Z}})$ to hold when $d = 3$.*) $\blacksquare$

Remark 3.4. Note that the bound presented in Theorem 3.3 is in fact sharp. This can be seen by noting from Eq. (2.86) that the squared Wasserstein distance between the two priors $\nu_\star, \nu_h$ contains the term $\|\mathcal{L}^{-1}\bar{f} - \hat{\mathcal{L}}_h^{-1}\bar{f}\|_{L^2(D)}^2$. Since this term expresses the deterministic FEM discretisation error the bound for the Wasserstein distance for the priors cannot be tighter than the rate obtained by the FEM method applied to the deterministic version of our elliptic BVP Eqs. (2.27) and (2.28). $\blacklozenge$

We now prove Theorem 3.4.

Proof of Theorem 3.4.

To start we note that the restriction of λ to the interval $(0, 1]$ when $d = 2$ and $(2/3, 1]$ when $d = 3$ is required as outlined in Theorem 3.2. Since both of the posteriors $\tilde{\nu}_\star, \tilde{\nu}_h$ are Gaussians we have that $\ell_{\#}\tilde{\nu}_\star \sim \mathcal{N}(\ell m_{u|\mathbf{v}}^{(\star)}, \ell \Sigma_{u|\mathbf{v}}^{(\star)} \ell^\dagger)$ and $\ell_{\#}\tilde{\nu}_h \sim \mathcal{N}(\ell m_{u|\mathbf{v}}^{(h)}, \ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger)$ are univariate Gaussians. Thus, we have via Eq. (2.86),

$$W^2(\ell_{\#}\tilde{\nu}_\star, \ell_{\#}\tilde{\nu}_h) = \left| \ell m_{u|\mathbf{v}}^{(\star)} - \ell m_{u|\mathbf{v}}^{(h)} \right|^2 + \left| \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(\star)} \ell^\dagger} - \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger} \right|^2. \quad (3.35)$$

We proceed to bound the two terms in the equation above. The following remark will be very useful throughout this proof.

Remark 3.5. Let $(\mathcal{X}, \|\cdot\|)$ be a normed vector space. Let $a, a_h \in \mathcal{X}$, where a_h depends on some discretisation parameter h . Suppose there are constants $\widetilde{M}, p > 0$ such that we have

$$\|a - a_h\| \leq \widetilde{M}h^p + \mathcal{O}(h^{p+1}) \quad (3.36)$$

as $h \rightarrow 0$. Then we can bound $\|a_h\|$ as follows:

$$\begin{aligned} \|a_h\| &= \|a + (a_h - a)\| \\ &\leq \|a\| + \|a_h - a\| \\ &\leq \|a\| + \widetilde{M}h^p + \mathcal{O}(h^{p+1}) \end{aligned} \quad (3.37)$$

$$\leq \|a\| + \mathcal{O}(h^p). \quad (3.38)$$

Both of the forms Eq. (3.37) and Eq. (3.38) will be useful in what follows. $\blacklozenge$

It will be helpful to first derive a bound on $\|\Sigma_\star - \Sigma_h\|_{L^2}$. We proceed as follows:

$$\begin{aligned} \|\Sigma_\star - \Sigma_h\|_{L^2} &= \|\mathcal{L}^{-1}K\mathcal{L}^{-\dagger} - \hat{L}_h^{-1}K\hat{L}_h^{-\dagger}\|_{L^2} \\ &\leq \|\mathcal{L}^{-1}K(\mathcal{L}^{-1} - \hat{L}_h^{-1})^\dagger\|_{L^2} + \|(\mathcal{L}^{-1} - \hat{L}_h^{-1})K\hat{L}_h^{-\dagger}\|_{L^2} \\ &\leq \|\mathcal{L}^{-1}\|_{L^2}\|K\|_{L^2}\|(\mathcal{L}^{-1} - \hat{L}_h^{-1})^\dagger\|_{L^2} + \|\mathcal{L}^{-1} - \hat{L}_h^{-1}\|_{L^2}\|K\|_{L^2}\|\hat{L}_h^{-\dagger}\|_{L^2} \\ &= (\|\mathcal{L}^{-1}\|_{L^2} + \underbrace{\|\hat{L}_h^{-1}\|_{L^2}}_{(a)})\|K\|_{L^2}\underbrace{\|\mathcal{L}^{-1} - \hat{L}_h^{-1}\|_{L^2}}_{(b)} \\ &\leq (2\|\mathcal{L}^{-1}\|_{L^2} + Ch^2 + \mathcal{O}(h^3))\|K\|_{L^2}Ch^2 \\ &\leq C_1h^2 + \mathcal{O}(h^4), \end{aligned} \quad (3.39)$$

where $C_1 := 2\|\mathcal{L}^{-1}\|_{L^2}\|K\|_{L^2}C > 0$ is a constant independent of h . Note we have utilised Remark 3.5 to bound (a) and Lemma 3.1 to bound (b). We have also made use of the fact that $\|L\|_{\mathcal{X} \rightarrow \mathcal{Y}} = \|L^\dagger\|_{\mathcal{Y} \rightarrow \mathcal{X}}$ for a bounded linear operator between two Hilbert spaces. With this result we can now begin to bound the two terms in Eq. (3.35). Starting with the mean term we have,

$$\begin{aligned} \left| \ell m_{u|\mathbf{v}}^{(\star)} - \ell m_{u|\mathbf{v}}^{(h)} \right| &= \left| \ell(m_{u|\mathbf{v}}^{(\star)} - m_{u|\mathbf{v}}^{(h)}) \right| \\ &\leq \|\ell\|_{L^2(D) \rightarrow \mathbb{R}} \|m_{u|\mathbf{v}}^{(\star)} - m_{u|\mathbf{v}}^{(h)}\|_{L^2(D)}. \end{aligned}$$

Using the definitions of the posterior means we can bound,

$$\begin{aligned}
\|m_{u|\mathbf{v}}^{(\star)} - m_{u|\mathbf{v}}^{(h)}\|_{L^2(D)} &= \|m_\star - m_h + \Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1}(\mathbf{v} - Sm_\star) - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1}(\mathbf{v} - Sm_h)\|_{L^2(D)} \\
&\leq \underbrace{\|m_\star - m_h\|_{L^2(D)}}_{(c)} + \underbrace{\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow L^2(D)}}_{(d)} \|\mathbf{v}\| \\
&\quad + \underbrace{\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} Sm_\star - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} Sm_h\|_{L^2(D)}}_{(e)}, \tag{3.40}
\end{aligned}$$

where we have defined $\mathbf{B}_{\epsilon,i} := (\epsilon^2 \mathbf{I} + S \Sigma_i S^\dagger)$, $i \in \{\star, h\}$. The term (c) was bounded by $Ch^2 \|\bar{f}\|_{L^2(D)}$ in the proof of Theorem 3.3. We now bound terms (d) and (e). For (d) we have:

$$\begin{aligned}
\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow L^2(D)} &\leq \|(\Sigma_\star - \Sigma_h) S^\dagger \mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow L^2(D)} \\
&\quad + \|\Sigma_h S^\dagger (\mathbf{B}_{\epsilon,\star}^{-1} - \mathbf{B}_{\epsilon,h}^{-1})\|_{\mathbb{R}^s \rightarrow L^2(D)} \\
&\leq \|\Sigma_\star - \Sigma_h\|_{L^2} \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \\
&\quad \underbrace{\|\Sigma_h\|_{L^2}}_{(f)} \|S\|_{C(D) \rightarrow \mathbb{R}^s} \underbrace{\|\mathbf{B}_{\epsilon,\star}^{-1} - \mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}}_{(g)}. \tag{3.41}
\end{aligned}$$

Using Remark 3.5 we can bound (f) by $\|\Sigma_\star\|_{L^2} + \mathcal{O}(h^2)$. In order to obtain a bound for (g) we first compute:

$$\begin{aligned}
\|\mathbf{B}_{\epsilon,\star} - \mathbf{B}_{\epsilon,h}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} &= \|S(\Sigma_\star - \Sigma_h) S^\dagger\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \\
&\leq \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 \|\Sigma_\star - \Sigma_h\|_{L^2} \\
&\leq C_1 \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 h^2 + \mathcal{O}(h^4). \tag{3.42}
\end{aligned}$$

Since $\mathbf{B}_{\epsilon,\star}$ and $\mathbf{B}_{\epsilon,h}$ are invertible we have for sufficiently small h that $\|\mathbf{B}_{\epsilon,\star} - \mathbf{B}_{\epsilon,h}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} < 1/\|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}$ since Eq. (3.42) gives us that $\|\mathbf{B}_{\epsilon,\star} - \mathbf{B}_{\epsilon,h}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \leq \mathcal{O}(h^2)$. Thus, we can utilise [66, Corollary 8.2] to deduce that:

$$\|\mathbf{B}_{\epsilon,\star}^{-1} - \mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \leq \frac{\|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}^2 \|\mathbf{B}_{\epsilon,\star} - \mathbf{B}_{\epsilon,h}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}}{1 - \|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \|\mathbf{B}_{\epsilon,\star} - \mathbf{B}_{\epsilon,h}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}} \leq C_2 h^2 + \mathcal{O}(h^4), \tag{3.43}$$

where $C_2 := \|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}^2 C_1 \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 > 0$ is a constant independent of h .

Returning to Eq. (3.41),

$$\begin{aligned}
\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow L^2(D)} &\leq (C_1 h^2 + \mathcal{O}(h^4)) \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \\
&\quad + (\|\Sigma_\star\|_{L^2} + \mathcal{O}(h^2)) \|S\|_{C(D) \rightarrow \mathbb{R}^s} (C_2 h^2 + \mathcal{O}(h^4)) \\
&\leq C_3 h^2 + \mathcal{O}(h^4), \tag{3.44}
\end{aligned}$$

where $C_3 := C_1 \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\mathbf{B}_{\epsilon, \star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} + C_2 \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\Sigma_\star\|_{L^2} > 0$ is a constant independent of h . Now returning to Eq. (3.40), for (e) we have:

$$\begin{aligned}
& \|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} S m_\star - \Sigma_h S^\dagger \mathbf{B}_{\epsilon, h}^{-1} S m_h\|_{L^2(D)} \\
& \leq \|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} S m_\star - \Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} S m_h\|_{L^2(D)} + \|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} S m_h - \Sigma_h S^\dagger \mathbf{B}_{\epsilon, h}^{-1} S m_h\|_{L^2(D)} \\
& = \|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} S (m_\star - m_h)\|_{L^2(D)} + \|(\Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} - \Sigma_h S^\dagger \mathbf{B}_{\epsilon, h}^{-1}) S m_h\|_{L^2(D)} \\
& \leq \|\Sigma_\star\|_{L^2} \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 \|\mathbf{B}_{\epsilon, \star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \|m_\star - m_h\|_{L^2(D)} \\
& \quad + \underbrace{\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon, \star}^{-1} - \Sigma_h S^\dagger \mathbf{B}_{\epsilon, h}^{-1}\|_{\mathbb{R}^s \rightarrow L^2(D)}}_{\text{bound using Eq. (3.44)}} \|S\|_{C(D) \rightarrow \mathbb{R}^s} \underbrace{\|m_h\|_{L^2(D)}}_{\text{Remark 3.5}} \\
& \leq \|\Sigma_\star\|_{L^2} \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 \|\mathbf{B}_{\epsilon, \star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} C h^2 \|\bar{f}\|_{L^2(D)} \\
& \quad + (C_3 h^2 + \mathcal{O}(h^4)) \|S\|_{C(D) \rightarrow \mathbb{R}^s} (C h^2 \|\bar{f}\|_{L^2(D)} + \|m_\star\|_{L^2(D)}) \\
& \leq C_4 h^2 + \mathcal{O}(h^4), \tag{3.45}
\end{aligned}$$

where $C_4 := C \|\Sigma_\star\|_{L^2} \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 \|\mathbf{B}_{\epsilon, \star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} \|\bar{f}\|_{L^2(D)} + C_3 \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|m_\star\|_{L^2(D)} > 0$ is a constant independent of h . Putting the three bounds together we obtain:

$$\begin{aligned}
\|m_{u|\mathbf{v}}^{(\star)} - m_{u|\mathbf{v}}^{(h)}\|_{L^2(D)} & \leq C \|\bar{f}\|_{L^2(D)} h^2 + C_3 \|\mathbf{v}\| h^2 + C_4 h^2 + \mathcal{O}(h^4) \\
& = C_5 h^2 + \mathcal{O}(h^4),
\end{aligned}$$

where $C_5 := C \|\bar{f}\|_{L^2(D)} + C_3 \|\mathbf{v}\| + C_4 > 0$ is a constant independent of h . Thus, the mean term in Eq. (3.35) can be bounded, for sufficiently small h , as:

$$\left| \ell m_{u|\mathbf{v}}^{(\star)} - \ell m_{u|\mathbf{v}}^{(h)} \right| \leq C_5 \|\ell\|_{L^2(D) \rightarrow \mathbb{R}} h^2 + \mathcal{O}(h^4). \tag{3.46}$$

For the covariance term in Eq. (3.35) it will be useful to make use of the following identity which holds for all $x, y > 0$:

$$|\sqrt{x} - \sqrt{y}| = \frac{|x - y|}{\sqrt{x} + \sqrt{y}}. \tag{3.47}$$

This identity follows from writing $x - y = (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y})$.

Using identity Eq. (3.47) we have:

$$\begin{aligned}
\left| \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} - \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger} \right| &= \frac{|\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger - \ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger|}{\sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} + \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger}} \\
&= \frac{|\ell (\Sigma_{u|\mathbf{v}}^{(*)} - \Sigma_{u|\mathbf{v}}^{(h)}) \ell^\dagger|}{\sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} + \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger}} \\
&\leq \frac{\|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2 \|\Sigma_{u|\mathbf{v}}^{(*)} - \Sigma_{u|\mathbf{v}}^{(h)}\|_{L^2}}{\sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} + \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger}}.
\end{aligned}$$

In order to proceed with bounding the above we first bound the term involving the difference in posterior covariances as follows:

$$\begin{aligned}
\|\Sigma_{u|\mathbf{v}}^{(*)} - \Sigma_{u|\mathbf{v}}^{(h)}\|_{L^2} &= \|\Sigma_\star - \Sigma_h - \Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} S \Sigma_\star + \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} S \Sigma_h\|_{L^2} \\
&\leq \|\Sigma_\star - \Sigma_h\|_{L^2} + \|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} S \Sigma_\star - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} S \Sigma_h\|_{L^2}. \tag{3.48}
\end{aligned}$$

The first term above has already been bounded so we focus on the second term and compute:

$$\begin{aligned}
&\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} S \Sigma_\star - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} S \Sigma_h\|_{L^2} \\
&\leq \|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} S \Sigma_\star - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} S \Sigma_\star\|_{L^2} + \|\Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} S \Sigma_\star - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1} S \Sigma_h\|_{L^2} \\
&\leq \underbrace{\|\Sigma_\star S^\dagger \mathbf{B}_{\epsilon,\star}^{-1} - \Sigma_h S^\dagger \mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow L^2(D)}}_{\text{bound using Eq. (3.44)}} \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\Sigma_\star\|_{L^2} \\
&\quad + \underbrace{\|\Sigma_h\|_{L^2}}_{\text{Remark 3.5}} \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 \underbrace{\|\mathbf{B}_{\epsilon,h}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s}}_{\text{Remark 3.5}} \|\Sigma_\star - \Sigma_h\|_{L^2} \\
&\leq C_3 \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\Sigma_\star\|_{L^2} h^2 \\
&\quad + \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 (\|\Sigma_\star\|_{L^2} + \mathcal{O}(h^2)) (\|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} + \mathcal{O}(h^2)) (C_1 h^2 + \mathcal{O}(h^2)) + \mathcal{O}(h^4) \\
&\leq C_6 h^2 + \mathcal{O}(h^4), \tag{3.49}
\end{aligned}$$

where $C_6 := C_3 \|S\|_{C(D) \rightarrow \mathbb{R}^s} \|\Sigma_\star\|_{L^2} + \|S\|_{C(D) \rightarrow \mathbb{R}^s}^2 \|\Sigma_\star\|_{L^2} \|\mathbf{B}_{\epsilon,\star}^{-1}\|_{\mathbb{R}^s \rightarrow \mathbb{R}^s} C_1 > 0$ is a constant independent of h . We can now bound Eq. (3.48) as follows:

$$\|\Sigma_{u|\mathbf{v}}^{(*)} - \Sigma_{u|\mathbf{v}}^{(h)}\|_{L^2} \leq C_7 h^2 + \mathcal{O}(h^4), \tag{3.50}$$

where $C_7 := C_1 + C_6 > 0$ is a constant independent of h . Returning to Eq. (3.35):

$$\begin{aligned}
\left| \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} - \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger} \right| &\leq \frac{\|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2 \|\Sigma_{u|\mathbf{v}}^{(*)} - \Sigma_{u|\mathbf{v}}^{(h)}\|_{L^2}}{\sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} + \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger}} \\
&\leq \frac{\|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2}{\sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger}} \cdot \frac{C_7 h^2 + \mathcal{O}(h^4)}{1 + \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger} / \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger}} \\
&\leq \frac{C_7 \|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2}{\sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger}} \cdot h^2 + \mathcal{O}(h^4) \\
&= C_8(\ell) h^2 + \mathcal{O}(h^4),
\end{aligned} \tag{3.51}$$

where $C_8(\ell) := C_7 \|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2 / \sqrt{\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger} > 0$ is a constant independent of h (but dependent on ℓ). Note that we have used the fact that both of the posterior variance terms $\ell \Sigma_{u|\mathbf{v}}^{(*)} \ell^\dagger$ and $\ell \Sigma_{u|\mathbf{v}}^{(h)} \ell^\dagger$ are strictly greater than zero, to arrive at the penultimate line above, since this implies that the denominator is greater than 1. The positivity of the posterior variances follows since both of the posterior covariances are positive definite due to the sensor noise ϵ being strictly positive. Thus, putting the mean and variance bounds together we obtain:

$$\begin{aligned}
W^2(\ell_\# \tilde{\nu}_\star, \ell_\# \tilde{\nu}_\star) &\leq C_5^2 \|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2 h^4 + C_8(\ell)^2 h^4 + \mathcal{O}(h^6) \\
&\leq (C_5^2 \|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2 + C_8(\ell)^2) h^4 + \mathcal{O}(h^6),
\end{aligned} \tag{3.52}$$

from which it follows that for sufficiently small h we have $W(\ell_\# \tilde{\nu}_\star, \ell_\# \tilde{\nu}_\star) \leq \gamma'(\ell) h^2 + \mathcal{O}(h^4)$, where $\gamma'(\ell) := \sqrt{C_5^2 \|\ell\|_{L^2(D) \rightarrow \mathbb{R}}^2 + C_8(\ell)^2} > 0$ is a constant independent of h as claimed. $\blacksquare$

3.5 Extension to time-dependent problems

In this section we detail our work in extending the error estimates above to the case of parabolic (spatio-temporal) PDEs.

3.5.1 StatFEM for time-dependent problems

In this section we will present the StatFEM model for linear parabolic PDEs. We begin by reviewing FEM for parabolic PDEs. Then we will derive the StatFEM model over the solution of our parabolic PDE. In Section 3.5.2 we detail some results towards deriving rigorous error bounds.

FEM for parabolic PDEs

In this section we will briefly review the additional necessary concepts from FEM which we will require in order to be able to apply StatFEM to parabolic PDEs. To this end consider the following

linear parabolic PDE:

$$\begin{aligned} \frac{\partial u(\mathbf{x}, t)}{\partial t} + (\mathcal{L}u)(\mathbf{x}, t) &= f(\mathbf{x}, t), & (\mathbf{x}, t) \in D \times (0, T], \\ u(\mathbf{x}, t) &= 0, & (\mathbf{x}, t) \in \partial D \times (0, T], \\ u(\mathbf{x}, 0) &= u^{(0)}(\mathbf{x}), & \mathbf{x} \in D, \end{aligned} \quad (3.53)$$

where $\mathcal{L}$ is the elliptic differential operator associated with the elliptic PDE Eq. (2.27) from Section 2.2 but now applied to a function of both space and time, i.e.

$$(\mathcal{L}u)(\mathbf{x}, t) := -\nabla_{\mathbf{x}} \cdot (\boldsymbol{\kappa}(\mathbf{x}) \nabla_{\mathbf{x}} u(\mathbf{x}, t)). \quad (3.54)$$

Note that we have added an $\mathbf{x}$ -subscript to the gradient operator, $\nabla_{\mathbf{x}}$, to denote that this is the gradient with respect to the spatial coordinates only. We also note that the conductivity parameter, $\boldsymbol{\kappa} : D \rightarrow \mathbb{R}^{d \times d}$, is the same as in Eq. (2.27) while in contrast the forcing f is now allowed to be a function of time as well, $f \in L^2(D \times (0, T])$. The initial condition is specified by the function $u^{(0)}(\mathbf{x})$. We call, following [55], the differential operator, $\frac{\partial}{\partial t} + \mathcal{L}$, *uniformly parabolic* if the conductivity parameter $\boldsymbol{\kappa}$ in $\mathcal{L}$ satisfies Assumption 2.4. We assume this is the case throughout this section.

We will write $u(t)$ for $u(\cdot, t)$ and $f(t)$ for $f(\cdot, t)$. Just as in the elliptic case in Section 2.2.2 we consider the variational formulation of the parabolic PDE (3.53), namely we seek, for each $t \in (0, T]$, $u(t) \in H_0^1(D)$ satisfying:

$$\begin{aligned} \left\langle \frac{\partial u}{\partial t}, v \right\rangle + a(u, v) &= \langle f, v \rangle, \quad \forall v \in H_0^1(D), \\ u(0) &= u^{(0)}, \end{aligned} \quad (3.55)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $L^2(D)$.

Standard results [55, see Theorems 3, 4 in section 7.1] show that there is a unique weak solution solving Eq. (3.55). We will now present the Galerkin approximation of the variational problem (3.55). This approximation can be done either in a semidiscrete or in a completely discrete manner. These two approaches will be discussed below.

Semidiscrete Galerkin FEM

The semidiscrete Galerkin approximation to the variational problem (3.55) is to seek, for each $t \in (0, T]$, an approximate solution $u_h(t) \in V_h$, where V_h is a finite dimensional subspace of $H_0^1(D)$. Here V_h is again taken to be a subspace of the space of piecewise polynomial functions defined on the mesh associated with a triangulation $\mathcal{T}_h$ of D as discussed in Section 2.2.2. Thus,

the semidiscrete Galerkin problem is to find for, $t \in (0, T]$, a solution to:

$$\begin{aligned} \left\langle \frac{\partial u_h}{\partial t}, v \right\rangle + a(u_h, v) &= \langle f, v \rangle, \quad \forall v \in V_h, \\ u_h(0) &= u_h^{(0)}, \end{aligned} \quad (3.56)$$

where $u_h^{(0)} \in V_h$ is some approximation of the initial condition $u^{(0)}$. Since we have only discretised in space we refer to this as the *spatially semidiscrete* problem. We can now use the nodal basis $\{\varphi_i\}_{i=1}^{n_u}$, defined in Section 2.2.2, to express the approximate solution as $u_h(t) = \Phi^T \mathbf{u}(t)$ where Φ is the vector of the FE nodal basis functions and $\mathbf{u}(t)$ is the time-dependent vector of degrees of freedom (DOFs). By utilising the linearity of (3.56) we have that $\mathbf{u}(t)$ satisfies the following vector initial value problem (IVP):

$$\begin{aligned} \mathbf{M}\mathbf{u}'(t) + \mathbf{A}\mathbf{u}(t) &= \mathbf{f}(t), \quad t \in (0, T], \\ \mathbf{u}(0) &= \gamma, \end{aligned} \quad (3.57)$$

where $\mathbf{A}, \mathbf{M}$ are the stiffness and mass matrices Eqs. (2.46) and (2.48) from Section 2.2.2 respectively, $\mathbf{f}(t)$ is the time-dependent load vector whose i -th entry is given by $f_i(t) = \langle f(t), \varphi_i \rangle$ and γ denotes the nodal values of $u_h^{(0)}$.

Recall from Section 2.2.2 that the stiffness matrix is invertible; in fact it is also positive definite [89, see Section 5.2]. Similarly the mass matrix is also positive definite (and so invertible) [89, see Section 10.1]. We can thus rewrite the above IVP as follows:

$$\begin{aligned} \mathbf{u}'(t) + \mathbf{M}^{-1}\mathbf{A}\mathbf{u}(t) &= \mathbf{M}^{-1}\mathbf{f}(t), \quad t \in (0, T], \\ \mathbf{u}(0) &= \gamma. \end{aligned} \quad (3.58)$$

This IVP has a unique solution for $t \in (0, T]$ as the following theorem, following the presentation in [89, Section 7.1], shows.

Theorem 3.5. *There exists a unique solution to the following IVP associated with the linear, constant coefficient, system of ODEs:*

$$\mathbf{x}'(t) + \mathbf{B}\mathbf{x}(t) = \mathbf{g}(t), \quad t > 0, \quad (3.59)$$

$$\mathbf{x}(0) = \mathbf{x}_0, \quad (3.60)$$

where here $\mathbf{x} : [0, \infty) \rightarrow \mathbb{R}^d$, $\mathbf{B} \in \mathbb{R}^{d \times d}$, $\mathbf{g} : [0, \infty) \rightarrow \mathbb{R}^d$, and $\mathbf{x}_0 \in \mathbb{R}^d$. ♦

Proof. To solve this equation we must first introduce the exponential of a matrix via the following

power series:

$$e^{\mathbf{B}} := \sum_{n=0}^{\infty} \frac{1}{n!} \mathbf{B}^n, \quad (3.61)$$

where $\mathbf{B}^0 = \mathbf{I}_d$ is the $d \times d$ identity matrix. This definition is based on the Maclaurin series of the regular exponential function and the series in Eq. (3.61) can easily be shown to converge for any $d \times d$ matrix $\mathbf{B}$. We will require the following three properties of this matrix exponential, adopted from [89, Section 7.1]:

- If $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{d \times d}$ commute, i.e. $\mathbf{B}_1 \mathbf{B}_2 = \mathbf{B}_2 \mathbf{B}_1$, then we have:

$$e^{\mathbf{B}_1 + \mathbf{B}_2} = e^{\mathbf{B}_1} e^{\mathbf{B}_2} = e^{\mathbf{B}_2} e^{\mathbf{B}_1}. \quad (3.62)$$

- The time derivative of the matrix $e^{-t\mathbf{B}}$ is given by the following:

$$\frac{d}{dt} e^{-\mathbf{B}t} = \frac{d}{dt} \left(\sum_{n=0}^{\infty} \frac{1}{n!} t^n (-\mathbf{B})^n \right) = \sum_{n=1}^{\infty} \frac{1}{(n-1)!} t^{n-1} (-\mathbf{B})^n = -\mathbf{B} e^{-t\mathbf{B}}. \quad (3.63)$$

- Since any matrix commutes with itself we have the following result for any $t \in \mathbb{R}$:

$$\mathbf{B} e^{t\mathbf{B}} = e^{t\mathbf{B}} \mathbf{B}. \quad (3.64)$$

We can now make progress towards solving the IVP Eqs. (3.59) and (3.60). We start by premultiplying both sides of Eq. (3.59) by the matrix $e^{t\mathbf{B}}$:

$$e^{t\mathbf{B}} (\mathbf{x}'(t) + \mathbf{B}\mathbf{x}(t)) = e^{t\mathbf{B}} \mathbf{g}(t) \quad (3.65)$$

$$e^{t\mathbf{B}} \mathbf{x}'(t) + e^{t\mathbf{B}} \mathbf{B}\mathbf{x}(t) = e^{t\mathbf{B}} \mathbf{g}(t). \quad (3.66)$$

We can now use Eq. (3.64) to rewrite the above as follows:

$$e^{t\mathbf{B}} \mathbf{x}'(t) + \mathbf{B} e^{t\mathbf{B}} \mathbf{x}(t) = e^{t\mathbf{B}} \mathbf{g}(t). \quad (3.67)$$

Using Eq. (3.63) together with product rule of differentiation allows us to further rewrite the above as:

$$(e^{t\mathbf{B}} \mathbf{x}(t))' = e^{t\mathbf{B}} \mathbf{g}(t). \quad (3.68)$$

We can now integrate yielding the following:

$$e^{t\mathbf{B}} \mathbf{x}(t) = \mathbf{x}_0 + \int_0^t e^{s\mathbf{B}} \mathbf{g}(s) ds, \quad (3.69)$$

from which the solution finally follows as:

$$\mathbf{x}(t) = e^{-t\mathbf{B}}\mathbf{x}_0 + \int_0^t e^{-(t-s)\mathbf{B}}\mathbf{g}(s)ds, \quad (3.70)$$

where we have used Eq. (3.62) to simplify the expression above. We thus have established existence of a solution to the IVP Eqs. (3.59) and (3.60). To establish uniqueness suppose $\mathbf{x}_1, \mathbf{x}_2$ are two solutions to the IVP Eqs. (3.59) and (3.60). Let $\mathbf{w}(t) := \mathbf{x}_1(t) - \mathbf{x}_2(t)$ denote their difference. Since both $\mathbf{x}_1, \mathbf{x}_2$ are solutions to our IVP we have that $\mathbf{w}$ satisfies

$$\mathbf{w}'(t) + \mathbf{B}\mathbf{w}(t) = \mathbf{0}, \quad t > 0, \quad (3.71)$$

$$\mathbf{w}(0) = \mathbf{0}. \quad (3.72)$$

This can be converted into the corresponding integral equation:

$$\mathbf{w}(t) = - \int_0^t \mathbf{B}\mathbf{w}(s)ds, \quad (3.73)$$

which implies that the following holds for all $t > 0$:

$$|\mathbf{w}(t)| \leq \|\mathbf{B}\| \int_0^t |\mathbf{w}(s)|ds, \quad (3.74)$$

where $\|\mathbf{B}\|$ is the operator norm of the matrix $\mathbf{B}$. Note that this inequality is of the form

$$|\mathbf{w}(t)| \leq a + \|\mathbf{B}\| \int_0^t |\mathbf{w}(s)|ds, \quad (3.75)$$

with $a = 0$ and so it follows via Grönwall's inequality [69] that $|\mathbf{w}(t)| \leq ae^{\|\mathbf{B}\|t} = 0$ for all $t > 0$ which implies that $\mathbf{w}(t) = \mathbf{0}$ and so $\mathbf{x}_1(t) = \mathbf{x}_2(t)$ establishing uniqueness as required. ■

The above theorem can be applied to the IVP Eq. (3.58) to establish the existence of a unique solution as claimed. We thus obtain the semidiscrete Galerkin FEM approximation as $u_h(t) = \Phi^T \mathbf{u}(t)$ where $\mathbf{u}(t)$ is the unique solution to Eq. (3.58).

Completely discrete schemes

We now move on to discuss some simple schemes for discretising our parabolic problem with respect to time as well as space. We follow the presentation in [89, Chapter 10]. Most such schemes start off with the semidiscrete Galerkin problem (3.56). To this end we now introduce a discretisation $0 = t_0 < t_1 < \dots < t_N = T$ of the time interval $(0, T]$, where $t_n := n\Delta t$ (with $\Delta t = T/N$). We assume that T and Δt are such that N is an integer, though this assumption is not necessary.

Denote by $u_h^{(n)}$ the approximation of $u_h(t)$ at time t_n . Similarly, denote $f^{(n)} := f(t_n)$. We can now obtain completely discrete schemes by discretising the time derivative in (3.56) with different difference quotients. For example, we can utilise a forward difference quotient,

$$\frac{\partial u_h(t_n)}{\partial t} \approx \partial_t u_h^{(n)} := (\Delta t)^{-1}(u_h^{(n+1)} - u_h^{(n)}), \quad (3.76)$$

to obtain the so called **forward Euler-Galerkin method**:

$$\begin{aligned} \langle \partial_t u_h^{(n)}, v \rangle + a(u_h^{(n)}, v) &= \langle f^{(n)}, v \rangle, \quad \forall v \in V_h, \\ u_h^{(0)} &= u_h^{(0)}. \end{aligned} \quad (3.77)$$

In terms of our nodal basis this can be expressed in the following matrix form as:

$$\begin{aligned} \mathbf{M}\mathbf{u}^{(n+1)} &= (\mathbf{M} - (\Delta t)\mathbf{A})\mathbf{u}^{(n)} + (\Delta t)\mathbf{f}^{(n)} \quad n = 0, 1, \dots, N-1 \\ \mathbf{u}^{(0)} &= \boldsymbol{\gamma}, \end{aligned} \quad (3.78)$$

where $\mathbf{u}^{(n)} \approx \mathbf{u}(t_n)$ and $\mathbf{f}^{(n)} := \mathbf{f}(t_n)$. If we instead utilise a backward difference quotient

$$\frac{\partial u_h(t_n)}{\partial t} \approx \bar{\partial}_t u_h^{(n)} := (\Delta t)^{-1}(u_h^{(n)} - u_h^{(n-1)}), \quad (3.79)$$

we instead obtain the so-called **backward Euler-Galerkin method**, which can be written in matrix notation as:

$$\begin{aligned} (\mathbf{M} + (\Delta t)\mathbf{A})\mathbf{u}^{(n)} &= \mathbf{M}\mathbf{u}^{(n-1)} + (\Delta t)\mathbf{f}^{(n)}, \quad n = 0, 1, \dots, N-1, \\ \mathbf{u}^{(0)} &= \boldsymbol{\gamma}. \end{aligned} \quad (3.80)$$

Other completely discrete schemes are possible, such as the popular **Crank-Nicolson-Galerkin method**. In this section we will focus our attention on the **forward Euler-Galerkin method** though nothing prevents us from utilising another completely discrete scheme.

StatFEM model for time-dependent problems

In this section we will present the StatFEM approximation for the solution of a stochastic version of our parabolic PDE. We follow the approach taken in [50, see the supplementary material]. To this end, consider a stochastically-forced version of the parabolic PDE (3.53) where now the forcing f is taken to be random. In particular, let $f = \bar{f} + \xi$ where $\bar{f}$ is the mean of f and ξ is taken to be a zero-mean Gaussian process,

$$\xi \sim \mathcal{GP}(0, k_\xi). \quad (3.81)$$

We will assume, for mathematical simplicity that ξ has a separable covariance function of the form

$$k_\xi((\mathbf{x}, t), (\mathbf{x}', t')) = \delta(t - t')k(\mathbf{x}, \mathbf{x}'), \quad (3.82)$$

i.e. the noise ξ is white in time and spatially regular via $k(\cdot, \cdot)$. We now present the StatFEM model for the forward Euler-Galerkin method.

StatFEM model for Forward Euler-Galerkin Method

Since the forcing f is random, our solution u will also be random. In general, we do not have expressions for how the solution depends on the forcing and as such it is impossible to explicitly compute the resulting distribution of our solution. We will thus turn to an approximate distribution based on a FEM discretisation of the solution as was done for elliptic PDEs in Section 3.3.1. This procedure will now be outlined.

Denote by $\bar{\mathbf{f}}(t)$ the mean of the time-dependent load vector, i.e. the vector with i -th entry given by $\bar{f}_i(t) = \langle \bar{f}, \varphi_i \rangle$. Similarly, denote by $\boldsymbol{\xi}(t)$ the (random) vector with i -th entry $\xi_i(t) = \langle \xi(t), \varphi_i \rangle$. Our assumptions on the covariance structure of ξ yield the following covariance structure for the mean zero vector $\boldsymbol{\xi}(t)$:

$$\text{cov}(\boldsymbol{\xi}(t)) = \mathbb{E} [\boldsymbol{\xi}(t)\boldsymbol{\xi}(t')^T] = \delta(t - t')\mathbf{G}, \quad (3.83)$$

where $\mathbf{G} \in \mathbb{R}^{n_u \times n_u}$ is the matrix with ij -th entry given by:

$$G_{ij} = \int_D \int_D \varphi_i(\mathbf{x})k(\mathbf{x}, \mathbf{x}')\varphi_j(\mathbf{x}')d\mathbf{x}d\mathbf{x}'. \quad (3.84)$$

If we apply the completely discrete Forward Euler-Galerkin method Eq. (3.78) to our stochastically forced parabolic PDE we obtain the following recursion:

$$\begin{aligned} \mathbf{u}^{(n+1)} &= (\mathbf{I}_{n_u} - (\Delta t)\mathbf{M}^{-1}\mathbf{A})\mathbf{u}^{(n)} + (\Delta t)\mathbf{M}^{-1}\bar{\mathbf{f}}^{(n)} + (\Delta t)\mathbf{M}^{-1}\boldsymbol{\xi}^{(n)}, \quad n = 0, 1, \dots, N-1, \\ \mathbf{u}^{(0)} &= \boldsymbol{\gamma}, \end{aligned} \quad (3.85)$$

where $\bar{\mathbf{f}}^{(n)} := \bar{\mathbf{f}}(t_n)$ and $\boldsymbol{\xi}^{(n)} := \boldsymbol{\xi}(t_n)$. Observe that due to the covariance structure of $\boldsymbol{\xi}(t)$ we have that the collection of random vectors $\{\boldsymbol{\xi}^{(n)}\}_{n=0}^{N-1}$ are independent and identically distributed (i.i.d.) multivariate normal, $\boldsymbol{\xi}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$.

This update rule allows us to write down the StatFEM model associated with the forward Euler-Galerkin method, as the following conditional measure at time t_n :

$$\mathbf{u}^{(n)} | \mathbf{u}^{(n-1)} \sim \mathcal{N}((\mathbf{I}_{n_u} - (\Delta t)\mathbf{M}^{-1}\mathbf{A})\mathbf{u}^{(n-1)} + (\Delta t)\mathbf{M}^{-1}\bar{\mathbf{f}}^{(n-1)}, (\Delta t)^2\mathbf{M}^{-1}\mathbf{G}\mathbf{M}^{-T}). \quad (3.86)$$

Since our approximate FEM solutions at time t_n are given by $u_h^{(n)} = \Phi^T \mathbf{u}^{(n)}$ we see that the means of the conditional distributions (3.86) correspond to applying the forward Euler-Galerkin method to our parabolic PDE with the forcing taken to be *just* the mean $\bar{f}$. This observation will allow us to start working on proving some error estimates for the time-dependent case.

3.5.2 Discretisation Error Analysis

We can now present a result towards deriving error estimates for the time-dependent case. To this end let us suppose that we have the following bounds for the L^2 error of a general completely discrete scheme for our deterministic parabolic PDE:

Assumption 3.3. Let u be the unique *weak solution* to the parabolic PDE (3.53) and let $u_h^{(n)}$ by the FEM approximation at time t_n resulting from a completely discrete scheme. Assume that $f \in L^2(D \times (0, T])$ and that our domain D is sufficiently regular. We then assume we have the following error bound, for sufficiently small $h, \Delta t$:

$$\|u^{(n)} - u_h^{(n)}\|_{L^2(D)} \leq C_1 h^\alpha + C_2 (\Delta t)^\beta, \quad (3.87)$$

where $C_1, C_2 > 0$ are positive constants independent of $h, \Delta t$ which depend on the true solution u in general and $\alpha, \beta > 0$. ♦

If Assumption 3.3 holds for the Forward Euler-Galerkin scheme then we have that the StatFEM prior mean has the same $h, \Delta t$ error dependence, i.e. the following theorem also holds:

Theorem 3.6. (*Prior mean error analysis*) Let $m^{(n)}, m_h^{(n)}$ be the true and StatFEM mean functions at time t_n . We have the following error bound for sufficiently small $h, \Delta t$:

$$\|m^{(n)} - m_h^{(n)}\|_{L^2(D)} \leq C_3 h^\alpha + C_4 (\Delta t)^\beta, \quad (3.88)$$

where now the constants $C_3, C_4 > 0$ depend on the true solution of the deterministic parabolic PDE with forcing taken to be $\bar{f}$. ♦

Proof. This result follows from our observation that the means for the StatFEM prior correspond to applying the completely discrete scheme to the parabolic PDE with forcing taken to be $\bar{f}$. This allows us to utilise the bound in Assumption 3.3 with constants specific to the solution of our parabolic PDE with forcing taken to be $\bar{f}$ as required which yields:

$$\|m^{(n)} - m_h^{(n)}\|_{L^2(D)} \leq C_3 h^\alpha + C_2 (\Delta t)^\beta, \quad (3.89)$$

as required. ■

3.5.3 Future work for the time-dependent case

We now comment on potential avenues for future work in light of the result Theorem 3.6:

- We need to establish that a bound of the form given in Assumption 3.3 holds for the Forward Euler-Galerkin scheme. Such an error bound holds for the Backward Euler-Galerkin and Crank-Nicolson-Galerkin schemes [89, see section 10.2]. We have also been deliberately vague about the required regularity for these results to hold. Tying this down is one potential avenue for future research.
- It would be desirable to find error bounds for the time-dependent case which do not depend on the true solution.
- We are also interested in extending this analysis to the error between the covariance terms. This was something we attempted but which proved to be more challenging.
- If bounds on the covariance terms can be established it will be interesting to see if we can prove an analogue of Theorem 3.3 for the time-dependent case for the Wasserstein-2 distance or other distances e.g. the Wasserstein-1 distance.
- If the error analysis above is not possible another potential avenue would be to investigate numerical simulations detailing that such results hold.
- Once we investigate the error for the StatFEM prior we can also consider incorporating noisy solution observations and deriving error bounds for a Kalman filter like scheme as detailed in [50].

3.6 Discussion

In this chapter we have provided detailed error analysis for the StatFEM method first introduced in [65] for the case of spatial elliptic BVPs. In particular, we have proven that the StatFEM prior converges, with respect to the Wasserstein-2 distance, to the true prior at a rate of $\mathcal{O}(h^2)/\mathcal{O}(h^{2-\frac{d}{2}-\epsilon'})$ in the distinct cases where the forcing covariance k_f satisfies Assumption 3.1 with $\lambda \in (1, \infty)/\lambda \in [0, 1]$. We also showed that given noisy observations of linear quantities of interest, the resulting posterior approximation is also controlled at a rate of $\mathcal{O}(h^2)$. In both settings, the choice of the Wasserstein distance provides a convenient metric in which to quantify the error introduced through discretisation. We also presented some work towards detailed error analysis for the StatFEM method for parabolic PDEs. This work, while still at an early stage, is one potential avenue for future work as detailed in Section 3.5.3. Several other possible avenues for future work present themselves. The first would be to prove error bounds for the full posterior distributions,

extending Theorem 3.4. In a similar vein, it would be interesting to look into generalising these error bounds to non-linear problems or problems driven by non-Gaussian noise. Another avenue which would be interesting is to see if similar error bounds can be proven for distributions utilising approximations other than FEM. Generalising the results to utilise Wasserstein- p distance, for $p > 2$, would also be an interesting goal, as this will allow one to have even more control over other aspects of the distribution such as quantiles and extrema, which is especially important for risk-assessment applications.

In the next chapter we present the numerical experiments illustrating the theoretical results for the case of spatial elliptic BVPs.

Chapter 4

StatFEM Error Analysis

Numerical Experiments

4.1 Introduction

In this chapter we will present some numerical experiments demonstrating the theoretical results from Section 3.4 for the case of spatial elliptic BVPs. In Section 4.2 we will consider a one-dimensional example in which we have access to an explicit formula for the Green's function, while in Section 4.3 we will consider a more challenging two-dimensional problem in which the solution operator is not available. In both scenarios we will aim to demonstrate the StatFEM prior and posterior bounds. Note that while Theorem 3.4 shows only that $W(\ell_{\#}\tilde{\nu}_{\star}, \ell_{\#}\tilde{\nu}_h)$ is $\mathcal{O}(h^2)$ our experiments suggest that for the problems examined here it also holds that $W(\tilde{\nu}_{\star}, \tilde{\nu}_h)$ is $\mathcal{O}(h^2)$. Finally, in Section 4.4 we explore whether our convergence results extend beyond the linear setting by calculating the Wasserstein distance between the prior and posterior distributions of a *nonlinear* quantity of interest, given by $\max_{x \in D} u(x)$. The code for these numerical experiments is available online¹. Additional implementation details for the experiments are provided in Appendix C.1.

Remark 4.1. Note that in the prior case we will only demonstrate the bounds for the case when K is trace-class, i.e. the case that $\lambda \in (1, \infty)$ in Assumption 3.1. ♦

¹https://github.com/YanniPapandreou/statFEM_analysis

4.2 One-dimensional Poisson Equation

In this section we will present a numerical experiment involving the following one-dimensional Poisson equation:

$$\begin{aligned} -\frac{d^2 u}{dx^2}(x) &= f(x), \quad x \in [0, 1], \\ u(0) &= u(1) = 0, \end{aligned} \tag{4.1}$$

where the forcing f is distributed as, $f \sim \mathcal{GP}(\bar{f}, k_f)$, with mean and covariance function given by:

$$\bar{f}(x) = 1, \tag{4.2}$$

$$k_f(x, y) = \sigma_f^2 \exp(-\ell_f |x - y|). \tag{4.3}$$

We take $\sigma_f = 0.1$ and $\ell_f = 4$. Note that the BVP from Eq. (4.1) is an instance of the general problem Eqs. (2.27) and (2.28) with $d = 1$, $D = [0, 1]$, and $\kappa(x) \equiv 1$. The Green's function, $G(x, y)$, for the BVP from Eq. (4.1) is easily computed to be:

$$G(x, y) = x(1 - y)\Theta(y - x) + (1 - x)y\Theta(x - y), \quad \forall x, y \in [0, 1], \tag{4.4}$$

where $\Theta(x)$ is the Heaviside Step function. As discussed in Section 3.3.1 the solution u to problem Eq. (4.1) will be the Gaussian process defined by Eq. (3.1). Since we readily have the Green's function available for our 1D problem we can explicitly write down this distribution as:

$$u \sim \mathcal{N}(\mu_\star, k_\star), \tag{4.5}$$

where the mean and covariance functions are given by:

$$\mu_\star(x) = \int_0^1 G(x, w) \bar{f}(w) dw = \frac{1}{2}x(1 - x), \tag{4.6}$$

$$k_\star(x, y) = \int_0^1 \int_0^1 G(x, w) k_f(w, t) G(t, y) dt dw. \tag{4.7}$$

Note that we have explicitly computed the integral for the mean $\mu_\star$, but not for the covariance $k_\star$. The double integral for the covariance is approximated numerically using quadrature to accurately compute this on a reference grid.

4.2.1 Prior results for 1-D example

Since we can straightforwardly evaluate the true prior numerically, we can now demonstrate the error bound for the StatFEM prior given in Theorem 3.3. In order to do this the following remark, taken from [131], will be useful.

Remark 4.2. When we have access to the true distributions we can estimate the rate of convergence of the StatFEM priors and posteriors as follows. Let $\eta_\star, \eta_h$ denote the true and StatFEM prior (or posterior) respectively. If we assume that we have a rate of convergence p , i.e.

$$W(\eta_\star, \eta_h) \leq \mathcal{O}(h^p) \text{ as } h \rightarrow 0, \quad (4.8)$$

then we have that there is a constant $C > 0$, independent of h , such that,

$$W(\eta_\star, \eta_h) = Ch^p + \mathcal{O}(h^{p+1}), \quad (4.9)$$

from which it follows that,

$$\log W(\eta_\star, \eta_h) = p \log h + \log |C| + \mathcal{O}(h). \quad (4.10)$$

From this we can see that we can estimate the rate of convergence, p , when we know the true distributions by computing the Wasserstein distance for a range of sufficiently small h -values, plotting a log-log plot of these values, and estimating the slope of the line of best fit. $\blacklozenge$

As outlined in Remark 4.2 above we compute, for a range of sufficiently small h -values, an approximation of the Wasserstein distance between the two priors and then plot these results on a log-log scale. We take a reference grid of $N = 51$ equally spaced points in $[0, 1]$ to compare the covariances, and we take 30 h -values in $[0.02, 0.25]$. With these choices we obtain the plot shown in Fig. 4.1a. From this plot we can see that the results indeed lie along a straight line with a slope estimated to be $p = 1.99$ (to 2 decimal places) using `scipy`'s built in linear regression function.

4.2.2 Posterior results for 1-D example

For this example we will consider a specific case of the general Bayesian inverse problem outlined in Section 3.3.2 which will correspond to taking the observation operator S to be a pointwise evaluation operator. More specifically, we suppose that we have sensor data which correspond to noisy observations of the value of the solution u at some points $\{x_i\}_{i=1}^s \subset D$. We thus take S to be the operator which maps a function g to the vector $Sg := (g(x_1), \dots, g(x_s))^T \in \mathbb{R}^s$. For this choice of S the update equations for the posterior means and covariances given in Eqs. (3.10) and (3.11) respectively can be simplified to a form which is more amenable to computation. These forms, for

i being either the symbol $\star$ or h , are:

$$m_{u|\mathbf{v}}^{(i)}(x) = m_i(x) - \mathbf{k}^{(i)}(x)^T B_{\epsilon,i}^{-1}(\mathbf{m}^{(i)} - \mathbf{v}), \quad (4.11)$$

$$k_{u|\mathbf{v}}^{(i)}(x, y) = k_i(x, y) - \mathbf{k}^{(i)}(x)^T B_{\epsilon,i}^{-1} \mathbf{k}^{(i)}(y), \quad (4.12)$$

$$B_{\epsilon,i} := \epsilon^2 I + S \Sigma_i S^\dagger = \epsilon^2 I + C_{Y,i}, \quad (4.13)$$

$$\mathbf{k}^{(i)}(x) := (k_i(x, x_1), \dots, k_i(x, x_s))^T, \quad (4.14)$$

where $k_{u|\mathbf{v}}^{(i)}$ is the covariance function associated with $\Sigma_{u|\mathbf{v}}^{(i)}$ and where the matrix $C_{Y,i} := S \Sigma_i S^\dagger \in \mathbb{R}^{s \times s}$ has pq -th entry $k_i(x_p, x_q)$. The specific sensor observations used in the experiments were obtained by simulating a trajectory from $\nu_\star$ evaluated at the locations $x_1, \dots, x_s$.

Since we have the true posterior in an explicit form we can now demonstrate that the error bound for the full StatFEM posterior agrees with the error bound for linear quantities of interest from the posterior given in Theorem 3.4. We demonstrate this error bound for 4 different levels of sensor noise, $\epsilon = 0.00005, 0.0001, 0.01, 0.1$. These levels are chosen so that the lower values are comparable to the prior variances. We take $s = 10$ sensor readings equally spaced in the interval $[0.01, 0.99] \subset D$. For each level of sensor noise we again follow the argument outlined in Remark 4.2 and compute, for a range of sufficiently small h -values, an approximation of the Wasserstein distance between the two posteriors, and plot these results on a log-log scale. We take a reference grid of $N = 41$ equally spaced points in $[0, 1]$ to compare the covariances, and we take 28 h -values in $[0.025, 0.25]$. With these choices we obtain the plot shown in Fig. 4.1b. We also present the estimates for the slopes and intercepts of the lines of best fit in Table 4.1.

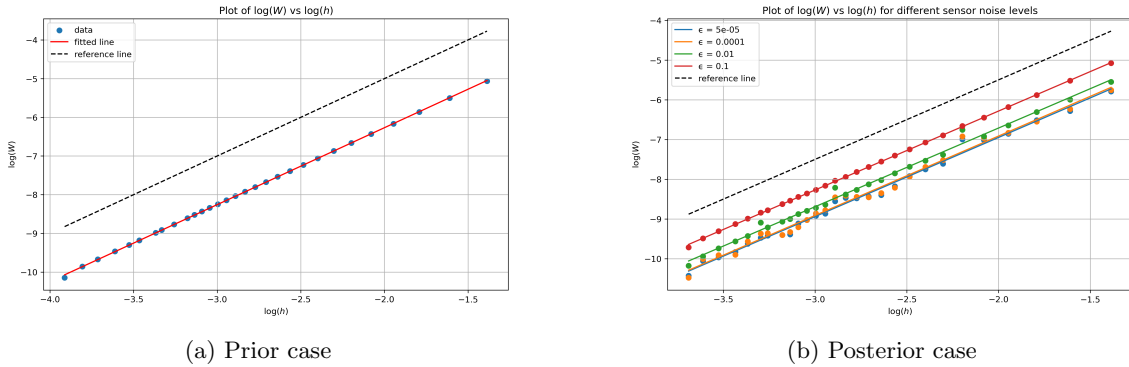


Figure 4.1: 1-D prior and posterior results. Fig. 4.1a shows a log-log plot of Wasserstein distance against h for the priors. The slope of the fitted line is estimated to be $p = 1.99$, in agreement visually with the reference line. Fig. 4.1b shows a log-log plot of Wasserstein distance against h for the posteriors for varying levels of sensor noise. In both sub-figures, the reference line has a slope of 2. Notice how the best fit lines are all parallel to the reference line.

We can see from Fig. 4.1b and Table 4.1 that we obtain lines of best-fit which are all parallel to each other, each with a slope of around $p = 2$. The intercepts are slightly different for each level of sensor noise, as is to be expected since the constant in Theorem 3.4 depends upon ϵ . These results

Table 4.1: Estimate of slopes and intercepts for the different levels of sensor noise ϵ . All of the slopes are around 2.

ϵ	slope	intercept
0.00005	1.9903	-2.9681
0.00010	1.9976	-2.9178
0.01000	1.9796	-2.7523
0.10000	1.9859	-2.3126

show that the error bound for the full posterior distributions is also of order $p = 2$, suggesting that it is possible to extend Theorem 3.4 to the Wasserstein distance between the full posteriors. Proving this is left to future research.

4.3 Two-dimensional Poisson Equation

In this section we will present a numerical example involving the following two-dimensional Poisson equation:

$$\begin{aligned} -\Delta u(\mathbf{x}) &= f(\mathbf{x}), & \mathbf{x} \in D &:= [0, 1]^2, \\ u(\mathbf{x}) &= 0, & \mathbf{x} \in \partial D, \end{aligned} \tag{4.15}$$

where the forcing f is distributed as, $f \sim \mathcal{GP}(\bar{f}, k_f)$, with mean and covariance given by

$$\bar{f}(\mathbf{x}) = 1, \tag{4.16}$$

$$k_f(\mathbf{x}, \mathbf{y}) = \sigma_f^2 \exp(-\ell_f \|\mathbf{x} - \mathbf{y}\|). \tag{4.17}$$

Again we take $\sigma_f = 0.1$ and $\ell_f = 4$. Note that the BVP from Eq. (4.15) is an instance of the general problem Eqs. (2.27) and (2.28) with $d = 2$, $D = [0, 1]^2$, and $\kappa(\mathbf{x}) \equiv 1$. For this problem we use the approach outlined in Remark 4.3 below to demonstrate the error bounds. We present the error estimates for the prior and posterior separately in the following sections.

4.3.1 Prior results for 2-D example

Following the notation of Theorem 3.3 we denote the StatFEM prior by ν_h . Since we do not have access to the true distributions here we will need an alternative approach to that of Remark 4.2. This approach is outlined in the remark, taken from [131], below:

Remark 4.3. When we do not have access to the true distributions we can proceed as follows. Again assuming that we have a rate of convergence, p , as given by Eq. (4.8), we can obtain an estimate as follows. First we utilise the triangle inequality to obtain,

$$W(\eta_h, \eta_{h/2}) \leq W(\eta_h, \eta_\star) + W(\eta_\star, \eta_{h/2}). \tag{4.18}$$

Together with Eq. (4.9) this yields,

$$W(\eta_h, \eta_{h/2}) = Ch^p + C(h/2)^p + \mathcal{O}(h^{p+1}). \quad (4.19)$$

Similarly we have,

$$W(\eta_{h/2}, \eta_{h/4}) = C(h/2)^p + C(h/4)^p + \mathcal{O}(h^{p+1}). \quad (4.20)$$

Dividing the two equations above yields,

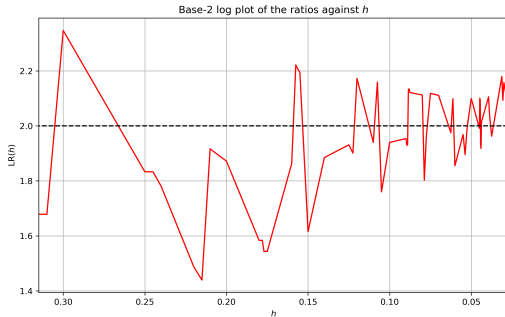
$$\begin{aligned} \frac{W(\eta_h, \eta_{h/2})}{W(\eta_{h/2}, \eta_{h/4})} &= \frac{Ch^p + C(h/2)^p + \mathcal{O}(h^{p+1})}{C(h/2)^p + C(h/4)^p + \mathcal{O}(h^{p+1})} \\ &= \frac{1 + 2^{-p} + \mathcal{O}(h)}{2^{-p} + 2^{-2p} + \mathcal{O}(h)} \\ &= 2^p + \mathcal{O}(h), \end{aligned} \quad (4.21)$$

from which it follows that,

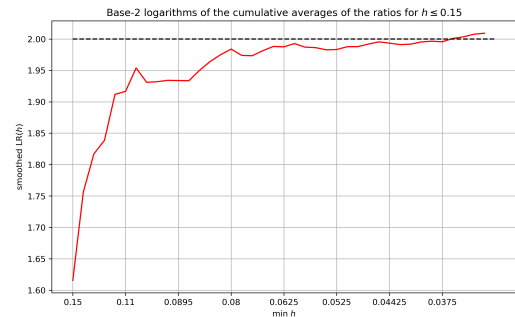
$$\text{LR}(h) := \log_2 \frac{W(\eta_h, \eta_{h/2})}{W(\eta_{h/2}, \eta_{h/4})} = p + \mathcal{O}(h). \quad (4.22)$$

From the above equation we can see that we can obtain an estimate of the rate of convergence by computing, for a range of sufficiently small h -values, the base-2 logarithm of the ratio $W(\eta_h, \eta_{h/2})/W(\eta_{h/2}, \eta_{h/4})$ and estimating the limit as h gets very small. ♦

We take a reference grid of $M = 41^2 = 1681$ equally spaced points in $[0, 1]^2$ to compare the covariances. We compute the logarithms for 59 h -values in the interval $[0.0275, 0.315]$. Note that these h -values are not equally spaced; instead they are arranged according to a refinement strategy which was utilised for efficient use of memory. With these choices we obtain the plot shown in Fig. 4.2a.



(a) Plot of $\text{LR}(h)$ against h .



(b) Plot of smoothed $\text{LR}(h)$ for $h \leq 0.15$.

Figure 4.2: 2-D prior results. Figure 4.2a shows a plot of $\text{LR}(h)$ against h . Figure 4.2b show a plot of smoothed $\text{LR}(h)$ for $h \leq 0.15$. The horizontal axis indicates the smallest h value taken in the cumulative averages.

From this plot we can see that the logarithms seem to be approaching $p = 2$ as $h \rightarrow 0$, as our theory suggests. Note that the points are somewhat oscillatory, though they seem to be oscillating around 2. To illustrate this more we smooth the above results by first discarding the results for larger h -values and then computing a cumulative average of the ratios before applying the base-2 logarithm. Note that this cumulative average is taken by starting with the results for large h first. We take our cutoff point to be $h = 0.15$. These smoothed results are shown in Fig. 4.2b.

From Fig. 4.2b we can see that by discarding the values corresponding to $h > 0.15$ the base-2 logarithms of the rolling averages converge to a value slightly greater than 2. The final logarithm for these rolling averages is $p = 2.01$ (to 2 decimal places). Thus, we obtain results in close agreement with Theorem 3.3.

4.3.2 Posterior results for 2-D example

For this example we will consider the same specific case of the general Bayesian inverse problem which we chose in Section 4.2.2. That is, we will again take the observation operator S to be an evaluation operator. Since we do not have the true prior in an explicit form, we also do not have the true posterior and as such we will utilise the argument in Remark 4.3 to demonstrate the StatFEM posterior error bound. To generate sensor readings, since we do not have access to the true prior, we instead simulated trajectories from the StatFEM prior with a very fine grid of mesh size $h = 1/120$.

We take a reference grid of $M = 41^2 = 1681$ equally spaced points in $[0, 1]^2$ to compare the covariances. We set the sensor noise level to $\epsilon = 0.001$ and take $s = 25$ sensor readings equally spaced in $[0.01, 0.99]^2 \subset D$. The level of sensor noise is chosen to be comparable to the prior variances. We follow the notation of Theorem 3.4 and denote the StatFEM posterior by $\tilde{\nu}_h$. We compute the logarithms $\log_2(W(\tilde{\nu}_h, \tilde{\nu}_{h/2})/W(\tilde{\nu}_{h/2}, \tilde{\nu}_{h/4}))$ for 53 h -values in the interval $[0.03125, 0.315]$, which are spread out according to the refinement strategy mentioned in Section 4.3.1. With these choices we obtain the plot shown in Fig. 4.3a.

From this plot we can see that the logarithms seem to be approaching $p = 2$ as h gets small. However, just as in Section 4.3.1, the results are oscillatory but seem to be oscillating around $p = 2$. We will thus smooth the results as explained in Section 4.3.1. We again take the cutoff point to be $h = 0.15$. These smoothed results are shown in Fig. 4.3b. From this figure we can see that by discarding the values corresponding to $h > 0.15$ the base-2 logarithms of the rolling averages converge to a value slightly less than 2. The final logarithm of these rolling averages is $p = 1.99$ (to 2 decimal places). Thus, just as in the 1D example we see results which suggest that it is possible to extend Theorem 3.4 to the case of the full posteriors.

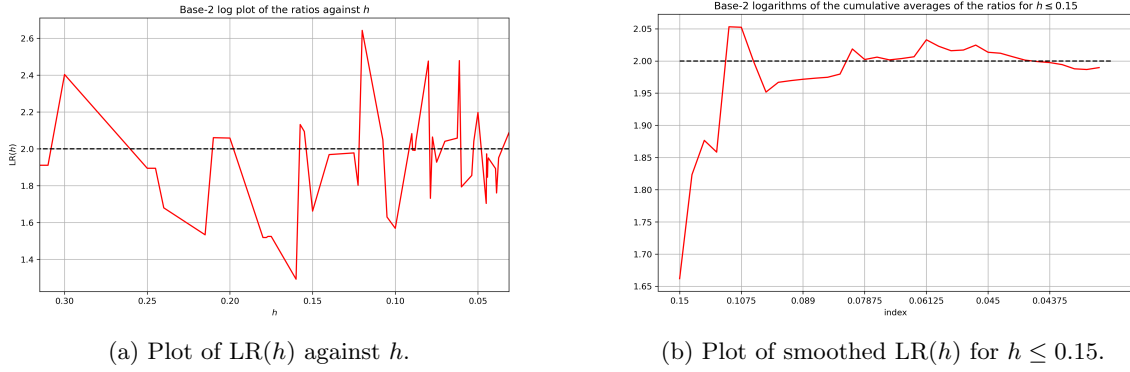


Figure 4.3: 2-D posterior results. Figure 4.3a shows a plot of $LR(h)$ against h . Figure 4.3b shows a plot of smoothed $LR(h)$ for $h \leq 0.15$. The horizontal axis indicates the smallest h value taken in the cumulative averages.

4.4 Distribution of the maximum

In this section we will present a numerical experiment involving the same one-dimensional Poisson equation Eq. (4.1). The forcing f will follow the same distribution as given in Section 4.2. Since the solution, u , is random the maximum, $\max_{x \in [0,1]} u(x)$, is a univariate random variable. Since taking the maximum is not a linear functional we have no guarantees that the distribution of the maximum will be Gaussian. More importantly our results from Theorem 3.4 will not necessarily hold for this example.

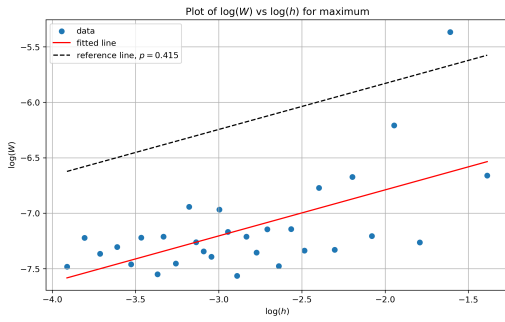
We will investigate the errors for the true and StatFEM prior/posterior distributions for the maximum via a Monte Carlo approach as we now outline in the following remark.

Remark 4.4. Let η_* , η_h denote the true and StatFEM prior (or posterior) for the maximum of u respectively. We are interested in investigating the rate of convergence of $W(\eta_*, \eta_h)$ as $h \rightarrow 0$. This is done as follows: in both the prior and posterior case we will first fix a reference grid $\{\tilde{x}_i\}_{i=1}^N$ and obtain samples of trajectories of the solution on this grid from the true distribution. We then take the maximum values of these trajectories to obtain an (approximate) sample from the true distribution for $\max_{x \in [0,1]} u(x)$. For a range of h -values we will then simulate trajectories of the solution on the same reference grid from the StatFEM distribution and then take the maximum values of these to obtain an (approximate) sample for the StatFEM case. We will then compute an approximation of the Wasserstein distance $W(\eta_*, \eta_h)$ by computing the Wasserstein distance between these two empirical distributions for each h -value by utilising the Python package POT [57]. We then investigate the rate of convergence as outlined previously in Remark 4.2. ♦

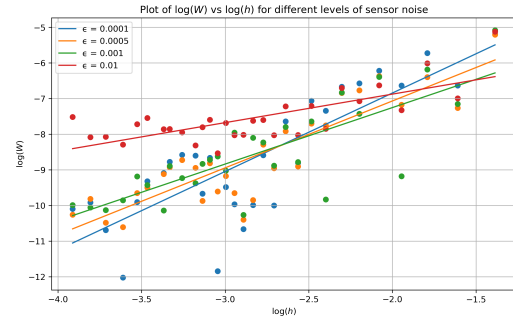
The results for the StatFEM priors and posteriors will be presented separately in the following sections.

4.4.1 Prior results for maximum example

As outlined in Remarks 4.2 and 4.4, we compute, for a range of sufficiently small h -values, a Monte Carlo approximation to the Wasserstein distance between the two priors for the maximum. We take a reference grid of $N = 100$ equally spaced points in $[0, 1]$ and 30 h -values in $[0.02, 0.25]$. We simulate 1000 trajectories from both priors to obtain approximate samples of the maximum. With these choices we obtain the plot shown in Fig. 4.4a. From this plot we can see that we indeed have convergence of the Wasserstein distance for this non-linear quantity of interest to 0, but at a different rate than that given in Theorem 3.3. The slope is estimated to be $p = 0.42$ (to 2 decimal places).



(a) Prior case for maximum



(b) Posterior case for maximum

Figure 4.4: 1-D prior and posterior results for the maximum. Figure 4.4a shows a log-log plot of Wasserstein distance against h for the priors for the maximum. The reference line has a slope of 0.415. The slope of the fitted line is estimated to be $p = 0.415$, in agreement visually with the reference line. Figure 4.4b shows a log-log plot of Wasserstein distance against h for the posteriors for the maximum for different levels of sensor noise.

4.4.2 Posterior results for maximum example

For the posterior case we take a reference grid of $N = 41$ equally spaced points in $[0, 1]$ to simulate 1000 trajectories and 30 h -values in $[0.02, 0.25]$. We take $s = 10$ sensor readings equally spaced in the interval $[0.01, 0.99]$. This is repeated for 4 different levels of sensor noise, $\epsilon = 0.0001, 0.0005, 0.001, 0.01$, again chosen so that the lower noise levels are comparable to the prior variances. For each level of noise we follow the argument outlined in Remarks 4.2 and 4.4. With these choices we obtain the plot shown in Fig. 4.4b. We also present the estimates for the slopes and intercepts of the lines of best fit in Table 4.2.

From Fig. 4.4b and Table 4.2 we can see that for each level of sensor noise we do indeed have convergence to 0 as h goes to 0. However, we observe that the rate is not always the same as that given in Theorem 3.4. From Table 4.2 it is particularly clear that as the sensor noise level decreases the rate of convergence increases. In particular, for $\epsilon = 0.0001, 0.0005$ we have rates around $p = 2$, similar to the case of linear quantities of interest. We hypothesise that this is because as the sensor

Table 4.2: Estimate of slopes and intercepts for the different levels of sensor noise ϵ .

ϵ	slope	intercept
0.00010	2.2008	-2.4444
0.00050	1.8764	-3.3165
0.00100	1.5844	-4.0858
0.01000	0.7999	-5.2773

noise decreases the posterior distribution concentrates more as the solution is better identified and this, combined with the fact that we take sensor readings close to the where the true prior mean Eq. (4.6) has a maximum, forces the maximum of the simulated trajectories to always be at the argmax of the prior mean in the exact case, i.e. at $x = 0.5$. Thus, for very low levels of sensor noise taking the maximum of the posterior draws is almost identical to evaluating the solution at $x = 0.5$; such an evaluation is a linear quantity of interest and so one would expect the results from Theorem 3.4 to hold.

4.5 Discussion

In this chapter numerical experiments demonstrating the theoretical results for the case of spatial elliptic BVPs from Section 3.4 on standard FEM test problems were presented. These demonstrated the rates for the prior theory, while also showing that it may be possible to extend the posterior results to the full distribution as opposed to just linear quantities of interest. Note also that the numerical experiments presented in this chapter demonstrate the sharpness of the error bounds presented in Theorems 3.3 and 3.4 as was noted in Remark 3.4. An example involving a non-linear quantity of interest was also presented, showcasing an example which lies out of the remit of our theory and showing that the rates proved may be limited to linear quantities of interest. One possible avenue for future work would be to perform numerical simulations investigating if any of these results carry over to the case of spatio-temporal problems.

Chapter 5

Background for Probabilistic Modelling of Localised Active Subspaces

In this chapter we will introduce additional relevant and required background knowledge for the second project this thesis will present. We will follow the notation from Chapter 2 unless otherwise stated.

5.1 The Matrix Normal Distribution

Having introduced the multivariate normal distribution in Section 2.3.2 we can now define the *Matrix Normal Distribution* which will be useful in Chapter 6. Before presenting this definition we first introduce the *vectorisation operator*, which we will denote by vec . This operator takes a matrix $\mathbf{M} \in \mathbb{R}^{d \times n}$ and returns a vector in $\mathbb{R}^{dn}$ by stacking the columns of $\mathbf{M}$ on top of each other, i.e. if $\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_n]$ where $\mathbf{m}_i \in \mathbb{R}^d$ for $i = 1, \dots, n$ then,

$$\text{vec}(\mathbf{M}) := \begin{pmatrix} \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_n \end{pmatrix}. \quad (5.1)$$

We will denote the inverse of the vectorisation operator by $\mathbf{v} \mapsto \text{unvec}(\mathbf{v}; d, n)$. Note that in order for this to be well-defined we have specified the dimensions of the matrices involved, though we will typically suppress the dimensions from unvec when they are clear from the context. We can now present the definition of the Matrix Normal Distribution as follows:

Definition 5.1 (*Matrix Normal Distribution*). A random matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ is said to follow a *matrix normal distribution* with parameters $\mathbf{M} \in \mathbb{R}^{d \times n}$, $\mathbf{U} \in \mathbb{R}^{d \times d}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ written

$$\mathbf{X} \sim \mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V}), \quad (5.2)$$

if and only if the vectorisation of $\mathbf{X}$ has the following multivariate normal distribution:

$$\text{vec}(\mathbf{X}) \sim \mathcal{N}(\text{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U}). \quad (5.3)$$

In the case where *both* $\mathbf{U}$ and $\mathbf{V}$ are invertible $\mathbf{X}$ has the following p.d.f.:

$$p(\mathbf{X}; \mathbf{M}, \mathbf{U}, \mathbf{V}) = \frac{1}{(2\pi)^{dn/2} |\mathbf{V}|^{d/2} |\mathbf{U}|^{n/2}} \exp \left(-\frac{1}{2} \text{tr} [\mathbf{V}^{-1} (\mathbf{X} - \mathbf{M})^T \mathbf{U}^{-1} (\mathbf{X} - \mathbf{M})] \right). \quad (5.4)$$

◆

Remark 5.1. As in Remark 2.10 we will occasionally subscript the matrix normal distribution symbol to make explicit the dimension of the random matrix, i.e. we will write $\mathbf{X} \sim \mathcal{MN}_{d,n}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ to denote that $\mathbf{X}$ is a random matrix in $\mathbb{R}^{d \times n}$ following a matrix normal distribution. ◆

Remark 5.2. As in Remark 2.9 we allow the degenerate cases of $\mathbf{U}$ and/or $\mathbf{V}$ to be singular in Definition 5.1. This will imply that the covariance matrix of $\text{vec}(\mathbf{X})$, $\mathbf{V} \otimes \mathbf{U}$, is singular and hence the centred random vector, $\text{vec}(\mathbf{X}) - \text{vec}(\mathbf{M})$, is constrained to lie in some lower-dimensional subspace as discussed in Remark 2.9. ◆

Since the matrix normal distribution is defined in terms of the multivariate normal distribution a similar transformation result analogous to Proposition A.1 holds for it:

Proposition 5.1. *If $\mathbf{X} \sim \mathcal{MN}_{d,n}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ and $\mathbf{A} \in \mathbb{R}^{m \times d}$, $\mathbf{B} \in \mathbb{R}^{n,k}$, $\mathbf{C} \in \mathbb{R}^{m \times k}$ are fixed deterministic matrices then,*

$$\mathbf{Y} = \mathbf{AXB} + \mathbf{C} \sim \mathcal{MN}_{m,k}(\mathbf{AMB} + \mathbf{C}, \mathbf{AUA}^T, \mathbf{B}^T \mathbf{VB}). \quad (5.5)$$

◆

For a proof of this please consult Appendix B.3.

The LKJ distribution

In Chapter 6 we will have need of a distribution which can model correlation matrices. The Lewandowski-Kurowicka-Joe (LKJ) distribution [94, 63] is one such distribution, and is a popular choice as a hierarchical prior for correlation matrices in Bayesian modelling. This distribution is defined below:

Definition 5.2 (LKJ Distribution). A random matrix $\mathbf{L} \in \mathbb{R}^{d \times d}$ is said to follow a *Lewandowski-Kurowicka-Joe (LKJ)* distribution with shape parameter $\eta > 0$ if and only if it has the following p.d.f.:

$$p(\mathbf{L}; \eta) = C (\det(\mathbf{L}))^{\eta-1}, \quad (5.6)$$

where C is a normalising constant given by the following expression:

$$C = 2^{\sum_{k=1}^d (2\eta-2+d-k)(d-k)} \prod_{k=1}^{d-1} [B(\eta + (d-k-1)/2, \eta + (d-k-2)/2)]^{d-k}, \quad (5.7)$$

where $B(a, b)$ is the Beta function. For $\eta = 1$, the distribution coincides with the uniform distribution over the space of all correlation matrices. ♦

5.2 Matrix-Valued Gaussian Processes

In this section we will describe a particular class of matrix-valued GPs which will be very useful for the work presented in Chapter 6. Matrix-valued GPs have been studied before, see for example [23, 46], and are sometimes referred to as multivariate Gaussian processes, though this term is usually reserved for vector-valued GPs [18, 107, 6, 141, 116, 155]. We will be concerned here with a specific class of matrix-valued GPs which we define below:

Definition 5.3 (Matrix-valued Gaussian Process). Let $D \subset \mathbb{R}^k$ be an open subset. For a given positive definite kernel function $k : D \times D \rightarrow \mathbb{R}$, a positive semi-definite *parameter covariance matrix* $\mathbf{\Omega} \in \mathbb{R}^{dn \times dn}$, and a function $\boldsymbol{\mu} : D \rightarrow \mathbb{R}^{d \times n}$ we say that a random matrix-valued function $\mathbf{X} : D \rightarrow \mathbb{R}^{d \times n}$ is a *Matrix-valued Gaussian Process (MatGP)* with mean function $\boldsymbol{\mu}$, kernel function k , and parameter covariance matrix $\mathbf{\Omega}$ if the vectorisation of any arbitrary finite collection of evaluations of $\mathbf{X}$ at given locations has a joint multivariate normal distribution, i.e. if for any $N \in \mathbb{N}$ and for any $S := \{\mathbf{s}_i\}_{i=1}^N \subset D$ we have:

$$\text{vec}(\mathbf{X}_S) \sim \mathcal{N}(\text{vec}(\mathbf{M}_S), \mathbf{K}_{SS} \otimes \mathbf{\Omega}), \quad (5.8)$$

where,

$$\begin{aligned} \mathbf{X}_S &:= [\mathbf{X}(\mathbf{s}_1), \dots, \mathbf{X}(\mathbf{s}_N)] \in \mathbb{R}^{d \times nN}, \\ \mathbf{M}_S &:= [\boldsymbol{\mu}(\mathbf{s}_1), \dots, \boldsymbol{\mu}(\mathbf{s}_N)] \in \mathbb{R}^{d \times nN}, \\ \mathbf{K}_{SS} &\in \mathbb{R}^{N \times N}, (\mathbf{K}_{SS})_{ij} := k(\mathbf{s}_i, \mathbf{s}_j). \end{aligned} \quad (5.9)$$

Note that this is equivalent to stating that the random matrix $\mathbf{X}_S$ has the following matrix normal

distribution (see Definition 5.1):

$$\mathbf{X}_S \sim \mathcal{MN}(\mathbf{M}_S, \mathbf{\Omega}, \mathbf{K}_{SS}). \quad (5.10)$$

The functions $\boldsymbol{\mu}$, k and matrix $\mathbf{\Omega}$ completely determine the matrix-valued GP. We write:

$$\mathbf{X} \sim \text{Mat } \mathcal{GP}(\boldsymbol{\mu}, k, \mathbf{\Omega}), \quad (5.11)$$

to denote the matrix-valued Gaussian Process with mean function $\boldsymbol{\mu}$, kernel function k , and parameter covariance matrix $\mathbf{\Omega}$. ♦

We now present some key properties of matrix-valued GPs in the following remark:

Remark 5.3. Let $\mathbf{X} \sim \text{Mat } \mathcal{GP}(\boldsymbol{\mu}, k, \mathbf{\Omega})$ be a matrix-valued GP. Consider N evaluation locations $\{\mathbf{s}_i\}_{i=1}^N$ and the corresponding kernel matrix $\mathbf{K}_{SS} \in \mathbb{R}^{N \times N}$ with entries $(\mathbf{K}_{SS})_{ij} = k(\mathbf{s}_i, \mathbf{s}_j)$. One can show that the covariance matrix of the random vector $\text{vec}(\mathbf{X}(\mathbf{s}_i))$ is given by:

$$\text{cov}(\text{vec}(\mathbf{X}(\mathbf{s}_i))) = (\mathbf{K}_{SS})_{ii} \mathbf{\Omega} = k(\mathbf{s}_i, \mathbf{s}_i) \mathbf{\Omega}, \quad (5.12)$$

while the covariance between the two random vectors $\text{vec}(\mathbf{X}(\mathbf{s}_i)), \text{vec}(\mathbf{X}(\mathbf{s}_j))$ for $i \neq j$ is given by:

$$\text{cov}(\text{vec}(\mathbf{X}(\mathbf{s}_i)), \text{vec}(\mathbf{X}(\mathbf{s}_j))) = (\mathbf{K}_{SS})_{ij} \mathbf{\Omega} = k(\mathbf{s}_i, \mathbf{s}_j) \mathbf{\Omega}. \quad (5.13)$$

Thus, we can view $\mathbf{\Omega}$ as determining the correlation structure between the $d \times n$ entries of $\mathbf{X}$ while the kernel function k determines the correlation structure across the inputs $\mathbf{s} \in D$.

Similarly the mean of the random matrix $\mathbf{X}(\mathbf{s}_i)$ is given by:

$$\mathbb{E}(\mathbf{X}(\mathbf{s}_i)) = \boldsymbol{\mu}(\mathbf{s}_i). \quad (5.14)$$

♦

Predictions for matrix-valued GPs

In Section 6.6.3 we will need to be able to make predictions from matrix-valued GPs. This can be done in a very similar manner as described in Section 2.3.5 for regular GPs. We now outline this procedure.

Suppose we have a matrix-valued function $\mathbf{X} : D \rightarrow \mathbb{R}^{d \times n}$ upon which we place a matrix-valued GP prior:

$$\mathbf{X} \sim \text{Mat } \mathcal{GP}(\boldsymbol{\mu}, k, \mathbf{\Omega}). \quad (5.15)$$

We assume we make $N \in \mathbb{N}$ potentially noisy observations of $\mathbf{X}$, i.e. we observe:

$$\mathbf{Y}_i = \mathbf{X}(\mathbf{s}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{MN}(\mathbf{0}, \mathbf{I}_d, \sigma^2 \mathbf{I}_n), \quad \sigma \geq 0, \quad (5.16)$$

for $i = 1, \dots, N$ and where the $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ are independent of the prior Eq. (5.15). Note we allow $\sigma = 0$ here to denote the case where we make noiseless observations $\boldsymbol{\epsilon}_i \equiv \mathbf{0}$ which will be the case in some of the experiments in Section 6.7. We will denote the observation locations by $S := \{\mathbf{s}_i\}_{i=1}^N \subset D$, and denote the observations and the noise terms by:

$$\mathbf{Y}_S := [\mathbf{Y}_1, \dots, \mathbf{Y}_N], \quad (5.17)$$

$$\boldsymbol{\epsilon} := [\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N]. \quad (5.18)$$

We wish to make predictions from the posterior $\mathbf{X}|\mathbf{Y}_S$ for the values of our function of interest on a grid of $m \in \mathbb{N}$ new test points, $S^* := \{\mathbf{s}_i^*\}_{i=1}^m \subset D$. Utilising Definitions 5.1 and 5.3 together with standard properties of the multivariate normal distributions and the independence of the $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ both between themselves and with the matrix-valued GP prior, Eq. (5.15), we can deduce:

$$\begin{pmatrix} \text{vec}(\mathbf{X}_{S^*}) \\ \text{vec}(\mathbf{Y}_S) \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \text{vec}(\mathbf{M}_{S^*}) \\ \text{vec}(\mathbf{M}_S) \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{S^*S^*} \otimes \boldsymbol{\Omega} & \mathbf{K}_{S^*S} \otimes \boldsymbol{\Omega} \\ \mathbf{K}_{SS^*} \otimes \boldsymbol{\Omega} & \mathbf{K}_{SS} \otimes \boldsymbol{\Omega} + \sigma^2 \mathbf{I}_{Nnd} \end{pmatrix} \right), \quad (5.19)$$

where we use the notation $\mathbf{M}_{S^*}, \mathbf{K}_{S^*S^*}, \mathbf{K}_{SS^*}, \mathbf{K}_{S^*S}$ similarly as in Eq. (5.9). We can now use Proposition A.4 to deduce that the conditional distribution of $\text{vec}(\mathbf{X}_{S^*})|\text{vec}(\mathbf{Y}_S)$ is:

$$\text{vec}(\mathbf{X}_{S^*})|\text{vec}(\mathbf{Y}_S) \sim \mathcal{N}(\tilde{\mathbf{M}}, \tilde{\boldsymbol{\Sigma}}), \quad (5.20)$$

where,

$$\tilde{\mathbf{M}} := \text{vec}(\mathbf{M}_{S^*}^*) + (\mathbf{K}_{S^*S} \otimes \boldsymbol{\Omega})(\mathbf{K}_{SS} \otimes \boldsymbol{\Omega} + \sigma^2 \mathbf{I}_{Nnd})^-(\text{vec}(\mathbf{Y}_S - \mathbf{M}_S)), \quad (5.21)$$

$$\tilde{\boldsymbol{\Sigma}} =: \mathbf{K}_{S^*S^*} \otimes \boldsymbol{\Omega} - (\mathbf{K}_{S^*S} \otimes \boldsymbol{\Omega})(\mathbf{K}_{SS} \otimes \boldsymbol{\Omega} + \sigma^2 \mathbf{I}_{Nnd})^-(\mathbf{K}_{SS^*} \otimes \boldsymbol{\Omega}), \quad (5.22)$$

where we have kept the pseudo-inverse in the expressions above instead of the regular inverse to allow for the case of $\sigma = 0$ and degenerate kernel matrices. Thus, in order to make predictions for $\mathbf{X}_{S^*}$ from the posterior $\mathbf{X}|\mathbf{Y}_S$ we can sample from the multivariate normal Eq. (5.20) and then undo the vectorisation with the unvec operator.

Remark 5.4. Note that for matrix-valued GPs we have not gone down the route of deriving the functional posterior $\mathbf{X}|\mathbf{Y}_S$ in closed form, instead concentrating on deriving the conditional distributions for point evaluations since we are ultimately interested in simulating predictions from

the posterior. In principle this should be possible albeit with more complicated expressions than in the case of predictions for regular GPs, Proposition 2.4. This will be suitable for our purposes in Chapter 6. ♦

Remark 5.5. Note that the formulae for the posterior mean and covariance, Eqs. (5.21) and (5.22), simplify in the case of noiseless observations, i.e. when $\sigma = 0$. In particular, one obtains:

$$\tilde{\mathbf{M}} := \text{vec}(\mathbf{M}_S^*) + (\mathbf{K}_{S^*S} \mathbf{K}_{SS}^- \otimes \boldsymbol{\Omega} \boldsymbol{\Omega}^-) (\text{vec}(\mathbf{Y}_S - \mathbf{M}_S)), \quad (5.23)$$

$$\tilde{\boldsymbol{\Sigma}} := (\mathbf{K}_{S^*S^*} - \mathbf{K}_{S^*S} \mathbf{K}_{SS}^- \mathbf{K}_{SS^*}) \otimes \boldsymbol{\Omega}. \quad (5.24)$$

This follows from the distributive property of the Kronecker product together with the property that $(\mathbf{A} \otimes \mathbf{B})^- = \mathbf{A}^- \otimes \mathbf{B}^-$. These simplified formulae allow us to more efficiently sample predictions since we can work with matrices of smaller dimension. In particular, we can form the Cholesky decomposition of $\tilde{\boldsymbol{\Sigma}}$ more efficiently for Eq. (5.24) by using the property

$$\text{chol}(\mathbf{A} \otimes \mathbf{B}) = \text{chol}(\mathbf{A}) \otimes \text{chol}(\mathbf{B}), \quad (5.25)$$

where $\text{chol}(\mathbf{A})$ denotes the Cholesky factor of the matrix $\mathbf{A}$, i.e. $\mathbf{A} = \text{chol}(\mathbf{A}) \text{chol}(\mathbf{A})^T$. Note that even in the case of noisy observations we can make some efficiency gains in computing the inverse (since $\sigma \neq 0$ in this case the pseudo-inverse becomes an inverse) in Eqs. (5.21) and (5.22) by utilising the following result from [144] which exploits the compatibility of a Kronecker product plus a constant diagonal term with the eigenvalue decomposition:

$$\mathbf{A} \otimes \mathbf{B} + \sigma^2 \mathbf{I} = (\mathbf{U}_\mathbf{A} \otimes \mathbf{U}_\mathbf{B}) (\boldsymbol{\Lambda}_\mathbf{A} \otimes \boldsymbol{\Lambda}_\mathbf{B} + \sigma^2 \mathbf{I}) (\mathbf{U}_\mathbf{A}^T \otimes \mathbf{U}_\mathbf{B}^T), \quad (5.26)$$

where $\mathbf{A} = \mathbf{U}_\mathbf{A} \boldsymbol{\Lambda}_\mathbf{A} \mathbf{U}_\mathbf{A}^T$ is the eigenvalue decomposition of the symmetric matrix $\mathbf{A}$, and similarly for $\mathbf{B}$. Note that since $\mathbf{A}, \mathbf{B}$ are assumed to be symmetric the matrices $\mathbf{U}_\mathbf{A}, \mathbf{U}_\mathbf{B}$ are both orthogonal. This, together with the fact that the term $\boldsymbol{\Lambda}_\mathbf{A} \otimes \boldsymbol{\Lambda}_\mathbf{B} + \sigma^2 \mathbf{I}$ is diagonal allows us to easily and efficiently form the inverse of Eq. (5.26) provided we first compute the eigendecompositions of $\mathbf{A}$ and $\mathbf{B}$. ♦

5.3 Hamiltonian Monte Carlo

As discussed in Remark 2.11 the HMC algorithm, originally introduced as Hybrid Monte Carlo [48], is a popular and effective method for estimating complex, high-dimensional posterior distributions. In this section we will briefly introduce the HMC algorithm, which we will utilise in Chapter 6. We follow the presentations given in [63, 129, 110].

Metropolis-Hastings Algorithm

To start with we first recall the well-known Metropolis-Hastings (MH) algorithm [108, 71]; the HMC algorithm being a particular variant of the MH algorithm. The MH algorithm, as with any MCMC method, constructs a *Markov Chain* with a unique stationary distribution which coincides with the required posterior. To be more specific, suppose that we wish to estimate the posterior distribution Eq. (2.61) by obtaining samples from it. The MH algorithm constructs a Markov Chain $\{\boldsymbol{\theta}_n\}_{n=1}^{\infty}$ for which as $n \rightarrow \infty$ we have $\boldsymbol{\theta}_n \sim p(\boldsymbol{\theta}|\mathbf{y})$. We can thus obtain samples from $p(\boldsymbol{\theta}|\mathbf{y})$ by running the chain for a sufficiently long time and then taking the required number of samples. Note that we often run the chain for a pre-specified *burn-in* time to ensure that we have reached the stationary distribution. For brevity, we will drop $\mathbf{y}$ and denote by $\pi^\dagger(\boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$. Thus, the posterior $p(\boldsymbol{\theta}|\mathbf{y}) \propto \pi^\dagger(\boldsymbol{\theta})$.

Let $Q(\boldsymbol{\theta}_n, \cdot)$ be a *proposal distribution*, which we assume we can sample from, with its own (potentially unnormalised) density $Q(\boldsymbol{\theta}_n, d\boldsymbol{\theta}_m) \propto q(\boldsymbol{\theta}_n, \boldsymbol{\theta}_m)d\boldsymbol{\theta}_m$. The MH algorithm constructs this Markov chain as follows:

1. choose some initialisation $\boldsymbol{\theta}_0$.
2. For $n = 0, \dots, N$ do the following:
 - (a) given $\boldsymbol{\theta}_n$, sample a proposal $\boldsymbol{\theta}' \sim Q(\boldsymbol{\theta}_n, \cdot)$.
 - (b) Compute the *acceptance ratio*, $\alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}')$, given by the following equation:

$$\alpha(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \min \left[1, \frac{\pi^\dagger(\tilde{\boldsymbol{\theta}})q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})}{\pi^\dagger(\boldsymbol{\theta})q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})} \right]. \quad (5.27)$$
 - (c) Generate a number uniformly in $[0, 1]$, i.e. draw $V \sim U(0, 1)$.
 - (d) If $V \leq \alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}')$ then accept $\boldsymbol{\theta}'$, i.e. set $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}'$, else reject and set $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n$.

For completeness, these steps are described in pseudo-code in Algorithm 8 in Appendix A.3.2. Notice that $\pi^\dagger(\tilde{\boldsymbol{\theta}})/\pi^\dagger(\boldsymbol{\theta}) = p(\tilde{\boldsymbol{\theta}}|\mathbf{y})/p(\boldsymbol{\theta}|\mathbf{y})$ justifying why we do not need the normalising constant for the posterior distribution. The transition kernel of the chain formed above follows as,

$$P(\boldsymbol{\theta}_n, d\boldsymbol{\theta}_m) = q(\boldsymbol{\theta}_n, \boldsymbol{\theta}_m)\alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}_m)d\boldsymbol{\theta}_m, \quad (5.28)$$

since in order to move from state $\boldsymbol{\theta}_n$ to state $\boldsymbol{\theta}_m$ we must first propose $\boldsymbol{\theta}_m$, corresponding to the term $q(\boldsymbol{\theta}_n, \boldsymbol{\theta}_m)$ above, and then accept this proposal, corresponding to the term $\alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}_m)$ above. Using this transition kernel, together with the form of the acceptance ratio one can show that the chain formed is *reversible* which is sufficient for $\pi^\dagger(\cdot)$ to be an invariant distribution of the

chain. For more theoretical details on proving this, together with convergence issues, the reader is directed to [129].

Remark 5.6. We have yet to specify how to choose the proposals in the MH algorithm above. There are several well-known classes of proposals [129]; we detail some of these below:

- **Symmetric Metropolis Algorithm:** Choosing any symmetric proposal density,

$$q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}), \quad (5.29)$$

results in the following simplification of the acceptance probability Eq. (5.27):

$$\alpha(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \min \left[1, \frac{\pi^\dagger(\tilde{\boldsymbol{\theta}})}{\pi^\dagger(\boldsymbol{\theta})} \right]. \quad (5.30)$$

- **Random Walk Metropolis-Hastings (RWMH):** Choosing a proposal density which is a function of the difference between points, i.e.

$$q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = g(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}), \quad (5.31)$$

yields what is known as a Random Walk Metropolis-Hastings algorithm. Perhaps one of the most common choices for such a proposal density is a Gaussian, $Q(\boldsymbol{\theta}, \cdot) = \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I})$ for some $\sigma > 0$. Another commonly used choice, in the case of a Markov Chain with a scalar state space, is the uniform distribution, $Q(\theta, \cdot) = U(\theta - 1, \theta + 1)$.

- **Independence Sampler:** Choosing a proposal density which does not depend on the current state, $q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = g(\tilde{\boldsymbol{\theta}})$ yields the so called Independence Sampler.
- **Metropolis Adjusted Langevin Algorithm (MALA):** Consider the following over-damped Langevin Itô diffusion process:

$$d\boldsymbol{\theta}_t = \frac{1}{2} \nabla \log \pi^\dagger(\boldsymbol{\theta}_t) dt + d\mathbf{W}_t, \quad (5.32)$$

where here $\mathbf{W}_t$ is a standard multi-dimensional Brownian motion. The stationary distribution of this diffusion process can be shown to be the posterior $p(\boldsymbol{\theta}|\mathbf{y})$. By discretising the dynamics via an Euler-Maruyama method we obtain the following proposal distribution which yields the Metropolis Adjusted Langevin Algorithm (MALA):

$$Q(\boldsymbol{\theta}, \cdot) = \mathcal{N} \left(\boldsymbol{\theta} + \frac{\delta}{2} \nabla \log \pi^\dagger(\boldsymbol{\theta}), \delta \mathbf{I} \right), \quad (5.33)$$

for some (small) step-size $\delta > 0$.



Remark 5.7. The MH algorithm with step-sized based proposals suffers from a trade-off between the diversity of samples and the acceptance ratio; larger moves are rejected more frequently, while smaller moves lead to more autocorrelation of the chain. The HMC algorithm also suffers from this, though this is somewhat mitigated by allowing one to move much more rapidly through the target posterior distribution, with distant proposals being much more likely to be accepted than in the MH algorithm. ◆

Hamiltonian Monte Carlo Algorithm

We now briefly motivate and detail the HMC algorithm which is a variation on the MH algorithm where proposals are made via computing a trajectory specified by Hamiltonian dynamics. By following such dynamics we are able to propose more distant states, while retaining high acceptance probabilities (see the trade-offs of MH in Remark 5.7), leading to more efficient exploration of the so-called *typical set* [9] where the majority of the probability mass is concentrated.

HMC starts off by introducing an auxiliary “momentum” variable, $\boldsymbol{\varphi}$, which is paired with the original variable of interest, $\boldsymbol{\theta}$, now viewed as a “position” variable. HMC then proceeds by translating the problem of sampling from the required posterior distribution to a physical system which evolves with respect to a Hamiltonian given by the required posterior. Thus, the key ingredient to being able to utilise Hamiltonian dynamics for MCMC is a way of translating a distribution into a potential energy function. Once the problem is translated to this physical representation we can simulate a Markov chain where each iteration involves a randomly sampled momentum followed by a Metropolis style update using the proposal from the Hamiltonian dynamics.

This translation is achieved by borrowing a concept from statistical mechanics; given a potential energy function $E(\mathbf{x})$, for the state, $\mathbf{x}$, of our physical system, one has that the associated *canonical distribution* [110] over states has the following p.d.f.:

$$P(\mathbf{x}) = \frac{1}{W} \exp\left(-\frac{E(\mathbf{x})}{T}\right), \quad (5.34)$$

where T is the *temperature* of the physical system and W is a normalizing constant. By inverting this relationship we can obtain, for a given p.d.f., the corresponding description of a physical system by choosing the energy to be given as:

$$E(\mathbf{x}) = -\log P(\mathbf{x}) - \log W, \quad (5.35)$$

where we have set $T = 1$.

The *Hamiltonian* is then introduced as the energy function of the joint state of position and

momentum, $\mathbf{x} := (\boldsymbol{\theta}, \boldsymbol{\varphi})$, as follows:

$$H(\boldsymbol{\theta}, \boldsymbol{\varphi}) := U(\boldsymbol{\theta}) + K(\boldsymbol{\varphi}), \quad (5.36)$$

with the associated joint distribution being:

$$\tilde{\pi}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \propto \exp(-H(\boldsymbol{\theta}, \boldsymbol{\varphi})). \quad (5.37)$$

The function $U(\boldsymbol{\theta})$ in Eq. (5.36) corresponds to the potential energy of our system and in the Bayesian context is taken to be

$$U(\boldsymbol{\theta}) = -\log(\pi^\dagger(\boldsymbol{\theta})), \quad (5.38)$$

while the *kinetic energy* term $K(\boldsymbol{\varphi})$ is typically chosen to correspond to an independent Gaussian distribution, $\pi(\boldsymbol{\varphi})$, for $\boldsymbol{\varphi}$:

$$K(\boldsymbol{\varphi}) = -\log(\pi(\boldsymbol{\varphi})) = \frac{1}{2}\boldsymbol{\varphi}^T \mathbf{M}^{-1} \boldsymbol{\varphi}, \quad (5.39)$$

where $\mathbf{M}$ is commonly referred to as the “*mass matrix*” and is usually taken to be diagonal.

Remark 5.8. Note that the choice of Hamiltonian Eq. (5.36) implies that the position and momentum variables are independent from each other. This is not the only possible choice; for more details see section 4.2 of [9]. Also note that the marginal distribution of $\boldsymbol{\theta}$ in Eq. (5.37) is by construction the posterior distribution we require. Thus, HMC samples the joint state, $(\boldsymbol{\theta}, \boldsymbol{\varphi})$, and by keeping only the position variables we obtain the required posterior samples of interest. ♦

The time evolution of the joint state $(\boldsymbol{\theta}, \boldsymbol{\varphi})$ is then given by *Hamilton’s equations*:

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= \frac{\partial H}{\partial \boldsymbol{\varphi}}, \\ \frac{d\boldsymbol{\varphi}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}}, \end{aligned} \quad (5.40)$$

which for the choices of potential and kinetic energies Eqs. (5.38) and (5.39) above yields:

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= \mathbf{M}^{-1} \boldsymbol{\varphi}, \\ \frac{d\boldsymbol{\varphi}}{dt} &= \frac{\partial \log(\pi^\dagger(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}. \end{aligned} \quad (5.41)$$

In the proposal step of HMC we start from the current joint state, $(\boldsymbol{\theta}, \boldsymbol{\varphi})$, and propose a new state by integrating Hamilton’s equations Eq. (5.40) for a fixed amount of time to move along a trajectory in phase space. Since in general this coupled system of differential equations cannot be solved analytically HMC instead discretises the system using a leapfrog scheme [110]. This leapfrog integrator, also known as the Verlet integrator, defines Hamiltonian dynamics with respect to an

approximation of the original Hamiltonian. As a result, it has the important property of being a *symplectic integrator*, i.e. it preserves volumes in phase space. A single step of the leapfrog scheme is given by the following equations:

$$\boldsymbol{\varphi}(t + \epsilon/2) = \boldsymbol{\varphi}(t) - \frac{\epsilon}{2} \frac{\partial}{\partial \boldsymbol{\theta}} U(\boldsymbol{\theta}(t)), \quad (5.42)$$

$$\boldsymbol{\theta}(t + \epsilon) = \boldsymbol{\theta}(t) + \epsilon \frac{\partial}{\partial \boldsymbol{\varphi}} K(\boldsymbol{\varphi}(t + \epsilon/2)), \quad (5.43)$$

$$\boldsymbol{\varphi}(t + \epsilon) = \boldsymbol{\varphi}(t + \epsilon/2) - \frac{\epsilon}{2} \frac{\partial}{\partial \boldsymbol{\theta}} U(\boldsymbol{\theta}(t + \epsilon)). \quad (5.44)$$

Having laid the foundations for HMC we can now detail the main steps of the algorithm:

1. choose some initialisation for position $\boldsymbol{\theta}_0$.
2. For $n = 0, \dots, N$ do the following:
 - (a) sample a new momentum from the momentum distribution, $\boldsymbol{\varphi}_{n+1} \sim \pi(\cdot)$.
 - (b) Update the current joint state to $(\boldsymbol{\theta}_n, \boldsymbol{\varphi}_{n+1})$.
 - (c) Starting from the current joint state, $(\boldsymbol{\theta}_n, \boldsymbol{\varphi}_{n+1})$, propose a new state, $(\boldsymbol{\theta}', \boldsymbol{\varphi}')$, by evolving via Hamilton's equations Eq. (5.40). This is done by performing L leapfrog steps, each of size ϵ , according to Eqs. (5.42) to (5.44) resulting in an evolution for time ϵL .
 - (d) Negate the resulting momentum $\boldsymbol{\varphi}' \leftarrow (-\boldsymbol{\varphi}')$.
 - (e) Compute the *acceptance ratio*, $\alpha((\boldsymbol{\theta}_n, \boldsymbol{\varphi}_{n+1}), (\boldsymbol{\theta}', \boldsymbol{\varphi}'))$ given by the corresponding version of Eq. (5.27):

$$\begin{aligned} \alpha((\boldsymbol{\theta}, \boldsymbol{\varphi}), (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\varphi}})) &= \min \left[1, \frac{\tilde{\pi}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\varphi}})}{\tilde{\pi}(\boldsymbol{\theta}, \boldsymbol{\varphi})} \right] \\ &= \min \left[1, \exp \left(H(\boldsymbol{\theta}, \boldsymbol{\varphi}) - H(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\varphi}}) \right) \right], \end{aligned} \quad (5.45)$$

where we note that the negation of the resulting momentum from the leapfrog steps together with the time-reversibility of the leapfrog scheme means the terms corresponding to the proposal in Eq. (5.27) cancel out.

- (f) Generate a number uniformly in $[0, 1]$, i.e. draw $V \sim U(0, 1)$.
- (g) If $V \leq \alpha((\boldsymbol{\theta}_n, \boldsymbol{\varphi}_{n+1}), (\boldsymbol{\theta}', \boldsymbol{\varphi}'))$ then accept the proposal, i.e. set $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}'$, else reject and set $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n$. Technically we should also set $\boldsymbol{\varphi}_{n+1} \leftarrow \boldsymbol{\varphi}'$ but this will just be replaced by a random draw of momentum in step (a) above.

For completeness, these steps are given in pseudo-code in Algorithm 9 in Appendix A.3.2. There are three routes to tuning the performance of the resulting algorithm:

- The choice of distribution corresponding to the kinetic energy, i.e. the choice of $\pi(\boldsymbol{\varphi})$. In the case of an independent Gaussian discussed above this involves tuning the mass matrix $\mathbf{M}$ with an ideal choice being to set the mass matrix to the posterior covariance matrix. Since this is assumed to be inaccessible, as otherwise we would not need MCMC methods, [9] suggests computing an empirical estimate of the posterior covariance matrix using an initial warm-up run of the HMC algorithm with some other mass matrix.
- Tuning the leapfrog step size, ϵ .
- Tuning the number of leapfrog steps L .

In [63] the authors argue that it is optimal and efficient to set ϵ and L so that $\epsilon L = 1$ and then tuning these two parameters to achieve an approximate acceptance rate of around 65%. A recent and increasingly popular alternative to manually tuning the leapfrog parameters is to utilise the NUTS algorithm which adaptively determines L by utilising a recursive algorithm which is able to stop early by detecting when the trajectory is about to double back on itself (i.e. when a “U-turn” is about to happen). In the experiments in Chapter 6 we will utilise the NUTS algorithm implemented in NumPyro. It is also possible to incorporate posterior geometry into the proposal to yield further gains in efficiency [64].

Remark 5.9. Note that we require the derivatives of the logarithms of both the prior and likelihood in Hamilton’s equations Eq. (5.41) in order to compute the derivative of $\log(\pi^\dagger(\boldsymbol{\theta}))$. These can be computed analytically for simple models, however for more complex models we will have to resort to techniques such as auto-diff to compute these gradients. For the work in Chapter 6 we utilise a HMC implementation found in the Python package NumPyro which uses the popular JAX framework [19] to provide auto-diff capabilities as well as efficient linear algebra routines. ♦

For more details on HMC the interested reader is directed to the works [63, 129, 110, 9, 10]. In particular, [9] provides a very accessible and intuitive discussion of the motivation for the use of Hamiltonian dynamics for MCMC.

5.4 Differential Geometry and Manifolds

5.4.1 Basics of Differential Geometry

In this section we introduce and discuss the essential concepts from differential geometry which we will require for this thesis. We follow the presentations given in [165, 17].

Differentiable Manifolds

The focus of differential geometry is the study of *differentiable manifolds*, which are generalisations of the concept of curves and surfaces (one and two dimensional objects respectively) to higher dimensions. An n -dimensional differentiable manifold $\mathcal{M}$ can intuitively be thought of as a topological space which locally resembles $\mathbb{R}^n$ and which satisfies certain smoothness properties. In this thesis, we will especially be concerned with numerical applications; the concept of *embedded submanifolds in a Euclidean space* is of special relevance for such applications. In particular, the Stiefel manifold, which we will introduce in Section 5.4.3 is an example of an embedded submanifold which will be vital for this thesis.

For the purposes of this thesis we assume the reader is familiar with the definitions of differentiable manifolds and embedded submanifolds. A precise treatment of these concepts is beyond the scope of this thesis; the interested reader is directed to [17, 153, 92, 77] for a more in-depth treatment of differential geometry. The reader is also directed towards [2] for a treatment of differential geometry specialised to matrix manifold applications. For convenience we include some additional details regarding these concepts in Appendix A.4, including a definition of n -dimensional embedded submanifolds of $\mathbb{R}^{n+d}$ in Definition A.14.

Tangent spaces

We will require the concept of a *tangent space* of a differentiable manifold. In particular, we will require this concept for embedded submanifolds and for *quotient manifolds* which we will soon define. We start with the definition of tangent spaces for differentiable submanifolds, which we adopt from Definition 2.3 in [165]:

Definition 5.4 (*Tangent space of a differentiable submanifold*). Let $\mathcal{M} \subset \mathbb{R}^{n+d}$ be an n -dimensional submanifold of $\mathbb{R}^{n+d}$. The *tangent space* of $\mathcal{M}$ at a point $p \in \mathcal{M}$, denoted by $T_p\mathcal{M}$, is the space of velocity vectors of differentiable curves passing through p , i.e.

$$T_p\mathcal{M} = \{\dot{c}(t_0) | c : I \rightarrow \mathcal{M}, c(t_0) = p\}. \quad (5.46)$$

In the above, $I \subset \mathbb{R}$ is an arbitrarily small open interval with $t_0 \in I$. ♦

Note that it is possible to define tangent spaces for general differentiable manifolds, though we will not require this concept for the thesis. One can easily show that the tangent space is actually a vector space with dimension equal to the dimension of the manifold, i.e.

$$\dim T_p\mathcal{M} = \dim \mathcal{M} = n, \quad (5.47)$$

for all $p \in \mathcal{M}$. For different characterisations of tangent spaces and a justification of the dimension of the tangent space being equal to that of the manifold please see Appendix A.4 and Remark A.14 respectively. Definition 5.4 is illustrated in Fig. 5.1 which is taken from [165, Figure 2.1].

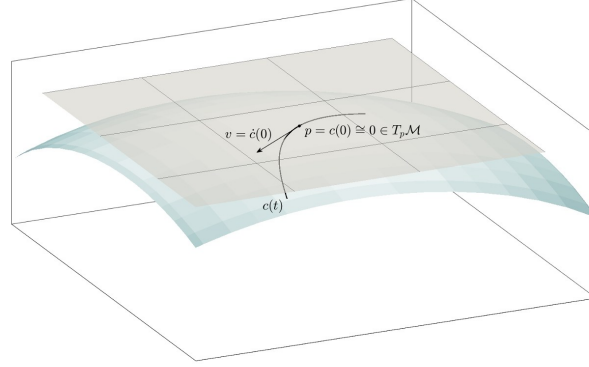


Figure 5.1: Visualization of the tangent space $T_p \mathcal{M}$ of a manifold (shown here as a curved surface) at a point $p \in \mathcal{M}$. Note that the tangent vector $v = \dot{c}(0) \in T_p \mathcal{M}$ is the velocity vector of a curve on the manifold $c : I \rightarrow \mathcal{M}$. This image is taken from [165].

Tangent spaces for quotient manifolds will be discussed later when we introduce Riemannian quotient manifolds.

Riemannian Manifolds

Having defined tangent spaces we are now in a position to introduce the concept of a *Riemannian metric* on an embedded submanifold, adopting Definition 2.6 from [165]:

Definition 5.5 (*Riemannian Metric on a submanifold*). Let $\mathcal{M}$ be a differentiable submanifold of $\mathbb{R}^{n+d}$. A *Riemannian metric* on $\mathcal{M}$ is a family $(\langle \cdot, \cdot \rangle_p)_{p \in \mathcal{M}}$ of inner products $\langle \cdot, \cdot \rangle_p : T_p \mathcal{M} \times T_p \mathcal{M} \rightarrow \mathbb{R}$ that is smooth in variations of the base point p . ♦

A manifold $\mathcal{M}$ equipped with a Riemannian metric is called a *Riemannian manifold*. One can define the concept of length on a Riemannian manifold as follows: the length of a tangent vector $v \in T_p \mathcal{M}$ is given by $\|v\|_p := \sqrt{\langle v, v \rangle_p}$. Similarly, the length of a curve on $\mathcal{M}$, $c : [a, b] \rightarrow \mathcal{M}$, can be defined as:

$$L(c) = \int_a^b \|\dot{c}(t)\|_{c(t)} dt = \int_a^b \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}} dt. \quad (5.48)$$

Having introduced the length of curves on a Riemannian manifold, we are now in a position to define the important concept of *Riemannian distance* between two points $p, q \in \mathcal{M}$:

Definition 5.6 (*Riemannian distance*). Let $\mathcal{M}$ be a Riemannian manifold. The *Riemannian distance* between two points $p, q \in \mathcal{M}$ with respect to the given Riemannian metric is defined as

follows:

$$d_{\mathcal{M}}(p, q) = \inf\{L(c) \mid c : [a, b] \rightarrow \mathcal{M} \text{ piecewise smooth, } c(a) = p, c(b) = q\}, \quad (5.49)$$

where we adopt the convention that $\inf\{\emptyset\} = \infty$. $\blacklozenge$

We will refer to the shortest path between two points $p, q \in \mathcal{M}$ as a *geodesic*. Note that the actual definition of a geodesic is slightly more technical than is required for this thesis. As is remarked in [165, see Remark 1] geodesics can be defined via an IVP of second order. The local uniqueness and existence of geodesics for any points in a connected component of a manifold then follows from the well-known existence and uniqueness theorem for IVPs. The interested reader is directed to [17, sections 5.8, 10.1], and [2, Section 5.4] for more details.

Central to the work in Chapter 6 are the concepts of the *Riemannian exponential* and the *Riemannian logarithm* which we define below, adopting the notation from Section 2.4 in [165]:

Definition 5.7 (*Riemannian Exponential*). The *Riemannian Exponential* is a map which intuitively maps a given tangent vector $v \in T_p\mathcal{M}$ to the endpoints of a geodesic (which is always possible due to the local uniqueness and existence of geodesics) that starts from $p \in \mathcal{M}$ with velocity v . More formally, let $B_\epsilon(0) \subset T_p\mathcal{M}$ be the open ball of radius $\epsilon > 0$ centered at $0 \in T_p\mathcal{M}$. The Riemannian exponential is the map $\exp_p^{\mathcal{M}} : B_\epsilon(0) \rightarrow \mathcal{M}$ which maps v to $q := \exp_p^{\mathcal{M}}(v) := c_{p,v}(1)$, where $c_{p,v} : [0, 1] \rightarrow \mathcal{M}$ is the geodesic which starts from p with velocity v , i.e. $c_{p,v}(0) = p$ and $\dot{c}_{p,v}(0) = v$. $\blacklozenge$

Provided that $\epsilon > 0$ is small enough one can show that the Riemannian exponential is in fact a smooth diffeomorphism between $B_\epsilon(0)$ and an open domain $\mathcal{D}_p \subset \mathcal{M}$ around p . The Riemannian exponential is thus invertible and its smooth inverse map is called the *Riemannian logarithm*:

Definition 5.8 (*Riemannian Logarithm*). The *Riemannian logarithm* is the inverse of the Riemannian exponential map. Using the same notation as in Definition 5.7 the Riemannian logarithm is the map $\log_p^{\mathcal{M}} : \mathcal{D}_p \rightarrow B_\epsilon(0)$ which maps $q \in \mathcal{D}_p \subset \mathcal{M}$ to $v := (\exp_p^{\mathcal{M}})^{-1}(q)$, where v satisfies $c_{p,v}(1) = q$. $\blacklozenge$

The Riemannian logarithm is concerned with the so-called *geodesic endpoint problem*: given $p, q \in \mathcal{M}$ find a geodesic connecting p and q .

Remark 5.10. It is important to remember that the definition of both the Riemannian exponential and Riemannian logarithm maps are dependent on the definition of geodesics which, in turn, are determined by the choice of Riemannian metric; a different choice of metric can yield different exponential and logarithm mappings. $\blacklozenge$

Remark 5.11. To help illustrate the Riemannian exponential and logarithm consider the following two examples taken from [17, see Section 10.2]:

- In the Euclidean space, $\mathbb{R}^d$, with the standard inner product, the Riemannian exponential is given by:

$$\exp_{\mathbf{x}}^{\mathbb{R}^d}(\mathbf{v}) = \mathbf{x} + \mathbf{v}, \quad (5.50)$$

since geodesics in $\mathbb{R}^d$ are given by straight lines. Thus, the Riemannian logarithm is given by:

$$\log_{\mathbf{x}}^{\mathbb{R}^d}(\mathbf{y}) = \mathbf{y} - \mathbf{x}. \quad (5.51)$$

- On the unit sphere, S^{d-1} , in $\mathbb{R}^d$ (equipped the standard inner product of $\mathbb{R}^d$) the Riemannian exponential is given by:

$$\exp_{\mathbf{x}}^{S^{d-1}}(\mathbf{v}) = \cos(\|\mathbf{v}\|) \mathbf{x} + \frac{\sin(\|\mathbf{v}\|)}{\|\mathbf{v}\|} \mathbf{v}. \quad (5.52)$$

Provided we restrict the domain of $\exp_{\mathbf{x}}^{S^{d-1}}$ to contain only those tangent vectors $\mathbf{v} \in T_{\mathbf{x}}S^{d-1}$ with norm strictly less than π then we have that the Riemannian logarithm is given by:

$$\log_{\mathbf{x}}^{S^{d-1}}(\mathbf{y}) = \arccos(\mathbf{x}^T \mathbf{y}) \frac{\mathbf{y} - (\mathbf{x}^T \mathbf{y}) \mathbf{x}}{\|\mathbf{y} - (\mathbf{x}^T \mathbf{y}) \mathbf{x}\|}, \quad (5.53)$$

Note that this formula depends on the choice of domain for the Riemannian exponential and choosing other domains would yield different expressions. For a more detailed discussion of this example the interested reader is directed to Example 10.21 of [17].

◆

Quotient manifolds

The Grassmann manifold, which will be introduced in Section 5.4.3 and which will play a central role in Chapter 6, is an example of a *quotient manifold*. A quotient manifold is a manifold formed by considering certain points on another manifold to be equivalent, as will be made clear below.

In order to properly define quotient manifolds we will first require several concepts which we now outline. We call a differentiable manifold $\mathcal{G}$ a *Lie group* when it also has a group structure which is smooth in the sense that the group operations of ‘multiplication’ and ‘inversion’ are both smooth, i.e. the maps $(g, h) \mapsto g \cdot h$ and $g \mapsto g^{-1}$ are smooth maps from $\mathcal{G} \times \mathcal{G} \rightarrow \mathcal{G}$ and $\mathcal{G} \rightarrow \mathcal{G}$ respectively.

Let $\mathcal{G}$ be a Lie group, let $\mathcal{M}$ be a differentiable manifold, and consider a left action of $\mathcal{G}$ on $\mathcal{M}$ given by the mapping $(g, p) \mapsto g \cdot p \in \mathcal{M}$ for all $(g, p) \in \mathcal{G} \times \mathcal{M}$. Note that one can also consider right actions. In this thesis we will always take $\mathcal{M}$ to be a matrix manifold and so the operation “ $\cdot$ ” in the group action can be considered to be the standard matrix product. This left action gives

rise to an **equivalence relation** on $\mathcal{M}$ defined via,

$$p \sim q \iff \exists g \in \mathcal{G} \text{ such that } g \cdot p = q, \quad (5.54)$$

i.e. two points $p, q \in \mathcal{M}$ are equivalent if and only if there is some $g \in \mathcal{G}$ which left acts on p to yield q . The equivalence classes of this equivalence relation are called $\mathcal{G}$ -orbits and are denoted by $[p] := \mathcal{G}p := \{g \cdot p | g \in \mathcal{G}\}$. We denote by $\mathcal{M}/\mathcal{G} := \{[p] | p \in \mathcal{M}\}$ the orbit space, i.e. the quotient space of $\mathcal{M}$ by the (left) action of $\mathcal{G}$. The map $\Pi : \mathcal{M} \rightarrow \mathcal{M}/\mathcal{G}$ which sends a point $p \in \mathcal{M}$ to its $\mathcal{G}$ -orbit $[p]$ is called the *quotient map*. We say the action is *free*, if every *isotropy group* $\mathcal{G}_p := \{g \in \mathcal{G} | g \cdot p = p\}$ is the trivial group, i.e. $\mathcal{G}_p = \{e\}$, where e is the identity element of the group $\mathcal{G}$. Quotient manifolds are manifolds formed from the quotient spaces of such actions, and can be thought of as the manifold resulting from bundling/quotienting all points in a $\mathcal{G}$ -orbit into one. This is made precise by the *Quotient Manifold Theorem* from [165, see Theorem 2.9]:

Theorem 5.1 (Quotient Manifold Theorem). *Suppose $\mathcal{G}$ is a Lie group acting smoothly, freely, and properly on a smooth manifold $\mathcal{M}$. Then the orbit space $\mathcal{M}/\mathcal{G}$ is itself a manifold of dimension $\dim(\mathcal{M}/\mathcal{G}) = \dim(\mathcal{M}) - \dim(\mathcal{G})$, and has a unique smooth structure such that the quotient map $\Pi : \mathcal{M} \rightarrow \mathcal{M}/\mathcal{G}$ is a smooth submersion, i.e. it is a smooth surjective mapping with surjective differential at every point. We call $\mathcal{M}$ the total space and $\mathcal{M}/\mathcal{G}$ is called the quotient space. $\blacklozenge$*

Given any point q in a quotient manifold we call the preimage of q by the quotient map, $\Pi^{-1}(q) \subset \mathcal{M}$, a *fiber* and denote it by $\mathcal{M}_q := \Pi^{-1}(q)$. Recalling that Π is a smooth submersion one has that each fiber $\mathcal{M}_q$ is a closed embedded submanifold of $\mathcal{M}$, which follows via the well known Implicit Function Theorem.

Riemannian Quotient Manifolds

In this thesis we will be concerned exclusively with quotient manifolds in which the total space $\mathcal{M}$ is a Riemannian manifold. This extra structure on the total space transfers over to the resulting quotient manifold allowing us to nicely characterise the tangent spaces of the quotient manifold and to even define a Riemannian metric on the quotient manifold yielding a *Riemannian quotient manifold*. This will now be elaborated on. Let $\mathcal{M}$ be a Riemannian manifold equipped with a Riemannian metric $\langle \cdot, \cdot \rangle_p^{\mathcal{M}}$. Consider a quotient manifold of $\mathcal{M}$ given by Theorem 5.1, $\mathcal{N} := \mathcal{M}/\mathcal{G}$. Consider a point $p \in \mathcal{M}$ and its associated tangent space $T_p\mathcal{M}$. Since the fiber $\mathcal{M}_{\Pi(p)}$ is a closed embedded submanifold of $\mathcal{M}$ and since we have an inner product on $T_p\mathcal{M}$ via the Riemannian metric we can decompose the tangent space $T_p\mathcal{M}$ as an orthogonal direct sum:

$$T_p\mathcal{M} = T_p\mathcal{M}_{\Pi(p)} \oplus (T_p\mathcal{M}_{\Pi(p)})^\perp. \quad (5.55)$$

We call the tangent space of the fiber the *vertical space*, $V_p := T_p \mathcal{M}_{\Pi(p)}$ and we call its orthogonal complement the *horizontal space*, $H_p := V_p^\perp$. Recalling that the quotient map Π is a smooth submersion, we have that the differential of Π at p is surjective, i.e. $d\Pi_p : T_p \mathcal{M} \rightarrow T_{\Pi(p)} \mathcal{N}$ is surjective. We can thus identify the vertical space with the kernel of the differential, i.e. $V_p = \ker(d\Pi_p)$, while the horizontal space is isomorphic to the tangent space $T_{\Pi(p)} \mathcal{N}$, i.e. $H_p \cong T_{\Pi(p)} \mathcal{N}$. The quotient map is thus a local diffeomorphism and its differential can be used to represent a tangent vector $\xi \in T_{\Pi(p)} \mathcal{N}$ by a concrete vector in the horizontal space given uniquely by $(d\Pi(p)|_{H_p})^{-1}[\xi]$. This representation is called the *horizontal lift* of ξ at p and is denoted by $\text{lift}_p(\xi) := (d\Pi(p)|_{H_p})^{-1}[\xi]$. For more details the reader is referred to [17, Section 9.4]. Provided that the equivalence relation and the Riemannian metric satisfy the following relation,

$$p \sim q \implies \langle \text{lift}_p(\xi), \text{lift}_p(\zeta) \rangle_p = \langle \text{lift}_q(\xi), \text{lift}_q(\zeta) \rangle_q \quad \forall \xi, \zeta \in T_{\Pi(p)} \mathcal{N}, \quad (5.56)$$

one can define a Riemannian metric on the quotient manifold $\mathcal{N}$ via:

$$\langle \xi, \zeta \rangle_{\Pi(p)} := \langle \text{lift}_p(\xi), \text{lift}_p(\zeta) \rangle_p \quad \forall \xi, \zeta \in T_{\Pi(p)} \mathcal{N}. \quad (5.57)$$

For a more detailed discussion of this the reader is referred to [17, Section 9.7].

5.4.2 Statistics on Riemannian Manifolds

Since we will be dealing with manifold-valued random variables in Chapter 6 we now briefly discuss the basics of statistics for data on Riemannian manifolds. In particular, we will discuss how moments can be generalised from Euclidean spaces to Riemannian manifolds. We follow the presentation in [58].

We start with the first moment. In the Euclidean setting the sample mean is the standard point estimate for the “centre of mass” of a set of points. This statistic is perhaps the simplest to define; for a set of points $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ the sample mean is given by:

$$\bar{\mathbf{x}}_N := \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i. \quad (5.58)$$

Note that defining the sample mean via Eq. (5.58) relies on the vector space operation of addition in Euclidean space. Since Riemannian manifolds are not vector spaces in general, this definition cannot be directly used for data on a manifold.

Upon first glance it might seem possible to remedy this issue for manifolds embedded in Euclidean space, $\mathcal{M} \subset \mathbb{R}^d$, since we can just utilise the vector addition operator of the ambient space $\mathbb{R}^d$. This however also fails since the resulting point often lies outside the manifold $\mathcal{M}$. To see this

consider Example 1 from [58]: take the unit sphere in $\mathbb{R}^3$, $S^2 \subset \mathbb{R}^3$, and consider the two points $\mathbf{x}_1 := (1, 0, 0)^T$, $\mathbf{x}_2 := (0, 1, 0)^T \in S^2$. If we average these two points, viewing them as vectors in $\mathbb{R}^3$, we obtain the point $\bar{\mathbf{x}}_2 = (1/2, 1/2, 0)^T$ which has norm $\|\bar{\mathbf{x}}_2\| = \sqrt{2}/2$ and thus falls out of S^2 .

To make progress towards generalising the sample mean to Riemannian manifolds we can consider alternative characterisations of the sample mean in Euclidean space in the hopes that some of these might be more amenable to generalisation. There are several such alternative definitions for the sample mean of points $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, and we now outline two of these:

1. **Algebraic definition:** The sample mean can be defined algebraically as the unique point for which the *sum of residuals* is zero, i.e. the unique point $\tilde{\mathbf{x}}$ such that:

$$\sum_{i=1}^N (\mathbf{x}_i - \tilde{\mathbf{x}}) = \mathbf{0}, \quad (5.59)$$

is $\tilde{\mathbf{x}} = \bar{\mathbf{x}}_N$.

2. **Geometric definition:** The sample mean can be defined as the *least squares centroid* of the data points, i.e. it minimizes the following sum of squares:

$$\ell(\mathbf{x}) := \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{x}\|^2, \quad (5.60)$$

over all points $\mathbf{x} \in \mathbb{R}^d$.

Notice that the algebraic definition still relies on the vector space structure of $\mathbb{R}^d$ while the geometric definition only involves a notion of distance between points in $\mathbb{R}^d$. Since on Riemannian manifolds we have a concept of distance between points (see Definition 5.6) we can generalise the sample mean via this geometric characterisation. This leads to the following generalisation, called the Karcher mean [79, 83], which is defined as follows:

Definition 5.9. The *Karcher mean* or *Riemannian centre of mass* of a collection $\{x_i\}_{i=1}^N$ of points on a Riemannian manifold $\mathcal{M}$, if it exists, is defined as follows:

$$\bar{\mu}_N := \operatorname{argmin}_{p \in \mathcal{M}} \sum_{i=1}^N d_{\mathcal{M}}(x_i, p)^2. \quad (5.61)$$

◆

Note that the Karcher mean defined above is an example of a Fréchet mean [60] which is a generalisation of the expectation of probability measures to metric spaces.

The Karcher mean will be very useful in the work presented in Chapter 6. As such we now detail one possible method for estimating it (see Algorithm 1). Since the Karcher mean is a solution

of a Riemannian optimization problem it is possible to utilize a gradient-descent based approach for estimating it. For an accessible discussion of Riemannian optimization the reader is directed to [17]. The key fact such an approach relies on is the fact [79, 165] that the (Riemannian) gradient of the sum in Eq. (5.61) is given by:

$$-2 \sum_{i=1}^N \log_p^{\mathcal{M}}(x_i). \quad (5.62)$$

This leads to Algorithm 1 for estimating the Karcher mean which is adapted from [165, 126].

Algorithm 1: Riemannian Karcher Mean via gradient descent

Input: set of points $\{x_i\}_{i=1}^N \subset \mathcal{M}$ on a Riemannian manifold $\mathcal{M}$, initial guess for the Karcher mean μ_0 , max number of iterations M , tolerance ϵ

Output: Estimate of the Karcher mean μ'

```

1 Set  $\mu' = \mu_0$  ;
2 for  $n \leftarrow 1$  to  $M$  do
3   Compute  $\delta\mu = \frac{1}{N} \sum_{i=1}^N \log_{\mu'}^{\mathcal{M}}(x_i)$  ;
4   Compute  $\alpha := \|\delta\mu\|_{\mu'}$  ;
5    $\mu' \leftarrow \exp_{\mu'}^{\mathcal{M}}(\delta\mu)$  ; // geodesic step
6   if  $\alpha < \epsilon$  then
7     return  $\mu'$  ; // terminate early
8   end
9 end
10 return  $\mu'$ 
```

Remark 5.12. Note that in Algorithm 1 we have chosen to use a step size inversely proportional to the number of points N . It is possible to instead consider a sequence of variable step-sizes. Also note that one can adjust the algorithm presented to store all the intermediate points for μ' together with the corresponding value of the sum in Eq. (5.61). Doing so allows one to use the point which achieves the lowest value for this sum. As discussed further in Section 6.7 it is often advisable, as with any gradient-descent algorithm, to run Algorithm 1 for several different initializations and take the Karcher mean to be the one which achieves the smallest value for the sum. ♦

We can now discuss the second moment. Recall that the sample covariance of a collection of points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d$ is given by:

$$\hat{\Sigma}_N := \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}_N)(\mathbf{x}_i - \bar{\mathbf{x}}_N)^T. \quad (5.63)$$

On first glance this definition does not seem to be amenable to generalisation to Riemannian manifolds as it relies on the vector space operations of $\mathbb{R}^d$, in particular on vector subtraction, transpose, and matrix multiplication. However, recall from Remark 5.11 that the Riemannian

logarithm in Euclidean space is given by

$$\log_{\mathbf{x}}^{\mathbb{R}^d}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}} - \mathbf{x}. \quad (5.64)$$

Thus, we can generalise the sample covariance to Riemannian manifolds by utilising the vector-space structure of the tangent space, the Riemannian logarithm, and the Karcher mean to define the Karcher sample covariance of a collection of points $\{x_i\}_{i=1}^N$ on a Riemannian manifold $\mathcal{M}$ via:

$$\hat{\Sigma}_N := \frac{1}{N-1} \sum_{i=1}^N \log_{\bar{\mu}_N}^{\mathcal{M}}(x_i) \log_{\bar{\mu}_N}^{\mathcal{M}}(x_i)^T. \quad (5.65)$$

5.4.3 The Stiefel and Grassmann Manifolds

In this section we introduce the Stiefel and Grassmann manifolds which are essential ingredients to the work which will be presented in Chapter 6. There are countless references detailing properties of these manifolds; we refer the interested reader to [8, 1, 88, 53, 165, 17] for further details. We again follow the flavour of presentation given in [165, 17].

Stiefel Manifold

We start our discussion with the definition of the *Stiefel manifold*:

Definition 5.10 (*Stiefel Manifold*). For $n \leq d$, the (compact) *Stiefel manifold*, $\text{St}(d, n)$, is the manifold consisting of $d \times n$ orthonormal matrices, i.e.

$$\text{St}(d, n) = \{\mathbf{X} \in \mathbb{R}^{d \times n} | \mathbf{X}^T \mathbf{X} = \mathbf{I}_n\}, \quad (5.66)$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. ♦

Remark 5.13. There is also a *non-compact* Stiefel manifold consisting of the open subset of matrices of rank n in $\mathbb{R}^{d \times n}$. In this thesis when we say the Stiefel manifold we will always refer to the **compact** manifold defined here unless stated otherwise. ♦

The Stiefel manifold $\text{St}(d, n)$ can be shown to be an embedded submanifold of $\mathbb{R}^{d \times n}$ of dimension:

$$\dim \text{St}(d, n) = dn - \frac{n(n+1)}{2}. \quad (5.67)$$

For a justification of this please see Appendix A.4.

Note that points $\mathbf{X} \in \text{St}(d, n)$ can be interpreted as orthonormal bases of cardinality n by taking the columns of $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ where the $\{\mathbf{x}_i\}_{i=1}^n$ are each vectors in $\mathbb{R}^d$ satisfying

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^T \mathbf{x}_j = \delta_{ij}. \quad (5.68)$$

Remark 5.14. The Stiefel manifold yields two edge cases which can be recognized as other familiar mathematical objects:

- **Case $n = 1$:** $\text{St}(d, 1)$ is the unit-sphere in $\mathbb{R}^d$, S^{d-1} . This establishes the unit-sphere as an embedded submanifold in $\mathbb{R}^d$ of dimension $d - 1$.
- **Case $d = n$:** $\text{St}(d, d) = \{\mathbf{X} \in \mathbb{R}^{d \times d} | \mathbf{X}^T \mathbf{X} = \mathbf{I}_d\}$ can be recognized as the set of $d \times d$ *orthogonal matrices*. This matrix manifold is an important example of a Lie group, the *orthogonal group*, $O(d)$.

◆

We now collate several important properties/facts regarding the Stiefel manifold. For further discussion and more detailed proofs establishing these facts the interested reader is directed to [17, 165].

- **Tangent spaces:** The tangent spaces of $\text{St}(d, n)$ are the following subspaces of $\mathbb{R}^{d \times n}$:

$$T_{\mathbf{X}} \text{St}(d, n) = \{\mathbf{U} \in \mathbb{R}^{d \times n} | \mathbf{X}^T \mathbf{U} + \mathbf{U}^T \mathbf{X} = \mathbf{0}\}. \quad (5.69)$$

- **Riemannian metric structure:** Endowing $\mathbb{R}^{d \times n}$ with the standard Frobenius inner product $\langle \mathbf{U}, \mathbf{V} \rangle := \text{tr}(\mathbf{U}^T \mathbf{V})$ allows us to conveniently make $\text{St}(d, n)$ into a Riemannian manifold as a *Riemannian submanifold* of $\mathbb{R}^{d \times n}$, where we take the Riemannian metric for the Stiefel manifold to be the restriction of the Frobenius inner product to $\text{St}(d, n)$.
- **Riemannian exponential:** The Riemannian Stiefel exponential at a base point $\mathbf{X} \in \text{St}(d, n)$, $\exp_{\mathbf{X}}^{\text{St}}$, sends a Stiefel tangent vector $\mathbf{U} \in T_{\mathbf{X}} \text{St}(d, n)$ to the endpoint $\tilde{\mathbf{X}} \in \text{St}(d, n)$ of a geodesic on the Stiefel manifold starting from $\mathbf{X}$ with velocity vector $\mathbf{U}$. It thus also provides an expression for geodesic curves on the Stiefel manifold.
- **Riemannian logarithm:** The Riemannian Stiefel logarithm at a base point $\mathbf{X} \in \text{St}(d, n)$, $\log_{\mathbf{X}}^{\text{St}}$, takes another point $\tilde{\mathbf{X}} \in \text{St}(d, n)$ and finds a Stiefel tangent vector $\mathbf{U} \in T_{\mathbf{X}} \text{St}(d, n)$ such that the geodesic starting from $\mathbf{X}$ with velocity $\mathbf{U}$ hits $\tilde{\mathbf{X}}$ after moving an arc length of $\|\mathbf{U}\|_{\mathbf{X}}$. The Riemannian Stiefel logarithm thus provides two useful pieces of information: the solution to the geodesic endpoint problem of finding a geodesic starting at $\mathbf{X}$ and ending at $\tilde{\mathbf{X}}$ and the Riemannian distance between the two given points $\mathbf{X}, \tilde{\mathbf{X}}$.

Grassmann Manifold

We can now move on to defining the Grassmann manifold:

Definition 5.11 (*Grassmann Manifold*). For $n \leq d$, the *Grassmann manifold*, $\text{Gr}(d, n)$, is the manifold consisting of all n -dimensional (linear) subspaces of $\mathbb{R}^d$, i.e.

$$\text{Gr}(d, n) = \{\mathcal{X} \leq \mathbb{R}^d \mid \dim \mathcal{X} = n\}, \quad (5.70)$$

where the notation $\mathcal{A} \leq \mathcal{B}$ denotes that $\mathcal{A}$ is a linear subspace of the vector space $\mathcal{B}$. $\blacklozenge$

Since each point $\mathcal{X} \in \text{Gr}(d, n)$ is a subspace of dimension n we can represent each point $\mathcal{X}$ in the Grassmann manifold by an orthonormal basis $\{\mathbf{x}_i\}_{i=1}^n$ satisfying $\text{span}(\{\mathbf{x}_i\}_{i=1}^n) = \mathcal{X}$. That is to say, each point $\mathcal{X} \in \text{Gr}(d, n)$ can be represented by a $d \times n$ matrix $\mathbf{X} := [\mathbf{x}_1, \dots, \mathbf{x}_n]$ which is precisely a point on the Stiefel manifold, $\mathbf{X} \in \text{St}(d, n)$. In fact, the Grassmann manifold can be identified as a quotient manifold of the Stiefel manifold by applying Theorem 5.1 with $\mathcal{M} = \text{St}(d, n)$ and $\mathcal{G} = O(n)$, i.e.

$$\text{Gr}(d, n) = \text{St}(d, n)/O(n) = \{[\mathbf{X}] \mid \mathbf{X} \in \text{St}(d, n)\}, \quad (5.71)$$

where the $O(n)$ -orbits are $[\mathbf{X}] = \{\mathbf{X}\mathbf{R} \mid \mathbf{R} \in O(n)\}$. We are thus bundling all points $\mathbf{X}, \mathbf{Y} \in \text{St}(d, n)$ which are related by $\mathbf{Y} = \mathbf{X}\mathbf{R}$ for some $\mathbf{R} \in O(n)$ into a single equivalence class $[\mathbf{X}]$. If the representative $\mathbf{X}$ has columns which span a subspace $\mathcal{X} = \text{span}(\mathbf{X})$ then right acting on $\mathbf{X}$ via some orthogonal matrix $\mathbf{R} \in O(n)$ preserves the subspace represented, i.e. $\text{span}(\mathbf{X}\mathbf{R}) = \text{span}(\mathbf{X}) = \mathcal{X}$; the application of $\mathbf{R}$ on the right simply rotates/reflects the orthonormal basis vectors within the subspace $\mathcal{X}$ onto some other orthonormal basis for the subspace.

Just as for the Stiefel manifold earlier we now collate several important properties/facts regarding the Grassmann manifold. For further discussion and more detailed proofs establishing these facts the interested reader is again directed to [17, 165].

- **Dimension:** The dimension of the Grassmann manifold is:

$$\dim \text{Gr}(d, n) = \dim \text{St}(d, n) - \dim O(n) = n(d - n), \quad (5.72)$$

which follows from its quotient structure and Theorem 5.1.

- **Tangent spaces:** Let $\mathbf{X} \in \text{St}(d, n)$ be a point in the Stiefel manifold. The tangent space based at $\mathbf{X}$ to the Stiefel manifold can be decomposed into the vertical and horizontal spaces from Section 5.4.1 for the Riemannian metric on $\text{St}(d, n)$ given by $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{X}} = \text{tr}(\mathbf{A}^T \mathbf{B})$:

$$T_{\mathbf{X}} \text{St}(d, n) = V_{\mathbf{X}} + H_{\mathbf{X}}. \quad (5.73)$$

The vertical space is:

$$V_{\mathbf{X}} = \{\mathbf{X}\mathbf{\Omega} \mid \mathbf{\Omega} \in \text{Skew}(n)\}, \quad (5.74)$$

where $\text{Skew}(n) = \{\mathbf{A} \in \mathbb{R}^{n \times n} | \mathbf{A}^T = -\mathbf{A}\}$ is the set of skew-symmetric $n \times n$ matrices.

The horizontal space is:

$$H_{\mathbf{X}} = \{\mathbf{U} \in \mathbb{R}^{d \times n} | \mathbf{X}^T \mathbf{U} = \mathbf{0}\}. \quad (5.75)$$

The vertical space $V_{\mathbf{X}}$ is the tangent space to the fiber $\Pi^{-1}(\mathcal{X})$ where $\mathcal{X} = \Pi(\mathbf{X}) = [\mathbf{X}]$ and $\Pi : \text{St}(d, n) \rightarrow \text{Gr}(d, n)$ is the quotient map discussed in Section 5.4.1 for the Grassmann manifold as a quotient of the Stiefel by $O(n)$. Recalling that the horizontal space at $\mathbf{X}$ is isomorphic to the tangent space based at $[\mathbf{X}]$ to the Grassmann manifold allows us to **identify** $T_{[\mathbf{X}]} \text{Gr}(d, n)$ concretely as follows:

$$T_{[\mathbf{X}]} \text{Gr}(d, n) = \{\mathbf{U} \in \mathbb{R}^{d \times n} | \mathbf{X}^T \mathbf{U} = \mathbf{0}\}, \quad (5.76)$$

with the understanding that this actually means $T_{[\mathbf{X}]} \text{Gr}(d, n) \cong H_{\mathbf{X}}$ where the mapping between tangent vectors to the Grassmann and vectors in the horizontal space is given by the horizontal lift map, $\text{lift}_{\mathbf{X}} = (D\Pi(\mathbf{X})|_{H_{\mathbf{X}}})^{-1}$.

Remark 5.15. It is important to point out that it is possible to write every tangent vector $\mathbf{U} \in T_{[\mathbf{X}]} \text{Gr}(d, n)$ as $\mathbf{U} = (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{V}$ where $\mathbf{V} \in \mathbb{R}^{d \times n}$ is arbitrary; this fact will be very useful in Chapter 6. This is simply orthogonal projection to the tangent space which is possible once we have a Riemannian structure on the Grassmann manifold. The matrix $(\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)$ is precisely this orthogonal projector. For more details on this please consult [17]. ♦

- **Riemannian metric structure:** $\text{Gr}(d, n)$ can be turned into a Riemannian manifold by utilising the quotient structure described above to push the metric from $\text{St}(d, n)$ to $\text{Gr}(d, n)$ via the quotient map Π . Explicitly this means that for all $\xi, \zeta \in T_{[\mathbf{X}]} \text{Gr}(d, n)$ we have a Riemannian metric via $\langle \xi, \zeta \rangle_{[\mathbf{X}]} = \langle \mathbf{U}, \mathbf{V} \rangle_{\mathbf{X}} = \text{tr}(\mathbf{U}^T \mathbf{V})$, where $\mathbf{U} = \text{lift}_{\mathbf{X}}(\xi)$ and $\mathbf{V} = \text{lift}_{\mathbf{X}}(\zeta)$. We will abuse notation in the rest of the thesis and use the identification between the horizontal spaces and the tangents spaces of the Grassmann manifold to simply write: $\langle \mathbf{U}, \mathbf{V} \rangle_{[\mathbf{X}]} = \text{tr}(\mathbf{U}^T \mathbf{V})$ for all $\mathbf{U}, \mathbf{V} \in T_{[\mathbf{X}]} \text{Gr}(d, n)$ represented via Eq. (5.76).
- **Riemannian exponential:** The Riemannian Grassmann exponential at a base point $\mathcal{X} \in \text{Gr}(d, n)$, $\exp_{\mathcal{X}}^{\text{Gr}}$, sends a Grassmann tangent vector $\mathbf{U} \in T_{\mathcal{X}} \text{Gr}(d, n)$ to the endpoint $\tilde{\mathcal{X}} \in \text{Gr}(d, n)$ of a geodesic on the Grassmann manifold starting from $\mathcal{X}$ with velocity vector $\mathbf{U}$. It thus also provides an expression for geodesic curves on the Grassmann manifold. For an algorithm to compute the Riemannian Grassmann Exponential please see Algorithm 2 which is adopted from [165, Algorithm 9].
- **Riemannian logarithm:** The Riemannian Grassmann logarithm at a base point $\mathcal{X} \in$

$\text{Gr}(d, n)$, $\log_{\mathcal{X}}^{\text{Gr}}$, takes another point $\tilde{\mathcal{X}} \in \text{Gr}(d, n)$ and finds a Grassmann tangent vector $\mathbf{U} \in T_{\mathcal{X}} \text{Gr}(d, n)$ such that the geodesic starting from $\mathcal{X}$ with velocity $\mathbf{U}$ hits $\tilde{\mathcal{X}}$ after moving an arc length of $\|\mathbf{U}\|_{\mathcal{X}}$. The Riemannian Grassmann logarithm thus provides two useful pieces of information: the solution to the geodesic endpoint problem of finding a geodesic starting at $\mathcal{X}$ and ending at $\tilde{\mathcal{X}}$ **and** the Riemannian distance between the two given points $\mathcal{X}, \tilde{\mathcal{X}}$. For an algorithm to compute the Riemannian Grassmann logarithm please see Algorithm 3 which is adopted from [165, Algorithm 11].

- **Riemannian distance:** We will have need for computing the Riemannian distance (see Definition 5.6) between points on the Grassmann manifold. Algorithm 4 provides a straight-forward method for computing this distance. The algorithm is derived from the discussion following Algorithm 11 in [165].

Algorithm 2: Riemannian Grassmann Exponential

Input: Base point $\mathcal{X} = [\mathbf{X}] \in \text{Gr}(d, n)$, where $\mathbf{X} \in \text{St}(d, n)$, tangent vector $\Delta \in T_{[\mathbf{X}]} \text{Gr}(d, n)$

Output: $\tilde{\mathcal{X}} := \exp_{\mathcal{X}}^{\text{Gr}}(\Delta) = [\tilde{\mathbf{X}}] \in \text{Gr}(d, n)$

- 1 $\mathbf{USV}^T \stackrel{\text{SVD}}{:=} \Delta$, with $\mathbf{U} \in \text{St}(d, n)$; // (thin) SVD of tangent vector
- 2 $\tilde{\mathbf{X}} := \mathbf{XV} \cos(\mathbf{S})\mathbf{V}^T + \mathbf{U} \sin(\mathbf{S})\mathbf{V}^T$; // cos and sin act only on diagonal entries.
- 3 return $\tilde{\mathbf{X}}$

Algorithm 3: Riemannian Grassmann Logarithm

Input: Base point $\mathcal{X} = [\mathbf{X}] \in \text{Gr}(d, n)$, where $\mathbf{X} \in \text{St}(d, n)$, $\mathcal{Y} = [\mathbf{Y}] \in \text{Gr}(d, n)$, where $\mathbf{Y} \in \text{St}(d, n)$

Output: $\Delta := \log_{\mathcal{X}}^{\text{Gr}}(\mathcal{Y}) \in T_{\mathcal{X}} \text{Gr}(d, n)$

- 1 $\Psi \mathbf{SR}^T \stackrel{\text{SVD}}{:=} \mathbf{Y}^T \mathbf{X}$;
- 2 $\mathbf{Y}_* := \mathbf{Y}(\Psi \mathbf{R}^T)$; // ‘Transition to Procrustes representative’
- 3 $\mathbf{L} := (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{Y}_*$;
- 4 $\mathbf{Q}\Sigma\mathbf{V}^T \stackrel{\text{SVD}}{:=} \mathbf{L}$; // (thin) SVD
- 5 $\Delta := \mathbf{Q} \arcsin(\Sigma)\mathbf{V}^T$; // arcsin acts only on diagonal entries
- 6 return Δ

Algorithm 4: Riemannian Grassmann Distance

Input: Points $\mathcal{X} = [\mathbf{X}], \mathcal{Y} = [\mathbf{Y}] \in \text{Gr}(d, n)$, where $\mathbf{X}, \mathbf{Y} \in \text{St}(d, n)$

Output: The Riemannian distance between $\mathcal{X}, \mathcal{Y}$: $d_{\text{Gr}}(\mathcal{X}, \mathcal{Y})$

- 1 $\mathbf{USV}^T \stackrel{\text{SVD}}{:=} \mathbf{X}^T \mathbf{Y}$; // (thin) SVD
- 2 $d_{\text{Gr}}(\mathcal{X}, \mathcal{Y}) = \left(\sum_{k=1}^n \arccos(s_k)^2 \right)^{1/2}$; // here s_k are the singular values of $\mathbf{X}^T \mathbf{Y}$, i.e. the diagonal entries of $\mathbf{S}$
- 3 return $d_{\text{Gr}}(\mathcal{X}, \mathcal{Y})$

Chapter 6

Grassmann-valued Gaussian Processes for Probabilistic Localised Active Subspaces

6.1 Introduction

This chapter contains the contributions for the second project of this thesis which were outlined in Chapter 1. To help guide the exposition which follows it is useful to keep the following example in mind: consider a parameterised elliptic PDE:

$$-\nabla \cdot (\kappa(\mathbf{x}; \mathbf{p}) \nabla u(\mathbf{x}; \mathbf{p})) = f(\mathbf{x}), \quad \mathbf{x} \in D := [0, 1]^2, \quad (6.1)$$

$$u(\mathbf{x}; \mathbf{p}) = 0, \quad \mathbf{x} \in \partial D. \quad (6.2)$$

Note that in this problem the coefficient $\kappa = \kappa(\mathbf{x}, \mathbf{p})$, and hence the solution $u = u(\mathbf{x}, \mathbf{p})$, depends on a set of parameters $\mathbf{p} = (p_1, \dots, p_{d_p})^T \in \mathbb{R}^{d_p}$. When the number of parameters d_p is very large it would be interesting to see if some sort of dimension reduction is possible for this parameter space. For complicated parameter to solution maps it is conceivable that such a dimension reduction will depend on the spatial location $\mathbf{x} \in D$. This leads us to consider what we call a localised active subspace, $\mathcal{M}(\mathbf{x}) \in \text{Gr}(d_p, n)$, which can be thought of as the collection of directions in $\mathbb{R}^{d_p}$ along which $u(\mathbf{x}, \mathbf{p})$ is most sensitive to perturbations (on average) in $\mathbf{p}$. This concept will be expanded on in more detail in this chapter.

In particular, this chapter is structured as follows. In Section 6.2 we provide a literature review of related work to that presented in this chapter. In Section 6.3 we provide a review of the

concepts of optimal ridge approximation and active subspaces. Then in Section 6.4 we motivate the idea of localised active subspaces. Section 6.5 contains the presentation of our GrassGP model. We then conduct several numerical examples in Section 6.7 demonstrating the use of our GrassGP model, including two numerical experiments related to the example problem above in Sections 6.7.2 and 6.7.3.

6.2 Related Work

The main motivation of this chapter stems from the literature surrounding active subspaces. The early development of active subspaces can be attributed to the work of Samarov [133], while more recent developments are due to the works of Constantine et al. [33, 35, 34]. There are also several works which lie at the intersection of the study of *ridge functions* [121] (see Section 6.3 below) and active subspaces. Such works include [138, 158, 137, 36].

Ideas of dimension reduction and variable selection are commonplace within the field of statistics. One example which aims to perform dimension reduction via subspace-based methods is the field of sufficient dimension reduction. This field aims to replace the input vector with their projection onto a subspace formed from the inputs themselves [37]. In the construction of this dimension reducing subspace it is vital that no information regarding either the conditional mean, conditional variance, or in general the conditional distribution of the outputs with respect to the inputs, is lost. Several methods exist in the literature to estimate these subspaces including sliced inverse regression (SIR) [96], sliced average variance estimation (SAVE) [44], and contour regression (CR) [95], among others. For a detailed review of the field of sufficient dimension reduction the reader is directed to [102].

The idea of considering several different low dimensional subspaces instead of one global active subspace is not new and has been investigated in [159] where the authors develop a clustered active subspace method which first locates the “clusters” and then utilises local GP emulators to discover the localised low dimensional structure. The model we consider in this work sidesteps the need to consider an additional clustering step at the loss of only being able to consider the same dimension for the dimension reduction in different parameter regimes.

In [137] the authors introduce the idea of using spatially correlated polynomial ridge functions for quick estimation of flowfields in computational fluid dynamics applications. A key idea in this paper is to only compute the ridge functions at a small number of subsampled locations and then upsample these to new locations by utilising spatial correlations between the full dimensional data. However, this approach is not always applicable as correlations in the full dimensional data need not imply correlations after performing the dimension reduction. In this work we side step this

issue by only exploiting correlations between the projectors themselves.

There has been extensive research into the problem of learning a manifold-valued mapping. These works include non-parametric methods [162, see also references therein] as well as semi-parametric methods [139]. The approach we investigate in this chapter is motivated by that taken in [120] where we utilise a GP instead of a simple linear regression to allow us to more flexibly discover the spatial dependence of our manifold-valued field. The linear model approach taken in [120] builds on the work of [150] where the author generalises linear regression to manifold-valued responses with real covariates via *geodesic regression*. This work was later expanded to include multiple covariates in [84].

Other GP based methods for modelling manifold-valued responses have been considered in the literature. Wrapped Gaussian Processes (WGP) were introduced in [103] to approximate mappings to Riemannian manifolds via wrapped Gaussian distributions which are defined in terms of normal distributions on tangent spaces of the manifold. The model we consider in this work is motivated by WGP's though as the authors of [103, 163] mention the method proposed in [103] is analytically intractable for manifolds with finite injectivity radius, as is the case here for Grassmann manifolds. In [163] the authors introduce *Gaussian Process Subspace Regression*, a Bayesian non-parametric model for subspace-valued mappings which is at the same time both extrinsic and intrinsic.

There has also been proposals for sampling from distributions on manifolds in other works. For example in [21] the authors propose a geodesic based HMC approach which can be used to sample from probability distributions defined on a manifold. In [86, 157] the authors investigate posteriors for cointegration estimation arising from placing priors defined on spaces related to the Grassmann manifold.

6.3 Ridge Approximation and Active Subspaces

In this section we provide details for *Optimal ridge approximation* of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ as well as for dimension reduction using the so-called *active subspace*. Our presentation here follows those in [36, 138]. We start by first introducing the concept of a *ridge function*, adopting the definition from [121]:

Definition 6.1 (*Ridge function*). A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is called a *ridge function* if there exists a matrix $\mathbf{M} \in \mathbb{R}^{d \times n}$, with $n \leq d$, and a function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ such that:

$$f(\mathbf{x}) = g(\mathbf{M}^T \mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d. \quad (6.3)$$

We call the subspace associated with the span of the columns of $\mathbf{M}$ its *ridge subspace*. ♦

Remark 6.1. As noted in [36] the term ridge function is more commonly used in the situation where $\mathbf{M}$ is a single vector (i.e. the case $n = 1$). In [121] our definition Definition 6.1 is called a *generalized ridge function*, while the term *generalized* is reserved for models where $\mathbf{M}$ depends on the particular $\mathbf{x}$ in [82]. ♦

Remark 6.2. If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a ridge function for some $\mathbf{M} \in \mathbb{R}^{d \times n}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ then it is constant along directions which are orthogonal to its ridge subspace, $\text{span}(\mathbf{M})$. To see this let $\mathbf{v} \in (\text{span}(\mathbf{M}))^\perp \subset \mathbb{R}^d$. One then has $\mathbf{M}^T \mathbf{v} = \mathbf{0}$ and so we can compute for $\alpha \in \mathbb{R}$ and any $\mathbf{x} \in \mathbb{R}^d$:

$$f(\mathbf{x} + \alpha \mathbf{v}) = g(\mathbf{M}^T(\mathbf{x} + \alpha \mathbf{v})) \quad (6.4)$$

$$= g(\mathbf{M}^T \mathbf{x} + \alpha \mathbf{M}^T \mathbf{v}) \quad (6.5)$$

$$= g(\mathbf{M}^T \mathbf{x}) \quad (6.6)$$

$$= f(\mathbf{x}). \quad (6.7)$$

Thus, f only varies along the n directions contained in its ridge subspace, where we have assumed, without loss of generality, that $\mathbf{M}$ is of full column-rank. In particular, we can assume that $\mathbf{M}$ has orthonormal columns, i.e. $\mathbf{M} \in \text{St}(d, n)$ and take $\mathbf{N} \in \text{St}(d, d - n)$ to be a matrix whose columns are an orthonormal basis for $(\text{span}(\mathbf{M}))^\perp$. By concatenating the columns of $\mathbf{M}$ and $\mathbf{N}$ into a single matrix $\mathbf{G} := [\mathbf{M}, \mathbf{N}] \in \mathbb{R}^{d \times d}$ we have that $\mathbf{G} \in O(d)$ i.e. $\mathbf{G}$ is an orthogonal matrix. Thus, $\mathbf{G}\mathbf{G}^T = \mathbf{I}_d$ and so we have $\mathbf{M}\mathbf{M}^T + \mathbf{N}\mathbf{N}^T = \mathbf{I}_d$ allowing us to decompose f as follows:

$$f(\mathbf{x}) = f(\mathbf{G}\mathbf{G}^T \mathbf{x}) = f(\mathbf{M}\mathbf{M}^T \mathbf{x} + \mathbf{N}\mathbf{N}^T \mathbf{x}) = f(\mathbf{M}\mathbf{y} + \mathbf{N}\mathbf{z}), \quad (6.8)$$

where $\mathbf{y} := \mathbf{M}^T \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} := \mathbf{N}^T \mathbf{x} \in \mathbb{R}^{d-n}$. Thus, we can view the pair $(\mathbf{y}, \mathbf{z})$ as a space decomposition and there is no redundancy or degeneracy in replacing $\mathbf{x}$ with $(\mathbf{y}, \mathbf{z})$. ♦

Optimal Ridge Approximation

Having introduced ridge functions we can now introduce the (optimal) ridge approximation problem as follows. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function we are interested in approximating by a ridge function and equip its domain with a probability measure encoded by a p.d.f. $\rho : \mathbb{R}^d \rightarrow [0, \infty)$. The p.d.f. serves the role of modelling any uncertainty we might have in the inputs to f (which typically represents some simulation model) which is a common modelling decision in the field of *uncertainty quantification* [148].

Remark 6.3. To illustrate the role of ρ more clearly, one can recall the example introduced in Section 6.1: we might be interested in estimating the parameter vector $\mathbf{p}$ based on noisy observations of the solution u . To capture our uncertainty in the parameter vector $\mathbf{p}$ we can model it as

being random with some associated density ρ . ♦

Given this function f and p.d.f. ρ the ridge approximation problem is to find a ridge function which minimizes the $L^2(\mathbb{R}^d, \rho)$ (see Remark 2.1) approximation error, i.e. the problem is to find $g : \mathbb{R}^n \rightarrow \mathbb{R}$ and a full rank $\mathbf{M} \in \mathbb{R}^{d \times n}$ which minimizes the following error:

$$\|f(\cdot) - g(\mathbf{M}^T \cdot)\|_{L^2(\mathbb{R}^d, \rho)} := \left(\int_{\mathbb{R}^d} (f(\mathbf{x}) - g(\mathbf{M}^T \mathbf{x}))^2 \rho(\mathbf{x}) d\mathbf{x} \right)^{\frac{1}{2}}. \quad (6.9)$$

As noted in [36] we can now restrict our attention to matrices $\mathbf{M} \in \text{St}(d, n)$ since if $\mathbf{M} \in \mathbb{R}^{d \times n}$ is of full column-rank we can utilise a thin QR factorization with the triangular $\mathbf{R}$ factor representing some invertible change of coordinates in the domain of g . We now let $\mathbf{N} \in \text{St}(d, d - n)$ be an orthogonal basis for the complement of $\text{span}(\mathbf{M})$ and introduce the coordinates $\mathbf{y} := \mathbf{M}^T \mathbf{x}$ and $\mathbf{z} := \mathbf{N}^T \mathbf{x}$ as in Remark 6.2. The p.d.f. now yields joint, marginal, and conditional densities on the coordinates $\mathbf{y}, \mathbf{z}$ of the subspaces $\text{span}(\mathbf{M})$ and $\text{span}(\mathbf{N})$ respectively. Following the notation in [36] we denote these densities as follows:

- $\pi(\mathbf{y}, \mathbf{z}) := \rho(\mathbf{M}\mathbf{y} + \mathbf{N}\mathbf{z})$ (joint density),
- $\pi(\mathbf{y}) := \int_{\mathbb{R}^{d-n}} \pi(\mathbf{y}, \mathbf{z}) d\mathbf{z}$ (marginal density),
- $\pi(\mathbf{z}|\mathbf{y}) := \pi(\mathbf{y}, \mathbf{z})/\pi(\mathbf{y})$ (conditional density).

Using the conditional density above allows us to write the conditional expectation of f given active coordinates $\mathbf{y}$ as [36] follows:

$$\tilde{g}(\mathbf{y}; \mathbf{M}) := \mathbb{E}(f|\mathbf{y}, \mathbf{M}) = \int_{\mathbb{R}^{d-n}} f(\mathbf{M}\mathbf{y} + \mathbf{N}\mathbf{z}) \pi(\mathbf{z}|\mathbf{y}) d\mathbf{z}. \quad (6.10)$$

As noted in [36] the ridge function $\tilde{g}(\mathbf{M}^T \mathbf{x}; \mathbf{M})$ satisfies, by construction:

$$\int_{\mathbb{R}^{d-n}} (\tilde{g}(\mathbf{y}; \mathbf{M}) - f(\mathbf{M}\mathbf{y} + \mathbf{N}\mathbf{z})) \pi(\mathbf{z}|\mathbf{y}) d\mathbf{z} = 0, \quad (6.11)$$

for all $\mathbf{y}$ for which $\pi(\mathbf{y}) > 0$. From this it follows as a consequence of Pinkus' Theorem 8.3 [121] that $\tilde{g}(\mathbf{M}^T \mathbf{x}; \mathbf{M})$ is the unique optimal ridge approximation for the norm Eq. (6.9).

Remark 6.4. A key point which is emphasized in both [138, 36] is the *orthogonal invariance* associated with $\mathbf{M}$, i.e. we are only interested in the particular subspace $\mathbf{M}$ represents and not in the particular choice of basis represented by a given $\mathbf{M} \in \text{St}(d, n)$. To see this Constantine et al.

first note that the conditional density is invariant upon replacing $\mathbf{M}$ with $\mathbf{M}\mathbf{Q}$ for any $\mathbf{Q} \in O(n)$:

$$\pi(\mathbf{z}|\mathbf{y} = \mathbf{Q}^T \mathbf{M}^T \mathbf{x}) = \frac{\rho(\mathbf{M}\mathbf{Q}\mathbf{Q}^T \mathbf{M}^T \mathbf{x} + \mathbf{N}\mathbf{z})}{\int \rho(\mathbf{M}\mathbf{Q}\mathbf{Q}^T \mathbf{M}^T \mathbf{x} + \mathbf{N}\mathbf{z}) d\mathbf{z}} \quad (6.12)$$

$$= \frac{\rho(\mathbf{M}\mathbf{M}^T \mathbf{x} + \mathbf{N}\mathbf{z})}{\int \rho(\mathbf{M}\mathbf{M}^T \mathbf{x} + \mathbf{N}\mathbf{z}) d\mathbf{z}} \quad (6.13)$$

$$= \pi(\mathbf{z}|\mathbf{y} = \mathbf{M}^T \mathbf{x}). \quad (6.14)$$

In a similar manner it is then shown that:

$$\tilde{g}(\mathbf{Q}^T \mathbf{M}^T \mathbf{x}; \mathbf{M}\mathbf{Q}) = \tilde{g}(\mathbf{M}^T \mathbf{x}; \mathbf{M}). \quad (6.15)$$

From these we finally get that the optimal ridge function $\tilde{g}$ depends only on the subspace $\text{span}(\mathbf{M})$ as remarked. This can be encapsulated by searching for a subspace in the Grassmann manifold (see Definition 5.11), $\mathcal{M} \in \text{Gr}(d, n)$, as opposed to searching for a particular basis representation $\mathbf{M} \in \text{St}(d, n)$. ♦

Active Subspaces

It is important to note that so far all we have established is that $\tilde{g}$ is the optimal ridge approximation for f given a **particular choice of ridge subspace** $\mathcal{M} = [\mathbf{M}] \in \text{Gr}(d, n)$. An important question is how one can estimate a good choice for this subspace given a limited computational budget for evaluating f and perhaps its gradients. This leads us to consider the so-called *active subspace* [33, 36] derived from gradient information of f as a particular choice for the subspace $\mathcal{M}$. We note that in this thesis we will be exclusively concerned with the active subspace as opposed to ridge approximation.

We now present a brief discussion of the active subspace following the presentations in [33, 138]. Suppose the function f is now assumed to be differentiable and Lipschitz continuous, and consider its gradient:

$$\nabla f(\mathbf{x}) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{pmatrix}. \quad (6.16)$$

Note that Lipschitz continuity of f implies that $\|\nabla f(\mathbf{x})\| \leq L$ for all $\mathbf{x} \in \mathbb{R}^d$ for some $L > 0$, where $\|\cdot\|$ is the standard Euclidean norm on $\mathbb{R}^d$. This allows us to define the expectation of the outer product of the gradient of f with itself as the following matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$:

$$\mathbf{C} := \mathbb{E}_{\mathbf{x} \sim \rho} [(\nabla f(\mathbf{x}))(\nabla f(\mathbf{x}))^T] = \int_{\mathbb{R}^d} (\nabla f(\mathbf{x}))(\nabla f(\mathbf{x}))^T \rho(\mathbf{x}) d\mathbf{x}. \quad (6.17)$$

The matrix $\mathbf{C}$ is symmetric and positive semi-definite since:

$$\mathbf{v}^T \mathbf{C} \mathbf{v} = \int_{\mathbb{R}^d} (\mathbf{v}^T \nabla f(\mathbf{x}))^2 \rho(\mathbf{x}) d\mathbf{x} \geq 0, \quad \forall \mathbf{v} \in \mathbb{R}^d. \quad (6.18)$$

It thus has a real eigenvalue decomposition with non-negative eigenvalues:

$$\mathbf{C} = \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T, \quad (6.19)$$

where $\mathbf{\Lambda} := \text{diag}(\{\lambda_i\}_{i=1}^d)$ contains the eigenvalues, $\lambda_1 \geq \dots \geq \lambda_d \geq 0$, and $\mathbf{W} \in O(d)$ contains the corresponding normalised eigenvectors as its columns.

The following result adopted from [33, Lemma 3.1] quantitatively describes the relationship between the gradient of f and the eigendecomposition of $\mathbf{C}$:

Lemma 6.1. The expected squared directional derivative of f with respect to the eigenvector $\mathbf{w}_i$ is equal to the corresponding eigenvalue, i.e.

$$\mathbb{E}_{\mathbf{x} \sim \rho} [((\nabla f(\mathbf{x}))^T \mathbf{w}_i)^2] = \int_{\mathbb{R}^d} ((\nabla f(\mathbf{x}))^T \mathbf{w}_i)^2 \rho(\mathbf{x}) d\mathbf{x} = \lambda_i, \quad i = 1, \dots, d. \quad (6.20)$$

◆

Note that Lemma 6.1 implies that if any of the eigenvalues are zero then the directional derivative of f with respect to the corresponding eigenvector is zero almost everywhere in the support of ρ . Since f is continuous it follows that f is constant along the direction specified by the eigenvector for the parameter values of interest (specified by the support of ρ). This is precisely the information which we can utilise to find a suitable subspace $\mathcal{M}$ where most of the variation in the function occurs. This observation leads us to the following idea from [33]: split the eigendecomposition of $\mathbf{C}$ into two groups as follows:

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \\ & \mathbf{\Lambda}_2 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{M} & \mathbf{N} \end{pmatrix}, \quad (6.21)$$

where $\mathbf{\Lambda}_1 = \text{diag}(\{\lambda_i\}_{i=1}^n)$ with $n < d$, and $\mathbf{M}$ contains the corresponding first n eigenvectors associated with $\mathbf{\Lambda}_1$. Due to the ordering of the eigenvalues of $\mathbf{C}$ the subspaces represented by $\mathbf{M}$ and $\mathbf{N}$ split the domain of f into two sets $\text{span}(\mathbf{M})$ and $\text{span}(\mathbf{N})$: perturbations of the inputs of f in $\text{span}(\mathbf{M})$ change f more than perturbations in $\text{span}(\mathbf{N})$ on average. The subspace $\text{span}(\mathbf{M})$ is called the *active subspace* while $\text{span}(\mathbf{N})$ is called the *inactive subspace*. Let $\mathbf{y} := \mathbf{m}^T \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} := \mathbf{N}^T \mathbf{x} \in \mathbb{R}^{d-n}$, as in Remark 6.2, be the coordinates in the active and inactive subspaces respectively. We will refer to $\mathbf{y}$ and $\mathbf{z}$ as the *active* and *inactive coordinates* respectively.

The theorem below, adopted from [138, 33], links the active subspace with the optimal ridge approximation $\tilde{g}$ given in Eq. (6.10):

Theorem 6.1. *Assuming $\mathbf{C}$ has the eigendecomposition Eq. (6.19) which is partitioned according to Eq. (6.21) we have that the optimal ridge approximation error on choosing the ridge subspace $\mathcal{M} = [\mathbf{M}]$ where $\mathbf{M}$ is as in Eq. (6.21) can be bounded as follows:*

$$\|f(\cdot) - \tilde{g}(\mathbf{M}^T \cdot)\|_{L^2(\mathbb{R}^d, \rho)} \leq C(\text{tr}(\mathbf{\Lambda}_2))^{\frac{1}{2}}, \quad (6.22)$$

where $C > 0$ is a constant that depends on ρ . ♦

Thus, a criterion for choosing n in the partition to determine the active subspace can be based off the decay of the eigenvalues of $\mathbf{C}$. As is noted in [33] each element of the matrix $\mathbf{C}$ involves d -dimensional integrals. When we are in the scenario where d is large and we wish to perform dimension reduction using the active subspace it is thus infeasible and/or costly to assemble $\mathbf{C}$ via, for example numerical integration, due to the infamous curse of dimensionality. The approach adopted in [33, Algorithm 3.1] is a Monte Carlo method for approximating $\mathbf{C}$ and its eigendecomposition which we present in Algorithm 5. This algorithm assumes that we have a means to compute the gradient of f .

Algorithm 5: Active subspace estimation via Monte Carlo

Input: Number of Monte Carlo samples N , function f , p.d.f. ρ

Output: Approximate eigendecomposition containing information on the active subspace

- 1 Sample N i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^N$ from the p.d.f. ρ ;
- 2 Compute $\nabla f(\mathbf{x}_i)$ for $i = 1, \dots, N$;
- 3 Form the following Monte Carlo approximation to $\mathbf{C}$:
- 4

$$\mathbf{C} \approx \hat{\mathbf{C}} := \frac{1}{N} \sum_{i=1}^N (\nabla f(\mathbf{x}_i))(\nabla f(\mathbf{x}_i))^T \quad (6.23)$$

;

- 5 Compute the eigendecomposition $\hat{\mathbf{C}} = \hat{\mathbf{W}}\hat{\mathbf{\Lambda}}\hat{\mathbf{W}}^T$.

- 6 return $\hat{\mathbf{W}}, \hat{\mathbf{\Lambda}}$
-

Remark 6.5. As noted above Algorithm 5 assumes we have a means of computing the gradient of f . The active subspace can be estimated even when we do not have a means of computing gradients. For example Algorithms 1.2 and 1.3 in [33] use local and global linear models respectively to perform this estimation. For ridge approximation methods have also been proposed which do not require the ability to compute gradients. These include methods based on finite differences [32, 54], as well as methods performing finite differencing via compressed sensing [59] and low-rank recovery [154]. Other methods utilise variable projection (VP) [67] to form a polynomial ridge approximation [74]. In this thesis we will assume that we always can compute gradients either

analytically, via automatic-differentiation or via some adjoint method. $\blacklozenge$

Remark 6.6. A useful heuristic when working with Algorithm 5 is to look for *spectral gaps*; a gap between successive eigenvalues indicates that the estimated active subspace is accurate. For a more detailed discussion on this the interested reader is directed to Chapter 3 of [33]. $\blacklozenge$

For further details on active subspaces the interested reader is directed to [33] and the references contained therein. More recently, in [62] the authors proposed a fully Bayesian gradient-free supervised dimension reduction method based on GPs which draws inspiration from the AS method described in this section.

6.4 Localised Active Subspaces

In the discussion in Section 6.3 it was assumed throughout that there is a **single** global active subspace of the input space of the simulation model, f , responsible for most of the variability in the outputs. As noted in Chapter 1 this assumption does not always hold in practice, especially when dealing with highly complex models which can exhibit diverse behaviour across different parameter regimes; a key example of such behaviour includes phase transitions. This observation leads us to instead consider *localised active subspaces* where we allow the subspace to vary as a function of some (or all) of the inputs to our simulation model. In this section we outline a possible approach for estimating localised active subspaces based on the approach outlined above in Section 6.3.

Suppose we decompose the original input space $\mathbb{R}^d$ into a Cartesian product $\mathbb{R}^d = \mathbb{R}^m \times \mathbb{R}^{d_p}$ where $m, d_p \in \mathbb{N}$ are such that $m + d_p = d$, i.e. we consider decomposing the original mapping f as $f : \mathbb{R}^m \times \mathbb{R}^{d_p} \rightarrow \mathbb{R}$. The input to the function f will now be the pair $(\mathbf{s}, \mathbf{x})$ where $\mathbf{s} \in \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^{d_p}$. We will view $\mathbf{s}$ as the parameters upon which our localised active subspace depends on, i.e. we are interested in studying how the active subspace for the remaining $\mathbb{R}^{d_p}$ variables varies as a function of $\mathbf{s}$. As was assumed in the discussion of active subspaces in Section 6.3 we now assume f is differentiable and Lipschitz continuous with respect to the second argument $\mathbf{x} \in \mathbb{R}^{d_p}$ and we assume that we have some p.d.f. $\rho : \mathbb{R}^{d_p} \rightarrow [0, \infty)$ for these inputs. Denote the gradient with respect to the second argument as follows:

$$\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{s}, \mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_{d_p}}(\mathbf{s}, \mathbf{x}) \end{pmatrix}. \quad (6.24)$$

We can similarly define the $\mathbf{s}$ -dependent expectation of the outer product of the gradient of f (with

respect to $\mathbf{x}$) with itself as the following parametrised matrix $\mathbf{C}(\mathbf{s}) \in \mathbb{R}^{d_p \times d_p}$:

$$\mathbf{C}(\mathbf{s}) := \mathbb{E}_{\mathbf{x} \sim \rho} [(\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}))(\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}))^T] = \int_{\mathbb{R}^{d_p}} (\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}))(\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}))^T \rho(\mathbf{x}) d\mathbf{x}. \quad (6.25)$$

Similarly as in Section 6.3 this matrix admits an eigendecomposition:

$$\mathbf{C}(\mathbf{s}) = \mathbf{W}(\mathbf{s})\mathbf{\Lambda}(\mathbf{s})\mathbf{W}(\mathbf{s})^T, \quad (6.26)$$

where $\mathbf{\Lambda}(\mathbf{s}) := \text{diag}(\{\lambda_i(\mathbf{s})\}_{i=1}^{d_p})$ contains the $\mathbf{s}$ -dependent eigenvalues, $\lambda_1(\mathbf{s}) \geq \dots \geq \lambda_{d_p}(\mathbf{s}) \geq 0$, and $\mathbf{W}(\mathbf{s}) \in O(d)$ contains the corresponding normalised $\mathbf{s}$ -dependent eigenvectors as its columns. We can again partition the eigendecomposition of $\mathbf{C}(\mathbf{s})$ into two groups:

$$\mathbf{\Lambda}(\mathbf{s}) = \begin{pmatrix} \mathbf{\Lambda}_1(\mathbf{s}) & \\ & \mathbf{\Lambda}_2(\mathbf{s}) \end{pmatrix}, \quad \mathbf{W}(\mathbf{s}) = \begin{pmatrix} \mathbf{M}(\mathbf{s}) & \mathbf{N}(\mathbf{s}) \end{pmatrix}, \quad (6.27)$$

where $\mathbf{\Lambda}_1(\mathbf{s}) = \text{diag}(\{\lambda_i(\mathbf{s})\}_{i=1}^n)$ with $n < d_p$, and $\mathbf{M}(\mathbf{s})$ contains the corresponding first n eigenvectors associated with $\mathbf{\Lambda}_1(\mathbf{s})$. We call the subspace $\text{span}(\mathbf{M}(\mathbf{s}))$ the *localised active subspace* associated to f . The cut-off for the active subspace dimension n can again be chosen based off the decay of the eigenvalues of $\mathbf{C}(\mathbf{s})$. Note that in principle this dimension can be $\mathbf{s}$ -dependent however for the purposes of this thesis we will make the assumption that it remains constant for all $\mathbf{s}$ values of interest.

When we are in the regime of a large number of parameters d_p upon which we wish to perform dimension reduction we again cannot directly compute the matrix $\mathbf{C}(\mathbf{s})$. We can again turn to a Monte Carlo approach to estimate this matrix, and hence the localised active subspace. This analogue of Algorithm 5 is presented in Algorithm 6.

Algorithm 6: Localised active subspace estimation via Monte Carlo

Input: Number of Monte Carlo samples N , function f , p.d.f. ρ , location $\mathbf{s}$

Output: Approximate eigendecomposition containing information on the localised active subspace at $\mathbf{s}$

- 1 Sample N i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^N$ from the p.d.f. ρ ;
- 2 Compute $\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}_i)$ for $i = 1, \dots, N$;
- 3 Form the following Monte Carlo approximation to $\mathbf{C}(\mathbf{s})$:
- 4

$$\mathbf{C}(\mathbf{s}) \approx \hat{\mathbf{C}}(\mathbf{s}) := \frac{1}{N} \sum_{i=1}^N (\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}_i))(\nabla_{\mathbf{x}} f(\mathbf{s}, \mathbf{x}_i))^T \quad (6.28)$$

;

- 5 Compute the eigendecomposition $\hat{\mathbf{C}}(\mathbf{s}) = \hat{\mathbf{W}}(\mathbf{s})\hat{\mathbf{\Lambda}}(\mathbf{s})\hat{\mathbf{W}}(\mathbf{s})^T$.
 - 6 **return** $\hat{\mathbf{W}}(\mathbf{s}), \hat{\mathbf{\Lambda}}(\mathbf{s})$
-

We can thus utilise Algorithm 6 to approximate the localised active subspace at any points

$\{\mathbf{s}_j\}_{j=1}^k$ of interest. However, when we are in the scenario of a limited computational budget Algorithm 6 will only be able to compute this localised active subspace at a small number of locations (small k). This motivates the following idea which underlies the work in this chapter: given our limited computational budget we first estimate the localised active subspace at a small number of locations yielding the collection $\{\mathbf{M}(\mathbf{s}_j)\}_{j=1}^k$ of representatives of these subspaces, i.e. $\mathbf{M}(\mathbf{s}_j)$ represents a subspace, $[\mathbf{M}(\mathbf{s}_j)]$, which is a point on the Grassmann manifold $\text{Gr}(d_p, n)$. If we now want to obtain an estimate for the localised active subspace at some new point $\tilde{\mathbf{s}}$ we can fit a probabilistic model to approximate the mapping from $\mathbf{s}$ to the localised active subspace, i.e. we can view the computed pairs $\{(\mathbf{s}_j, [\mathbf{M}(\mathbf{s}_j)])\}_{j=1}^k$ as evaluations of some subspace-valued mapping $P : \mathbb{R}^m \rightarrow \text{Gr}(d_p, n)$ and fit a probabilistic model using these observed evaluations to allow us to predict the value at the new location $\tilde{\mathbf{s}}$. For this purpose we develop a probabilistic model based on GPs which we can use to interpolate a field of Grassmann-valued response variables. This model is presented in Section 6.5 and we then present two numerical examples illustrating the model's use in approximating localised active subspaces in Sections 6.7.2 and 6.7.3.

6.5 Grassmann Valued Gaussian Processes

In this section we will derive a probabilistic model based on GPs which can be used to interpolate a field of Grassmann-valued response variables. To be more specific we will develop a probabilistic model capable of approximating a subspace-valued mapping, $P : D \rightarrow \text{Gr}(d, n)$ where D is a subset $D \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$. Since $\text{Gr}(d, n)$ is not a Euclidean space we cannot directly apply GPs to model the mapping P . We instead exploit the Riemannian geometry of $\text{Gr}(d, n)$ to allow us to apply a GP based model to a linear tangent space of the manifold. In particular, under the assumption that the dispersion of the response variable on the manifold is not too large, we can map the response via the Grassmann logarithm, $\log_{\mathcal{X}}^{\text{Gr}}$, to the tangent space $T_{\mathcal{X}} \text{Gr}(d, n)$, perform interpolation on the linear tangent space via a GP model and then map back to the manifold via the Grassmann exponential map, $\exp_{\mathcal{X}}^{\text{Gr}}$. One important question which remains is which point on the manifold $\mathcal{X} \in \text{Gr}(d, n)$ shall we choose to anchor the tangent space at? In this thesis we will take $\mathcal{X}$ to be the Karcher mean of the observed response variables. This will be motivated in the derivation below.

6.5.1 Problem Formulation

We assume we have a dataset consisting of $N \in \mathbb{N}$ observations of a $\text{Gr}(d, n)$ -valued mapping, i.e. we assume we are given a dataset:

$$\mathcal{D}_N := \{(\mathbf{s}_i, \mathcal{W}_i)\}_{i=1}^N, \quad (6.29)$$

where $\{\mathbf{s}_i\}_{i=1}^N \subset D$ are our N input points and the $\{\mathcal{W}_i\}_{i=1}^N \subset \text{Gr}(d, n)$ are the corresponding observations of $\{P(\mathbf{s}_i)\}_{i=1}^N$, i.e. $\mathcal{W}_i = P(\mathbf{s}_i)$. Each of the Grassmann-valued observations can be represented using a matrix in the Stiefel manifold via $\mathcal{W}_i = [\mathbf{W}_i]$, $\mathbf{W}_i \in \text{St}(d, n)$ as is explained in Section 5.4.3. Thus, the dataset can be represented more concretely as:

$$\mathcal{D}_N := \{(\mathbf{s}_i, [\mathbf{W}_i])\}_{i=1}^N. \quad (6.30)$$

Our problem is now to learn a probabilistic model for the unobserved map P from the dataset $\mathcal{D}_N$. We now outline in detail the steps we take to obtain our probabilistic model.

Remark 6.7. To help link this statistical problem to an application it is again useful to recall the example introduced in Section 6.1. The dataset $\mathcal{D}_N$ for this example will be the set of tuples $\{(\mathbf{x}_i, \mathcal{M}_i)\}_{i=1}^N$ where $\mathcal{M}_i$ is the lower-dimensional subspace representing the localised active subspace for the $\mathbf{p}$ -dependence of the solution u of the parametrised elliptic PDE at the spatial location $\mathbf{x}_i$. ♦

6.5.2 Lifting of the Dataset to the Tangent Space

We will utilise the Riemannian geometry of the Grassmann manifold described in Section 5.4.3 to lift the dataset $\mathcal{D}_N$ to a tangent space to the manifold.

We start by computing the Karcher mean (see Definition 5.9), $\overline{\mathcal{W}}_N$, of our observed response variables:

$$\overline{\mathcal{W}}_N = \underset{[\mathbf{X}] \in \text{Gr}(d, n)}{\text{argmin}} \sum_{i=1}^N d_{\text{Gr}}([\mathbf{W}_i], [\mathbf{X}])^2, \quad (6.31)$$

where $d_{\text{Gr}} : \text{Gr}(d, n) \times \text{Gr}(d, n) \rightarrow \mathbb{R}$ is the Riemannian distance on the Grassmann manifold (see Definition 5.6). We will utilise the Karcher mean as our anchor point $\mathcal{X} = \overline{\mathcal{W}}_N$ to anchor our chosen tangent space. This is a sensible choice for the anchor point under the assumption that the dispersion of the observed response variables on the manifold is not too large. In what follows we will assume $\mathcal{X} = \overline{\mathcal{W}}_N$ unless otherwise stated.

We now utilise the Riemannian Grassmann logarithm to lift the response variables to the

tangent space $T_{\mathcal{X}} \text{Gr}(d, n)$ via:

$$\mathbf{U}_i := \log_{\mathcal{X}}^{\text{Gr}}([\mathbf{W}_i]), \quad i = 1, \dots, N. \quad (6.32)$$

This yields our lifted dataset:

$$\tilde{\mathcal{D}}_N := \{(\mathbf{s}_i, \mathbf{U}_i)\}_{i=1}^N. \quad (6.33)$$

6.5.3 Gaussian Process Model on the Tangent Space

Having lifted our dataset to the linear tangent space $T_{\mathcal{X}} \text{Gr}(d, n)$ we can now start building our GP model on this space. As discussed in Section 5.4.3 we can concretely represent tangent vectors to the Grassmann manifold via matrices in the horizontal space associated to the tangent space. Thus, we can view each $\mathbf{U}_i$ as a $d \times n$ matrix, $\mathbf{U}_i \in \mathbb{R}^{d \times n}$, which satisfies the constraint:

$$\mathbf{X}^T \mathbf{U}_i = \mathbf{0}, \quad (6.34)$$

where here $\mathbf{X} \in \text{St}(d, n)$ is an element of the Stiefel manifold which represents our anchor point's equivalence class, i.e. $\mathcal{X} = [\mathbf{X}]$.

In order to build a GP model for the lifted dataset $\tilde{\mathcal{D}}_N$ in Eq. (6.33) we are thus faced with two problems:

1. The lifted response variables, $\{\mathbf{U}_i\}_{i=1}^N$, are matrix-valued.
2. The lifted response variables are not free to vary in $\mathbb{R}^{d \times n}$, but are instead required to satisfy the constraint given by Eq. (6.34).

The first problem above can be solved by working with matrix-valued GPs, which we described in Section 5.2. The second problem is more nefarious since it is can be difficult to utilise GP models for data constrained to lie in a closed subspace of a vector space. Luckily, we can again exploit the geometry of the Grassmann manifold and parametrise the tangent vectors $\mathbf{U}_i$ using Remark 5.15 which allows us to write:

$$\mathbf{U}_i = (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{V}_i, \quad \mathbf{V}_i \in \mathbb{R}^{d \times n}, \quad (6.35)$$

where each $\mathbf{V}_i$ is arbitrary and is thus allowed to vary freely in $\mathbb{R}^{d \times n}$. We will thus build a GP model for the $\{\mathbf{V}_i\}_{i=1}^N$ and then simply project them to the tangent space using the projector $(\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)$ via Eq. (6.35), i.e. we introduce the random matrix-valued function $\mathbf{V} : D \rightarrow \mathbb{R}^{d \times n}$ and stipulate that it follows the matrix-valued GP,

$$\mathbf{V} \sim \text{Mat } \mathcal{GP}(\boldsymbol{\mu}, k, \boldsymbol{\Omega}). \quad (6.36)$$

We can then transfer this random function to the tangent space $T_{\mathcal{X}} \text{Gr}(d, n)$ via:

$$\mathbf{U}(\mathbf{s}) := (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{V}(\mathbf{s}), \quad (6.37)$$

yielding a random function $\mathbf{U} : D \rightarrow T_{\mathcal{X}} \text{Gr}(d, n)$. Note that this is simply an instance of the well-known *reparametrisation trick* [85].

6.5.4 GrassGP Model

We are now in a position to introduce and define what we call a *Grassmann-valued Gaussian Process*:

Definition 6.2 (*Grassmann-valued Gaussian Process*). Let $D \subset \mathbb{R}^k$ be an subset. For a given positive definite kernel function $k : D \times D \rightarrow \mathbb{R}$, a positive semi-definite *parameter covariance matrix* $\mathbf{\Omega} \in \mathbb{R}^{dn \times dn}$, and a function $\boldsymbol{\mu} : D \rightarrow \mathbb{R}^{d \times n}$ we say that a random Grassmann-valued function $P : D \rightarrow \text{Gr}(d, n)$ is a *Grassmann-valued Gaussian Process (GrassGP)*, if it has the following form:

$$P(\mathbf{s}) = \exp_{[\mathbf{X}]}^{\text{Gr}}(\mathbf{U}(\mathbf{s})), \quad \forall \mathbf{s} \in D, \quad (6.38)$$

where $\mathbf{X} \in \text{St}(d, n)$ is a representative of the *anchor point* $\mathcal{X} = [\mathbf{X}] \in \text{Gr}(d, n)$ and the random function $\mathbf{U} : D \rightarrow T_{[\mathbf{X}]} \text{Gr}(d, n)$ has the form:

$$\mathbf{U}(\mathbf{s}) = (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{V}(\mathbf{s}), \quad \forall \mathbf{s} \in D, \quad (6.39)$$

where the random function $\mathbf{V} : D \rightarrow \mathbb{R}^{d \times n}$ follows the matrix-valued GP distribution:

$$\mathbf{V} \sim \text{Mat } \mathcal{GP}(\boldsymbol{\mu}, k, \mathbf{\Omega}). \quad (6.40)$$

We write:

$$P \sim \text{Gr } \mathcal{GP}(\boldsymbol{\mu}, k, \mathbf{\Omega}, \mathbf{X}), \quad (6.41)$$

to denote the Grassmann-valued Gaussian Process with mean function $\boldsymbol{\mu}$, kernel function k , parameter covariance matrix $\mathbf{\Omega}$ and anchor point $[\mathbf{X}]$. ♦

The GrassGP model will be used as our prior model for the dataset $\mathcal{D}_N$ as will be expanded on in Section 6.6 below. Due to the GP structure on the fixed tangent space we expect this prior to be able to flexibly capture a wide range of input-output dependency structures.

6.6 Inference using GrassGP model

In this section we describe how to utilise the GrassGP model as a prior to perform Bayesian inference to learn the unknown function P from the dataset $\mathcal{D}_N$. Any Bayesian inference requires the specification of the following two ingredients:

- A prior distribution - this will be our GrassGP model.
- A likelihood - this will be detailed in Section 6.6.1 below.

6.6.1 Choice of Likelihood

A crucial ingredient in any Bayesian inference is the choice of likelihood. Our GrassGP model essentially specifies a prior over the tangent vectors $\{\mathbf{U}_i\}_{i=1}^N$ where each $\mathbf{U}_i$ is specified in terms of the corresponding $\mathbf{W}_i$ via Eq. (6.32). Since each $\mathbf{U}_i$ lies in a Euclidean space it is sensible to specify a likelihood stipulating that we observe each tangent vector with i.i.d. Gaussian noise, i.e. the likelihood has the following form:

$$p(\mathcal{W}_1, \dots, \mathcal{W}_N | \mathbf{U}_1, \dots, \mathbf{U}_N, \sigma) = \prod_{i=1}^N p(\mathcal{W}_i | \mathbf{U}_i, \sigma), \quad (6.42)$$

where each $p(\mathcal{W}_i | \mathbf{U}_i, \sigma)$ is as follows:

$$p(\mathcal{W}_i | \mathbf{U}_i) \propto \exp \left(- \frac{\|\mathbf{U}_i - \log_{[\mathcal{X}]}^{\text{Gr}}(\mathcal{W}_i)\|_{\mathcal{X}}^2}{2\sigma^2} \right), \quad (6.43)$$

and where $\sigma > 0$ is a parameter specifying the noise level.

Note that the Riemannian norm $\|\cdot\|_{\mathcal{X}}$ on the tangent space $T_{\mathcal{X}} \text{Gr}(d, n)$ is simply the Frobenius norm and is given by:

$$\|\Delta\|_{\mathcal{X}}^2 = \text{tr}(\Delta^T \Delta), \quad \forall \Delta \in T_{\mathcal{X}} \text{Gr}(d, n), \forall \mathcal{X} \in \text{Gr}(d, n). \quad (6.44)$$

Remark 6.8. Note that the likelihood specified by Eqs. (6.42) and (6.43) is equivalent to stipulating:

$$\log_{\mathcal{X}}^{\text{Gr}}(\mathcal{W}_i) \stackrel{\text{i.i.d.}}{\sim} \mathcal{MN}(\mathbf{U}_i, \mathbf{I}_d, \sigma^2 \mathbf{I}_n), \quad i = 1, \dots, N, \quad (6.45)$$

i.e. that each tangent vector is distributed as i.i.d. matrix normal random variables located at the observed $\mathbf{U}_i$ with noise level given by σ . ♦

Remark 6.9. Note that a naive approach to performing inference for subspace-valued response variables would be to simply consider Stiefel representatives of the subspace as elements of the ambient Euclidean space of real-valued matrices and simply introduce a likelihood in this space.

However, this would not guarantee that we produce valid projectors on the Grassmann manifold. As such we adopt the approach described here of introducing the likelihood on the tangent space of the Grassmann manifold and then transferring the tangent vectors to the manifold via the Riemannian exponential map. $\blacklozenge$

We also note that the introduction of stochasticity into our problem is justified for the following reasons:

- The computation of our dataset of localised active subspaces via Algorithm 6 relies on a Monte Carlo approach introducing uncertainty into the dataset.
- When we lift the dataset to the tangent space we choose one particular anchor point $\mathcal{X} \in \text{Gr}(d, n)$, in this thesis we take this to be the Karcher mean. Due to the fixed choice of tangent space it is possible that we introduce some model-misspecification via distortions for points which are not close enough to the Karcher mean. This introduces further uncertainty into our model which can be captured via a Gaussian likelihood.

6.6.2 The Full Bayesian Model

In this section we will specify the full Bayesian model which we will use to learn the unknown Grassmann-valued function, P , from the dataset $\mathcal{D}_N$. Before we can present this model we must first clearly list all the unknown parameters which are required to fully specify both the prior and likelihood:

- **Prior mean function, μ :** in this work we will always take $\mu = \mathbf{0}$ identically. This is a sensible choice when working with GP based priors.
- **Prior kernel function, k :** in this work we will take the kernel function to be from a fixed parametric family, $\{k_\ell\}_\ell$, parametrised by the vector ℓ . We will specify the particular family used for each example in Section 6.7. Thus, the unknown parameters related to the kernel function are specified by the vector ℓ and will be modelled via a hierarchical prior which will be specified for each example presented in Section 6.7.
- **Prior parameter covariance matrix, Ω :** this matrix parameter will be modelled via some parametric form; usually as a diagonal matrix. The particular form, together with its hierarchical prior, will be specified for each example presented in Section 6.7.
- **Prior anchor point, $\mathcal{X} = [\mathbf{X}]$:** as described in Section 6.5.2 this will be taken to be the Karcher mean of the observed response variables $\mathcal{X} = \overline{\mathcal{W}}_N$, where $\overline{\mathcal{W}}_N$ is specified by Eq. (6.31).

- **Likelihood noise level**, σ : this will either be fixed to a constant or will be modelled directly by specifying a hierarchical prior on σ . The particular value or prior will be specified for each example in Section 6.7.

We can now describe the full Bayesian model we are proposing:

$$\sigma, \boldsymbol{\Omega}, \boldsymbol{\ell} \sim p(\sigma)p(\boldsymbol{\Omega})p(\boldsymbol{\ell}), \quad (6.46)$$

$$P = \exp_{[\mathbf{X}]}^{\text{Gr}}(\mathbf{U}) | \{\boldsymbol{\ell}, \boldsymbol{\Omega}\} \sim \text{Gr } \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\ell}}, \boldsymbol{\Omega}, \mathbf{X}), \quad (6.47)$$

$$\log_{[\mathbf{X}]}^{\text{Gr}}(\mathcal{W}_i) | \{\mathbf{U}_i, \sigma\} \stackrel{\text{i.i.d.}}{\sim} \mathcal{MN}(\mathbf{U}_i, \mathbf{I}_d, \sigma^2 \mathbf{I}_n), \quad i = 1, \dots, N, \quad (6.48)$$

where the $\{\mathbf{U}_i\}_{i=1}^N$ are defined via $\mathbf{U}_i = \log_{[\mathbf{X}]}^{\text{Gr}}(P(\mathbf{s}_i))$ for $i = 1, \dots, N$ and $p(\sigma), p(\boldsymbol{\Omega}), p(\boldsymbol{\ell})$ are specified hierarchical priors for the unknown parameters, i.e. we are postulating that the observed response variables $\{\mathcal{W}_i\}_{i=1}^N$ in Eq. (6.29) are modelled as being drawn from the Bayesian model specified by Eqs. (6.46) to (6.48).

Thus, given the dataset $\mathcal{D}_N$ we wish to obtain the posterior $\{\sigma, \boldsymbol{\Omega}, \boldsymbol{\ell}\} | \mathcal{D}_N$ corresponding to the Bayesian model Eqs. (6.46) to (6.48) and using this posterior distribution we wish to make predictions for $P(\tilde{\mathbf{s}})$ at unobserved test points $\tilde{\mathbf{s}} \in D$. The procedure for making predictions will be explained in Section 6.6.3 below.

6.6.3 Prediction

In this section we will describe how to make predictions with our probabilistic model for new unobserved test points. Let $\tilde{\mathbf{s}}$ be a new input point in D . We wish to make a prediction for $P(\tilde{\mathbf{s}})$. Provided we can sample from the marginal posterior distribution $\{\boldsymbol{\Omega}, \boldsymbol{\ell}\} | \mathcal{D}_N$ we can make a prediction for $P(\tilde{\mathbf{s}})$ as follows:

1. Sample hyper-parameters, $\boldsymbol{\Omega}, \boldsymbol{\ell}$, from the marginal posterior distribution, $\{\boldsymbol{\Omega}, \boldsymbol{\ell}\} | \mathcal{D}_N$.
2. Make a prediction for $P(\tilde{\mathbf{s}})$ from the Grassmann-valued GP, $\text{Gr } \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\ell}}, \boldsymbol{\Omega}, \mathbf{X})$, with the sampled values of $\boldsymbol{\Omega}, \boldsymbol{\ell}$, conditioned on the dataset $\mathcal{D}_N$.

We will utilise HMC to obtain samples from the hyper-parameter posterior distributions needed for step 1 above. Generating predictions from the Grassmann-valued GP in step 2 proceeds as follows. Since P has the form given by Eq. (6.38) where $\mathbf{U}$ is specified by Eqs. (6.39) and (6.40) with $\boldsymbol{\mu} = \mathbf{0}$ and $k = k_{\boldsymbol{\ell}}$ we can view $\mathbf{U}$ itself as a matrix-valued GP as the following result states:

Proposition 6.1. *If $\mathbf{U} : D \rightarrow T_{[\mathbf{X}]} \text{Gr}(d, n)$ is given by Eq. (6.39) where*

$$\mathbf{V} \sim \text{Mat } \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\ell}}, \boldsymbol{\Omega}, \mathbf{X}), \quad (6.49)$$

then $\mathbf{U}$ is itself a matrix-valued GP:

$$\mathbf{U} \sim \text{Mat } \mathcal{GP}(\mathbf{0}, k_{\ell}, \mathbf{\Omega}_{\mathbf{X}}), \quad (6.50)$$

where $\mathbf{\Omega}_{\mathbf{X}}$ is the following matrix:

$$\mathbf{\Omega}_{\mathbf{X}} = (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{\Omega}(\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)^T. \quad (6.51)$$

◆

Proof of Proposition 6.1.

This follows from Definition 5.3 together with properties of the Matrix Normal distribution presented in Section 2.3.1. In order to show that $\mathbf{U}$ follows the matrix-valued GP given by Eq. (6.50) we must show, via Definition 5.3, that for any $N \in \mathbb{N}$ and for any $S := \{\mathbf{s}_i\}_{i=1}^N \subset D$ we have:

$$\text{vec}(\mathbf{U}_S) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{SS} \otimes \mathbf{\Omega}_{\mathbf{X}}), \quad (6.52)$$

where $\mathbf{U}_S, \mathbf{K}_{SS}$ are defined similarly as in Eq. (5.9), with the kernel function $k = k_{\ell}$. Since $\mathbf{U}(\mathbf{s})$ is given by an affine transformation of the matrix-valued GP, $\mathbf{V}$, via Eq. (6.39) we have, again by Definition 5.3, that:

$$\mathbf{V}_S \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Omega}, \mathbf{K}_{SS}). \quad (6.53)$$

Noting that $\mathbf{U}_S = (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{V}_S$ we have, via Proposition 5.1,

$$\mathbf{U}_S \sim \mathcal{MN}(\mathbf{0}, (\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)\mathbf{\Omega}(\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)^T, \mathbf{K}_{SS}), \quad (6.54)$$

i.e. $\mathbf{U}_S \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Omega}_{\mathbf{X}}, \mathbf{K}_{SS})$ and so Eq. (6.52) holds as required. ■

We can thus make predictions from the Grassmann-valued GP by first making predictions for $\mathbf{U}(\tilde{\mathbf{s}})$ from the matrix-valued GP Eq. (6.50) as outlined in Section 5.2. This prediction can then be pushed onto the Grassmann manifold to yield a prediction for $P(\tilde{\mathbf{s}})$ via the Grassmann Riemannian exponential map, $P(\tilde{\mathbf{s}}) = \exp_{[\mathbf{X}]}^{\text{Gr}}(\mathbf{U}(\tilde{\mathbf{s}}))$.

Remark 6.10. Note that the parameter covariance matrix $\mathbf{\Omega}_{\mathbf{X}}$ in the distribution of $\mathbf{U}$, Eq. (6.50) is singular since the matrix $(\mathbf{I}_d - \mathbf{X}\mathbf{X}^T)$ is an orthogonal projection onto the tangent space $T_{[\mathbf{X}]} \text{Gr}(d, n)$. Thus, we are in the special case of a positive semi-definite covariance matrix for the resulting multivariate normal distribution and so we must use the pseudo-inverse in the conditioning formula as detailed in Section 5.2. ◆

6.7 Numerical Experiments

In this section we will present some numerical experiments demonstrating the use of our GrassGP model. In Section 6.7.1 we will consider a low-dimensional example which we will use to both demonstrate and visualise the performance of our GrassGP model. Then in Sections 6.7.2 and 6.7.3 we will present two high-dimensional examples which illustrate the use of our GrassGP model for studying localised active subspaces as described in Section 6.4 for a parametrised ODE and PDE respectively.

The code for these numerical experiments is available online¹.

6.7.1 Visualisation of GrassGP Model

Dataset Description

In this section we will present a numerical experiment which will be used to both demonstrate and visualise our GrassGP model. This example is adopted from an example presented in [163, (see Section 6.1)] which was used to visualise the GP based subspace regression model the authors Zhang et al. considered. This example involves learning perhaps the simplest case of a subspace-valued mapping, $P : D \rightarrow \text{Gr}(2, 1)$, where D is a subset of $\mathbb{R}$. As the authors Zhang et al. note, such a mapping is the simplest type of Grassmann-valued function mapping a real number to a 1-dimensional linear subspace of the plane, $\mathbb{R}^2$. Note that a 1-d linear subspace of the plane, i.e. a line, can be uniquely identified by the angle made with the positive x -axis, $\theta \in [0, \pi)$. We can thus identify $\text{Gr}(2, 1)$ with the unit circle in the plane, S^1 , by treating antipodal points as equivalent. This equivalence is illustrated in Fig. 6.1. The low dimensions involved allow us to easily plot such a mapping which will allow us to visualise the performance of our Bayesian model more easily.

In particular, we will apply our GrassGP model to learn the following function:

$$\begin{aligned} P : [0, 1] &\rightarrow \text{Gr}(2, 1), \\ s &\mapsto P(s) := [\mathbf{v}(s)], \end{aligned} \tag{6.55}$$

where the unit vector $\mathbf{v}(s)$ is defined via,

$$\mathbf{v}(s) := \begin{pmatrix} \cos(2\pi s) \\ \sin(2\pi s) \end{pmatrix}. \tag{6.56}$$

Note that $P(s)$ indeed lies in $\text{Gr}(2, 1)$, since the vector $\mathbf{v}(s)$ lies on the unit circle in $\mathbb{R}^2$ for all $s \in [0, 1]$.

¹<https://github.com/YanniPapandreou/grassgp>

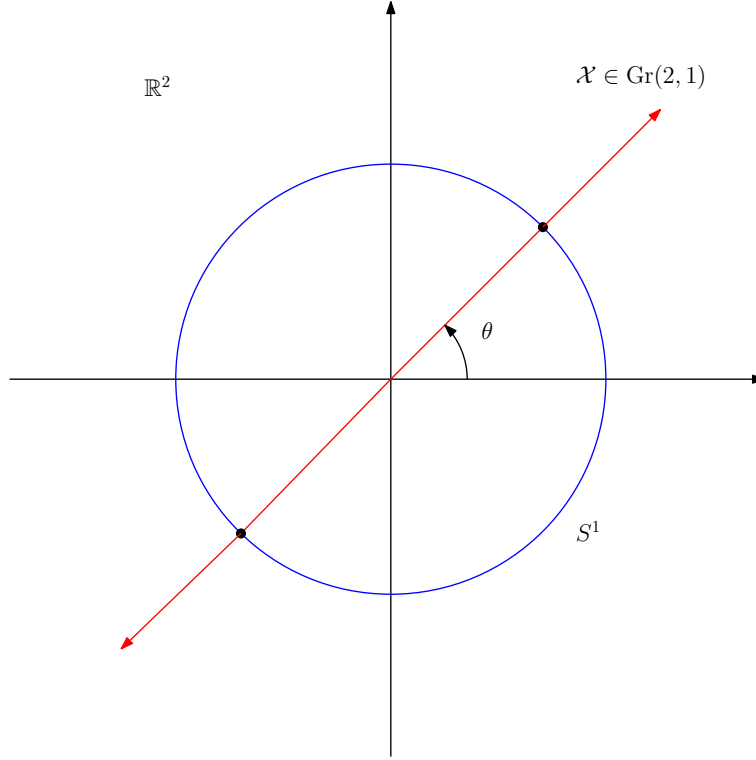


Figure 6.1: Illustration of equivalence of $\text{Gr}(2, 1)$ with the unit circle in the plane S^1 . The red ray denotes an element, $\mathcal{X}$, of $\text{Gr}(2, 1)$, i.e. a 1 dimensional subspace of the plane. This subspace can be uniquely identified with two antipodal points on the unit circle S^1 . By treating these antipodal points as equivalent we get a unique representation of $\mathcal{X}$ via an angle $\theta \in [0, \pi)$.

Remark 6.11. Note that the mapping Eq. (6.55) can be equivalently viewed as the following mapping:

$$\begin{aligned} \alpha : [0, 1] &\rightarrow [0, \pi), \\ s &\mapsto \alpha(s) := (2\pi s) \bmod \pi, \end{aligned} \tag{6.57}$$

where we represent the point $P(s) \in \text{Gr}(2, 1)$ by the angle the subspace makes with the positive x -axis in the plane, $\alpha(s) \in [0, \pi)$, where we treat antipodal points as equivalent. Thus, this function corresponds to a line in the plane which is being rotated about the origin with an angular velocity of 2π . ♦

We will use our GrassGP model to learn this function given $N = 14$ observations of $P(s)$ at equally spaced locations, $\{s_i\}_{i=1}^{14}$, between 0 and 1 inclusive, i.e. our dataset is:

$$\mathcal{D}_{14} = \{(s_i, P(s_i))\}_{i=1}^{14} = \{(s_i, [\mathbf{v}(s_i)])\}_{i=1}^{14}, \tag{6.58}$$

where $s_1 = 0, s_{14} = 1$. We augment the observed dataset $\mathcal{D}_{14}$ with an additional 26 points to yield a test dataset,

$$\mathcal{D}_{40}^* = \{(s_i^*, P(s_i^*))\}_{i=1}^{40} = \{(s_i^*, [\mathbf{v}(s_i^*)])\}_{i=1}^{40}, \tag{6.59}$$

which will be compared with predictions made from the model. Note that we first generate the test

dataset Eq. (6.59) and then take every other 3rd point to obtain our training dataset Eq. (6.58). Thus, the testing dataset also contains training points where we expect the prediction errors to be smaller. Figure 6.2 shows the training and test datasets Eqs. (6.58) and (6.59) in various representations.

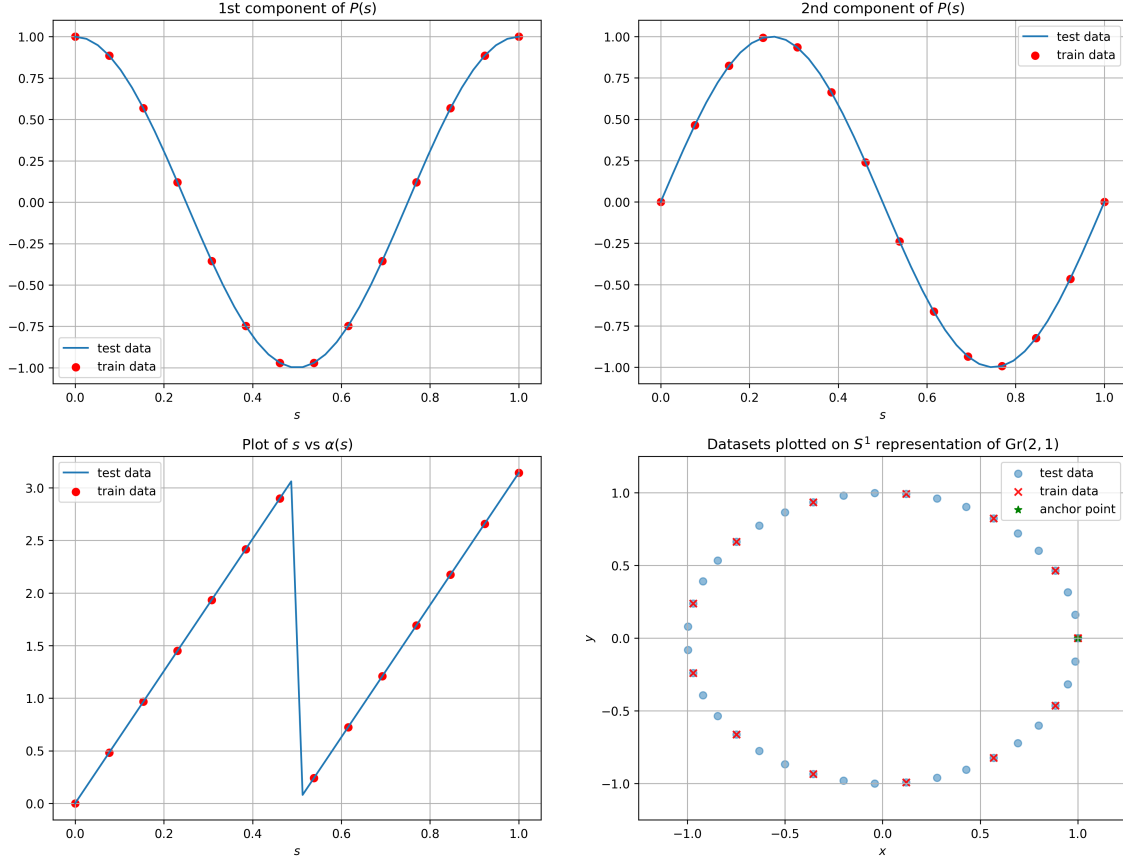


Figure 6.2: Plots of train and test datasets in various representations. *Top row*: plots of the components of $P(s)$. *Bottom row*: (left) plot of $\alpha(s)$, where $\alpha(s)$ is the angle the subspace $P(s)$ makes with positive x -axis, (right) plot of datasets on the unit circle representation of $\text{Gr}(2,1)$ together with the anchor point.

Experiment Parameters

We now detail the various choice we make for the unknown parameters described in Section 6.6.2 for this experiment. Note that anything not specified explicitly is assumed to be the same as described in Section 6.6.2.

- **Prior kernel function:** We choose to work with the family of Squared Exponential (SE) kernels:

$$k_\ell(x, y) := \exp\left(-\frac{1}{2\ell^2}|x - y|^2\right), \quad (6.60)$$

parameterised by the length scale, $\ell > 0$. We place a log-normal hierarchical prior on the

length scale:

$$\ell \sim \text{Log } \mathcal{N}(0, 0.5^2). \quad (6.61)$$

- **Prior parameter covariance matrix:** We will use the following parametric form for $\mathbf{\Omega} \in \mathbb{R}^{2 \times 2}$:

$$\mathbf{\Omega} := (\boldsymbol{\sigma} \boldsymbol{\sigma}^T) \odot \mathbf{L}, \quad (6.62)$$

where $\boldsymbol{\sigma} \in \mathbb{R}^2$, $\mathbf{L} \in \mathbb{R}^{2 \times 2}$ and $\odot$ denotes the Hadamard product between matrices, i.e. the element-wise product. The vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)^T$ can be interpreted as controlling the overall variance of the components of the 2×1 random matrix-valued GP, $\mathbf{V}$, specified in Eq. (6.40), while $\mathbf{L}$ can be viewed as determining the correlation structure between the components of $\mathbf{V}$. We place i.i.d. log-normal hierarchical priors on the components of $\boldsymbol{\sigma}$ and we place a LKJ distribution (see Definition 5.2) on $\mathbf{L}$, i.e.

$$\begin{aligned} \sigma_1, \sigma_2 &\stackrel{\text{i.i.d.}}{\sim} \text{Log } \mathcal{N}(0, 1), \\ \mathbf{L} &\sim \text{LKJ}(2, 1). \end{aligned} \quad (6.63)$$

Note that, as mentioned in Definition 5.2, modelling $\mathbf{L} \sim \text{LKJ}(2, 1)$ is equivalent to placing a uniform prior for the correlation matrix $\mathbf{L}$.

- **Prior anchor point:** As described in Section 6.6.2 the anchor point will be taken to be the Karcher mean of the observed data. Due to the low dimensions involved we can adopt a “brute-force” approach to computing the Karcher mean. This approach will be described below and pseudo-code for it is presented in Algorithm 7.
- **Likelihood noise level:** Since this example is intended for visualisation purposes and the observed dataset will be simulated exactly we fix the likelihood noise level to be $\sigma = 0.01$, a non-zero, albeit small value, chosen for numerical stability.

Brute-force computation of the Karcher mean

Since we can identify $\text{Gr}(2, 1)$ with the interval $[0, \pi)$, representing the angle a given subspace, $\mathcal{W} \in \text{Gr}(2, 1)$, makes with the positive x -axis we can express the Karcher mean (see Eq. (6.31)) as a minimisation problem over this angle $\theta \in [0, \pi)$. To be more precise, if we denote by $\mathbf{x}(\theta) := (\cos(\theta), \sin(\theta))^T$ we have:

$$\overline{\mathcal{W}}_N = \underset{\theta \in [0, \pi)}{\text{argmin}}(S(\theta)), \quad (6.64)$$

where $S(\theta)$ is defined as follows,

$$S(\theta) := \sum_{j=1}^N d_{\text{Gr}}(P(\mathbf{s}_j), [\mathbf{x}(\theta)])^2. \quad (6.65)$$

We can thus compute the Karcher mean via a “brute-force” approach as follows: we discretise the interval $[0, \pi)$ using $m \in \mathbb{N}$ equally spaced points, $\{\theta_i\}_{i=1}^m \subset [0, \pi)$. For each $\theta = \theta_i$ we compute the sum, $S(\theta)$, in Eq. (6.64) and store the resulting value. The Karcher mean is then given by $\bar{\mathcal{W}}_N = [\mathbf{x}(\theta_k)]$ where $k \in \{1, \dots, m\}$ is the index which achieves the lowest result for the sum in Eq. (6.64). Note that we compute the Riemannian Grassmann distance using Algorithm 4. The algorithm is presented in Algorithm 7 in pseudo-code.

Algorithm 7: Brute-force computation of the Karcher mean for points on $\text{Gr}(2, 1)$

Input: Points $\{\mathcal{W}_i\}_{i=1}^N \subset \text{Gr}(2, 1)$ and number of grid points, $m \in \mathbb{N}$, to discretise $[0, \pi)$
Output: Karcher mean $\bar{\mathcal{W}}_N$ of the points $\{\mathcal{W}_i\}_{i=1}^N$

```

1  $S = \{\}$  ;
2 for  $i \leftarrow 1$  to  $m$  do
3    $S_i := 0$  ;
4    $\theta_i := \frac{i-1}{m} \pi$  ; // Grid point in  $[0, \pi)$ 
5   for  $j \leftarrow 1$  to  $N$  do
6      $d_j := d_{\text{Gr}}(P(\mathbf{s}_j), [\mathbf{x}(\theta_i)])$  ; // Compute Grassmann distance via Algorithm 4
7      $S_i \leftarrow S_i + d_j^2$ 
8   end
9    $S \leftarrow S \cup \{S_i\}$ 
10 end
11  $k := \underset{i=1, \dots, m}{\text{argmin}}(S)$  ;
12 return  $[\mathbf{x}(\theta_k)]$ 
```

Using Algorithm 7 for large enough m allows us to obtain the following plot of θ vs $S(\theta)$ shown in Fig. 6.3.

Figure 6.3 allows us to identify the anchor point for our observed data with the point in $\text{Gr}(2, 1)$ with $\theta = 0$, i.e. we have:

$$\bar{\mathcal{W}}_N = [\mathbf{e}_1], \quad (6.66)$$

where $\mathbf{e}_1 = (1, 0)^T$ is the first standard basis vector of $\mathbb{R}^2$.

It is helpful to view this on the unit-circle representation of $\text{Gr}(2, 1)$ which we include in Fig. 6.4. This plot shows clearly that the Karcher mean represents a midpoint subject to the constraint that the observed data points lie on the unit circle.

Remark 6.12. The brute force approach taken to compute the Karcher mean here cannot be applied to $\text{Gr}(d, n)$ for general $d, n \in \mathbb{N}$. Instead one can use a gradient descent algorithm to attempt to solve the minimisation problem Eq. (6.31). One version of this algorithm is presented in Algorithm 1 in Section 5.4. Noting that the loss surface $S(\theta)$ shown in Fig. 6.3 contains several

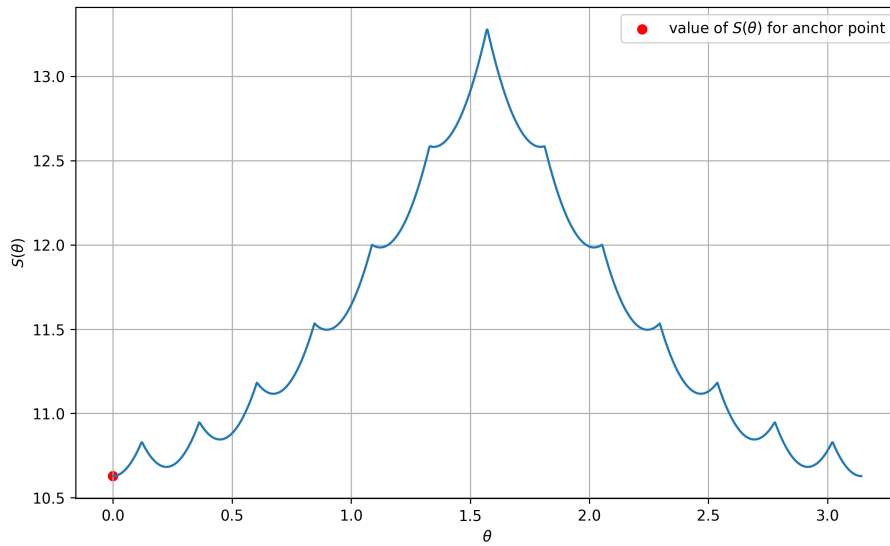


Figure 6.3: Plot of θ vs $S(\theta)$ showing the location of the minimum, and hence the location of the anchor point, in red.

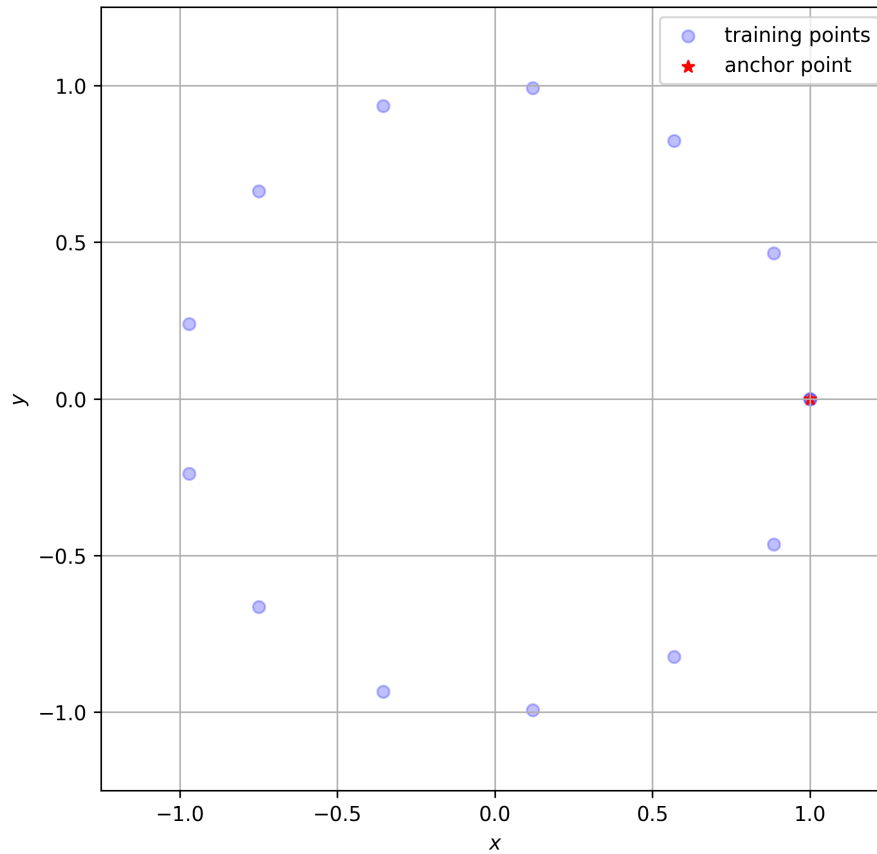


Figure 6.4: Plot of observed points with the Karcher mean on the circle representation of $\text{Gr}(2, 1)$

local minima indicates that any gradient based procedure for solving Eq. (6.31) can easily get stuck; it is thus advisable to run the algorithm for several different random initialisation points and choose the one which obtains the lowest value. ♦

Fitting the model

In order to fit the model we utilise HMC, as described in Section 5.3, to obtain samples from the hyper-parameter posterior distribution corresponding to the training dataset Eq. (6.58). We initialise the HMC sampler with the Maximum a Posteriori (MAP) estimate which is obtained using NumPyro's built-in SVI support. The HMC sampler is then run for $I = 7000$ iterations with an additional burn-in period of $I_b = 2000$ iterations. We thin the resulting chain by 2 to obtain $\tilde{I} = 3500$ samples from the hyper-parameter posterior distribution. Figure 6.5 shows traceplots and autocorrelation plots for the resulting samples for ℓ, σ_1, σ_2 , and the off-diagonal components of $\mathbf{L}$, i.e. L_{12}, L_{21} . Note that the diagonal components have been omitted since they are constrained to be 1 as $\mathbf{L}$ is a correlation matrix. From the traceplots in Fig. 6.5 we observe good mixing of the chains while the autocorrelation plots show a drop to acceptable levels of autocorrelation after a lag of at most 20. The split Gelman-Rubin diagnostic [63] was also computed to be 1.0 for these five hyper-parameters, giving us further evidence for convergence of the HMC.

In-sample performance

To assess the in-sample performance of our model we took the resulting values for the in-sample subspace angles, $\{\alpha(s_i)\}_{i=1}^{14}$, obtained from each sampled hyper-parameter values from the HMC and plotted a density plot of them together with the training and test values of $\alpha(s)$. Figure 6.6 shows the resulting plot. We can see from this plot that we obtain acceptable in-sample performance for our model with the HMC samples clustering closely around the training data.

Remark 6.13. Notice that at the endpoints $s = 0, 1$ in Fig. 6.6 the density plot reveals the presence of some samples which seem to completely miss the true data. However, these two clusters at each endpoint have a separation of around π , which is consistent with the fact that the true signal should have $\alpha(0) = \alpha(1) = 0$. These points are equivalent in terms of subspace angle and so the edges of Fig. 6.6 should be thought of as being “glued” together. ♦

To get a more quantitative assessment of the in-sample performance we then compute, for each training point, the Karcher mean, M_i , of the resulting points in $\text{Gr}(2, 1)$ corresponding to the sampled hyper-parameters from the HMC:

$$M_i := \underset{[\mathbf{X}] \in \text{Gr}(2,1)}{\text{argmin}} \sum_{j=1}^{\tilde{I}} d_{\text{Gr}}(\hat{P}^{(j)}(s_i), [\mathbf{X}]), \quad i = 1, \dots, 14, \quad (6.67)$$

where here $\{\hat{P}^{(j)}(s_i)\}_{j=1}^{\tilde{I}}$ are the points on $\text{Gr}(2, 1)$ generated from the model during the training phase corresponding to $P(s_i)$. We then compute the Riemannian Grassmann distance between M_i and $P(s_i)$:

$$E_i := d_{\text{Gr}}(M_i, P(s_i)), \quad (6.68)$$

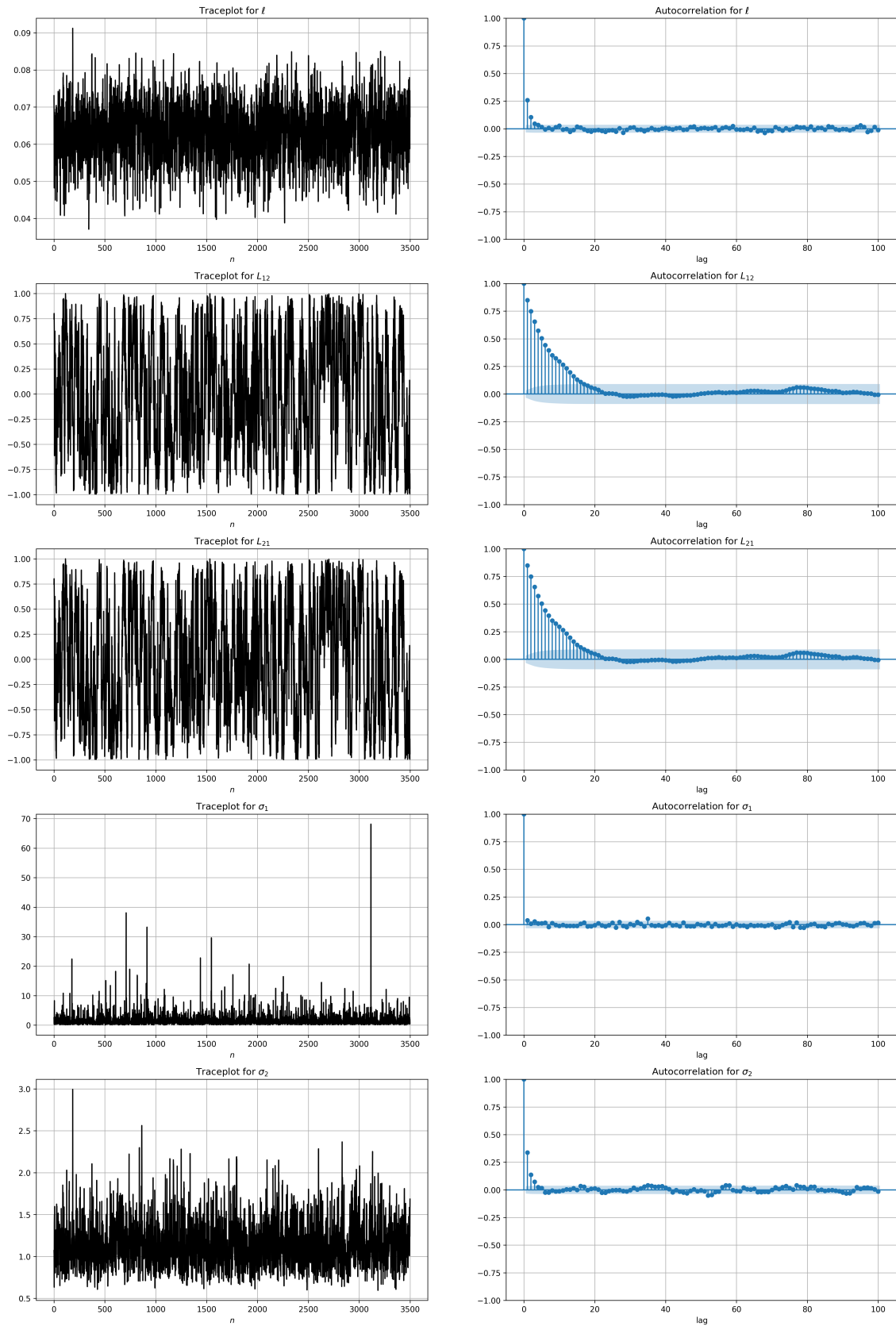


Figure 6.5: Diagnostic plots for HMC sampler (Section 6.7.1). *Left column*: Traceplots for the sampled hyper-parameters. *Right column*: Autocorrelation plots for the sampled hyper-parameters.

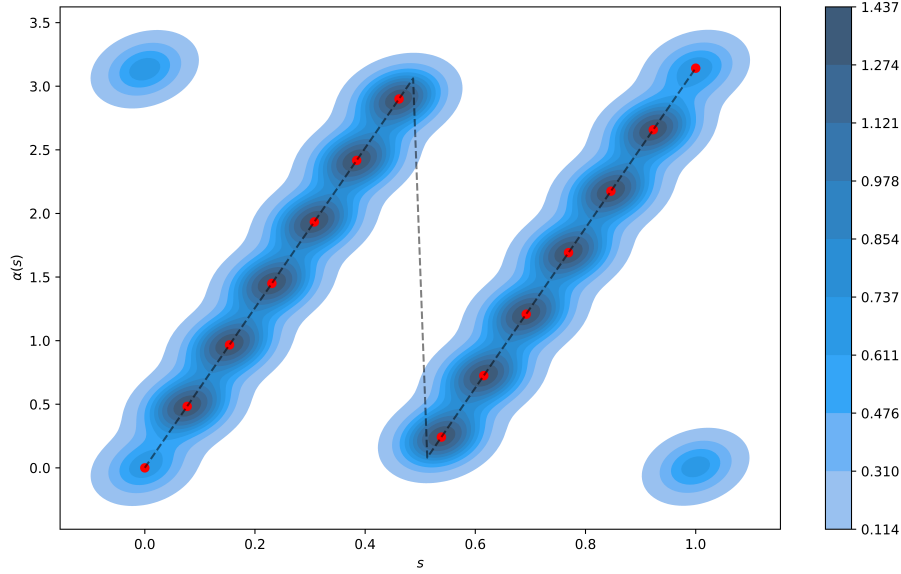


Figure 6.6: Plot of s vs $\alpha(s)$ showing the in-sample results from the HMC sampler (Section 6.7.1). The training points are plotted in red, while the test data is plotted as the dashed black line.

as a measure of the in-sample accuracy of our model. Table 6.1 presents summary statistics of the resulting in-sample errors $\{E_i\}_{i=1}^{14}$. From this table we can see that we achieve low errors overall indicating a high in-sample accuracy.

Table 6.1: Statistics for in-sample errors (Section 6.7.1).

	Errors
mean	0.000726
std	0.000502
min	0.000000
25%	0.000302
50%	0.000726
75%	0.001149
max	0.001451

Predictive performance

We now use our model to make predictions for the test dataset Eq. (6.59) as described in Section 6.6.3. To assess the resulting predictive performance of our model we include a plot of the predictions for the tangent vectors $\{\mathbf{U}(s_i^*)\}_{i=1}^{40}$ in Fig. 6.7. From this plot we can see that our model accurately tracks the tangent vectors for the test dataset Eq. (6.59). Notice that the first component of the tangent vector is constant and the model predicts this with very small uncertainty overall. For the varying second component we can see from the figure that the uncertainty drops near the train data points, $\{s_i\}_{i=1}^{14}$, as is to be expected. We also include in Fig. 6.8 a density plot of the corresponding predictions for the subspace angles, $\{\alpha(s_i^*)\}_{i=1}^{40}$, obtained from each sampled hyper-parameter value from the HMC. This plot is analogous to Fig. 6.6 but now for the

test dataset Eq. (6.59) and using the GP predictions from the model. We can see from this plot that we obtain acceptable predictive performance for our model with the predictions clustering closely around the test data.

For a more quantitative assessment of the predictive performance we compute, for each test point, the Karcher mean, $\tilde{M}_i$, of the predictions made by the model for $P(s_i^*)$ for each sampled hyper-parameter from the HMC. Since for each sampled hyper-parameter we use a GP-based model to make predictions we have two versions of errors:

- Errors between the Karcher mean using the mean GP predictions (i.e. the mean of the GP posterior for each sampled hyper-parameter).
- Errors between the Karcher mean using the actual predictions from the GP.

To indicate the level of uncertainty in the predictions we also compute a proxy for the standard deviation of our predictions via the following formula:

$$\sqrt{\frac{1}{\tilde{I}} \sum_{j=1}^{\tilde{I}} d_{\text{Gr}}(\tilde{P}^{(j)}(s_i^*), \tilde{M}_j)}, \quad i = 1, \dots, 40, \quad (6.69)$$

where $\tilde{P}^{(j)}(s_i^*)$ denotes the predictions (either the mean GP prediction or the actual prediction as appropriate). We illustrate these errors together with the associated uncertainties in Fig. 6.9. From this plot we can see that the error remains small overall with spikes at around $s = 0.3, 0.7$. These spikes of inaccuracy can be seen to correspond to the sharp jumps in the second component of the true tangent vectors at the same locations in Fig. 6.7. This is perhaps due to the fact that we are using the squared-exponential family of kernels in our model which cannot deal with such sharp jumps. Also note that the uncertainty depicted by the shaded regions drops to practically zero at the training points which are depicted with the vertical dashed green lines. We also include in Table 6.2 a table of summary statistics for these two types of errors.

Table 6.2: Statistics for predictive errors (Section 6.7.1).

	Errors (using mean)	Errors (using predictions)
mean	0.192035	0.193057
std	0.288104	0.288085
min	0.000000	0.000000
25%	0.001149	0.001149
50%	0.074869	0.076441
75%	0.264218	0.266577
max	0.969547	0.972691

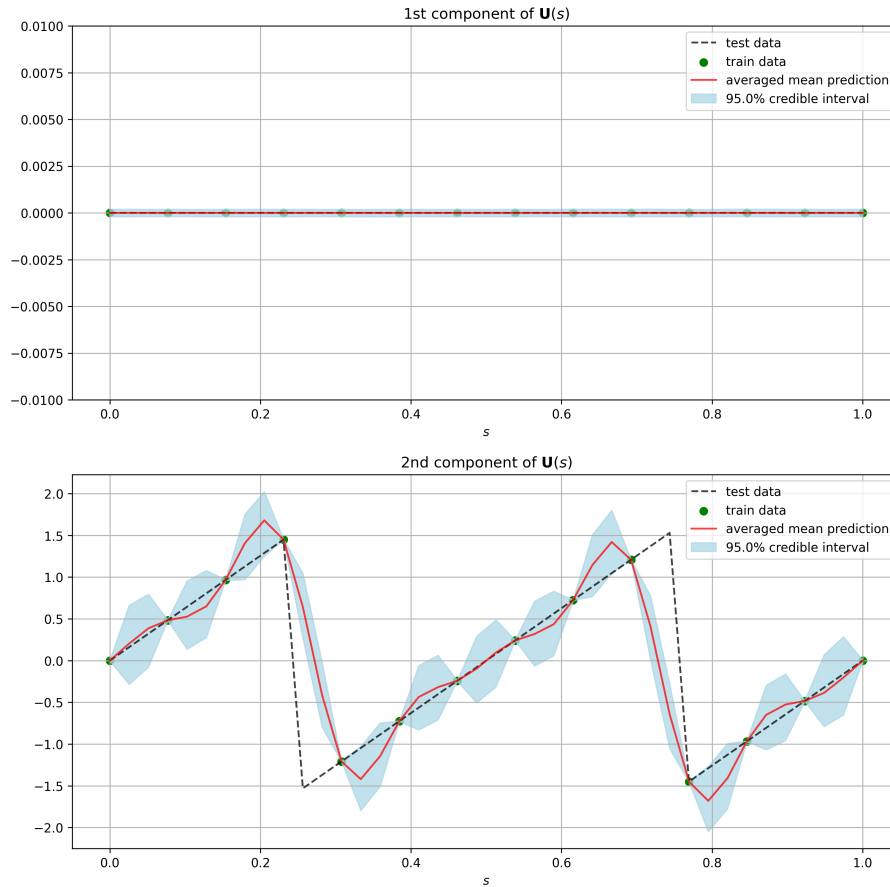


Figure 6.7: Predictions for tangent vectors $\mathbf{U}(s)$ for the dataset Eq. (6.59). The solid red curves show the averaged mean prediction, i.e. the average of the mean GP predictions for each sampled posterior hyper-parameter. The shaded blue region gives a 95% credible region over the GP predictions for each sampled posterior hyper-parameter.

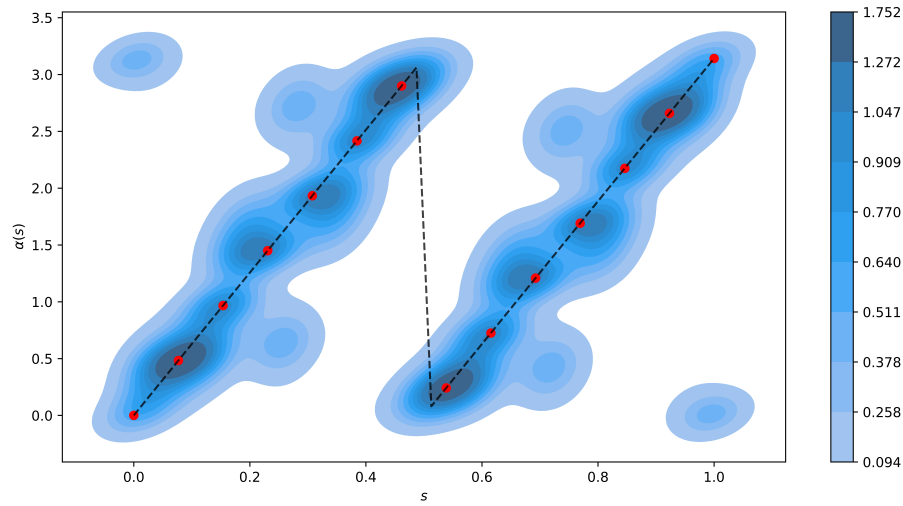


Figure 6.8: Plot of s vs $\alpha(s)$ showing the resulting predictions for the test dataset Eq. (6.59). The training points are plotted in red, while the test data is plotted as the dashed black line.

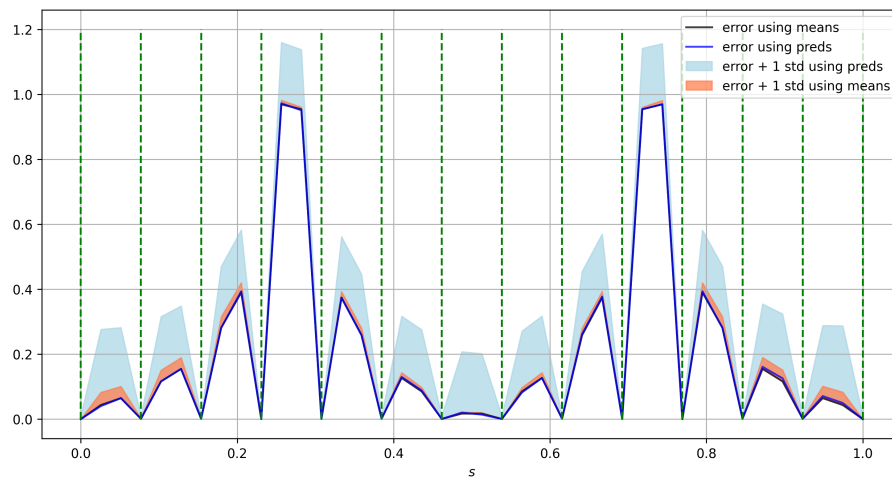


Figure 6.9: Plot of errors against s using both the GP predictive mean and predictions (Section 6.7.1). The shaded regions depict a measure of uncertainty given by the error plus 1 standard deviation computed via Eq. (6.69) for both the predictive mean and predictions respectively. The vertical dashed green lines indicate the locations of the training data.

6.7.2 Localised active subspaces for a parametrised ODE

Dataset Description

In this section we will present a numerical experiment which will be used to demonstrate the methodology for analysing localised active subspaces presented in Section 6.4. This example is motivated by Example 5.1 in [34]. We consider the following parametrised second-order ODE:

$$-\frac{d}{dx} \left(\kappa(x; \mathbf{p}) \frac{d}{dx} u(x; \mathbf{p}) \right) = f(x), \quad x \in [0, 1], \quad (6.70)$$

$$u(0; \mathbf{p}) = u(1; \mathbf{p}) = 0. \quad (6.71)$$

The logarithm of the coefficient $\kappa = \kappa(x; \mathbf{p})$ of the differential operator in Eq. (6.70) is given by the following truncated Karhunen-Lo  ve (KL) (see Theorem A.4) type expansion:

$$\log \kappa(x; \mathbf{p}) := \sum_{k=1}^{d_p} p_k \gamma_k \varphi_k(x), \quad (6.72)$$

where $\mathbf{p} = (p_1, \dots, p_{d_p})^T \in \mathbb{R}^{d_p}$ and the pairs $\{(\gamma_k, \varphi_k(x))\}_{k=1}^{\infty}$ are the eigenpairs associated with the correlation operator corresponding to the Mat  rn-1/2 covariance function:

$$c(x, y) := \exp(\beta^{-1}|x - y|). \quad (6.73)$$

We take the length scale in Eq. (6.73) to be $\beta = 1$. We choose to truncate the KL expansion of the field at $d_p = 100$ terms, which implies that our parameter space is $\mathbb{R}^{100}$ with associated p.d.f. ρ being a standard multivariate normal density function, i.e. the parameter vector $\mathbf{p} \sim \mathcal{N}_{100}(\mathbf{0}, \mathbf{I}_{100})$.

The function of interest for which we want to apply the localised active subspace methodology described in Section 6.4 is $u(x; \mathbf{p})$, i.e. we are interested in investigating if we can perform localised dimension reduction for the $\mathbf{p}$ -dependence of $u(x; \mathbf{p})$ as a function of the location x in the ODE's domain. In particular, we are interested in the leading eigenvector of the x -dependent matrix $C(x) \in \mathbb{R}^{100 \times 100}$:

$$C(x) := \int_{\mathbb{R}^{100}} (\nabla_{\mathbf{p}} u(x; \mathbf{p})) (\nabla_{\mathbf{p}} u(x; \mathbf{p}))^T \rho(x) d\mathbf{p}, \quad (6.74)$$

which corresponds to the matrix given by Eq. (6.25). Thus, we are interested in learning the mapping $P : [0, 1] \rightarrow \text{Gr}(100, 1)$ which maps $x \in [0, 1]$ to the localised active subspace $P(x) \in \text{Gr}(100, 1)$ given by the leading eigenspace of Eq. (6.74). We will utilise our GrassGP model to learn this function given $N = 10$ observations of $P(x)$ at equally spaced locations, $\{x_i\}_{i=1}^{10}$, between 0 and 1, i.e. our training dataset is:

$$\mathcal{D}_{10} = \{(x_i, P(x_i))\}_{i=1}^{10}. \quad (6.75)$$

We augment the observed dataset $\mathcal{D}_{10}$ with an additional 30 points to yield a test dataset,

$$\mathcal{D}_{40}^* = \{(x_i^*, P(x_i^*))\}_{i=1}^{40}, \quad (6.76)$$

which will be compared with predictions made from the model. Note that we first generate the test dataset Eq. (6.76) and then take every other 4th point to obtain our training dataset Eq. (6.75).

Remark 6.14. In order to compute $P(x)$ to obtain the training and test datasets we will approximate this matrix and its leading eigenspace via the Monte Carlo approach given in Algorithm 6 with 300 Monte Carlo samples from the p.d.f. ρ . We compute the gradient $\nabla_{\mathbf{p}} u(x; \mathbf{p})$ by solving an adjoint problem associated to the BVP from Eqs. (6.70) and (6.71). $\blacklozenge$

Figure 6.10 shows a plot illustrating how the localised active subspace drifts from $P(x_0^*)$ as we move along the ODE's domain $[0, 1]$. This plot shows the Riemannian Grassmann distance (see Definition 5.6) between the first test (and train) point $P(x_0^*)$ and the rest of the train and test points respectively. We can see that there is a substantial, yet smooth, change in the localised active subspace as we move from one end of $[0, 1]$ to the other.

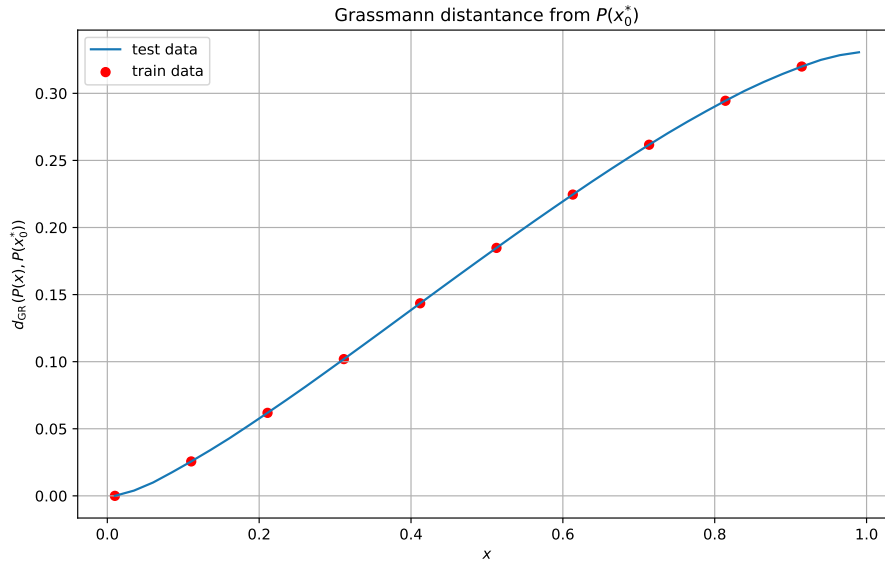


Figure 6.10: Plot illustrating the evolving localised active subspace by plotting the Riemannian Grassmann distance between $P(x_0^*)$ and the train and test datasets Eqs. (6.75) and (6.76) respectively against x .

Remark 6.15. In Example 5.1 of [34] the authors apply the global active subspace methodology to study a linear function of the solution of a parametrised elliptic PDE. In our example we have, for simplicity, restricted the dimension of the PDE to 1 dimension, i.e. we consider an ODE instead, noting that our model can be easily extended to PDEs, as we will see in Section 6.7.3. Further, instead of studying a linear function of the solution we instead investigate the parameter dependence at different points in the ODE's domain. We also consider Dirichlet boundary conditions

instead of the mixed Dirichlet and Neumann conditions considered in the example from [34]. ♦

Experiment Parameters

We now detail the various choices we make for the unknown parameters described in Section 6.6.2 for this experiment. Note that, as in Section 6.7.1, anything not specified explicitly is assumed to be the same as described in Section 6.6.2.

- **Prior kernel function:** As in Section 6.7.1 we again choose to work with the family of Squared Exponential (SE) kernels Eq. (6.60) parametrised by the length scale, $\ell > 0$. We again place a log-normal hierarchical prior on the length scale:

$$\ell \sim \text{Log } \mathcal{N}(0, 0.5^2). \quad (6.77)$$

- **Prior parameter covariance matrix:** We depart from the choice in Section 6.7.1 and use the following parametric form for $\mathbf{\Omega} \in \mathbb{R}^{100 \times 100}$:

$$\mathbf{\Omega} := \text{diag}(\{\omega_i^2\}_{i=1}^{100}), \quad (6.78)$$

i.e. we model $\mathbf{\Omega}$ as a diagonal matrix where the diagonal entries control the overall variance of the components of the 100×1 matrix-valued GP. We place i.i.d. log-normal hierarchical priors on the $\{\omega_i\}_{i=1}^{100}$, i.e.

$$\omega_1, \dots, \omega_{100} \stackrel{\text{i.i.d.}}{\sim} \text{Log } \mathcal{N}(0, 1). \quad (6.79)$$

The choice of a diagonal matrix for $\mathbf{\Omega}$ reduces the computational burden in the HMC sampler which helps speed up the inference, which is vital due to the high dimensions involved in this example.

- **Prior anchor point:** We will again take the anchor point to be the Karcher mean of the observed data. Due to the higher dimensions involved we can no longer adopt the “brute-force” approach taken in Section 6.7.1. We instead use the gradient descent based approach given by Algorithm 1 to estimate the Karcher mean. This algorithm was run for several random initialisations which had little effect on the resulting anchor point.
- **Likelihood noise level:** We again decide to fix the likelihood noise level to a small, albeit non-zero value, $\sigma = 0.001$. We use a smaller value than in Section 6.7.1 since the scale in the resulting tangent vectors of our dataset are now smaller due to the larger dimensions involved.

Fitting the model

In order to fit the model we again utilise HMC to obtain samples from the hyper-parameter posterior distribution corresponding to the training dataset Eq. (6.75). We again initialise the HMC sampler with the MAP estimate which is obtained using NumPyro’s built-in SVI support. The HMC sampler is then run for $I = 7000$ iterations with an additional burn-in period of $I_b = 5000$ iterations. We note that we had to use a longer burn-in period than in Section 6.7.1 in order to achieve convergence of the sampler. We again thin the resulting chain by 2 to obtain $\tilde{I} = 3500$ samples from the hyper-parameter posterior distribution. Figure 6.11 shows traceplots and autocorrelation plots for the resulting sample for $\ell, \omega_1, \omega_2, \omega_3$, and ω_4 . We exclude the plots for $\{\omega_i\}_{i=5}^{100}$ as they exhibited similar behaviour as the first four parameters. From the traceplots in Fig. 6.11 we observe good mixing of the chains while the autocorrelation plots show a drop to acceptable levels of autocorrelation after a lag of at most 20. The split Gelman-Rubin diagnostic was also computed to be 1.0 for all the hyper-parameters, except for the kernel length, ℓ , which had a value of 1.01. These values give us further evidence for convergence of the HMC.

In-sample performance

To assess the in-sample performance of our model we again compute the Riemannian Grassmann distance between the training points and the Karcher mean of the resulting points in $\text{Gr}(100, 1)$ corresponding to the sampled hyper-parameters from the HMC as was computed for the example in Section 6.7.1. Table 6.3 presents summary statistics for the resulting in-sample errors (this is analogous to Table 6.1 for the previous example). From this table we can see that we achieve low errors overall indicating a high in-sample accuracy.

Table 6.3: Statistics for in-sample errors (Section 6.7.2).

	Errors
mean	0.001124
std	0.001175
min	0.000401
25%	0.000493
50%	0.000711
75%	0.001068
max	0.004300

Predictive performance

We now use our model to make predictions for the localised active subspaces contained in the test dataset Eq. (6.76) as described in Section 6.6.3. As in Section 6.7.1 we again use the Riemannian Grassmann distance as a measure of predictive performance. As was done previously we compute

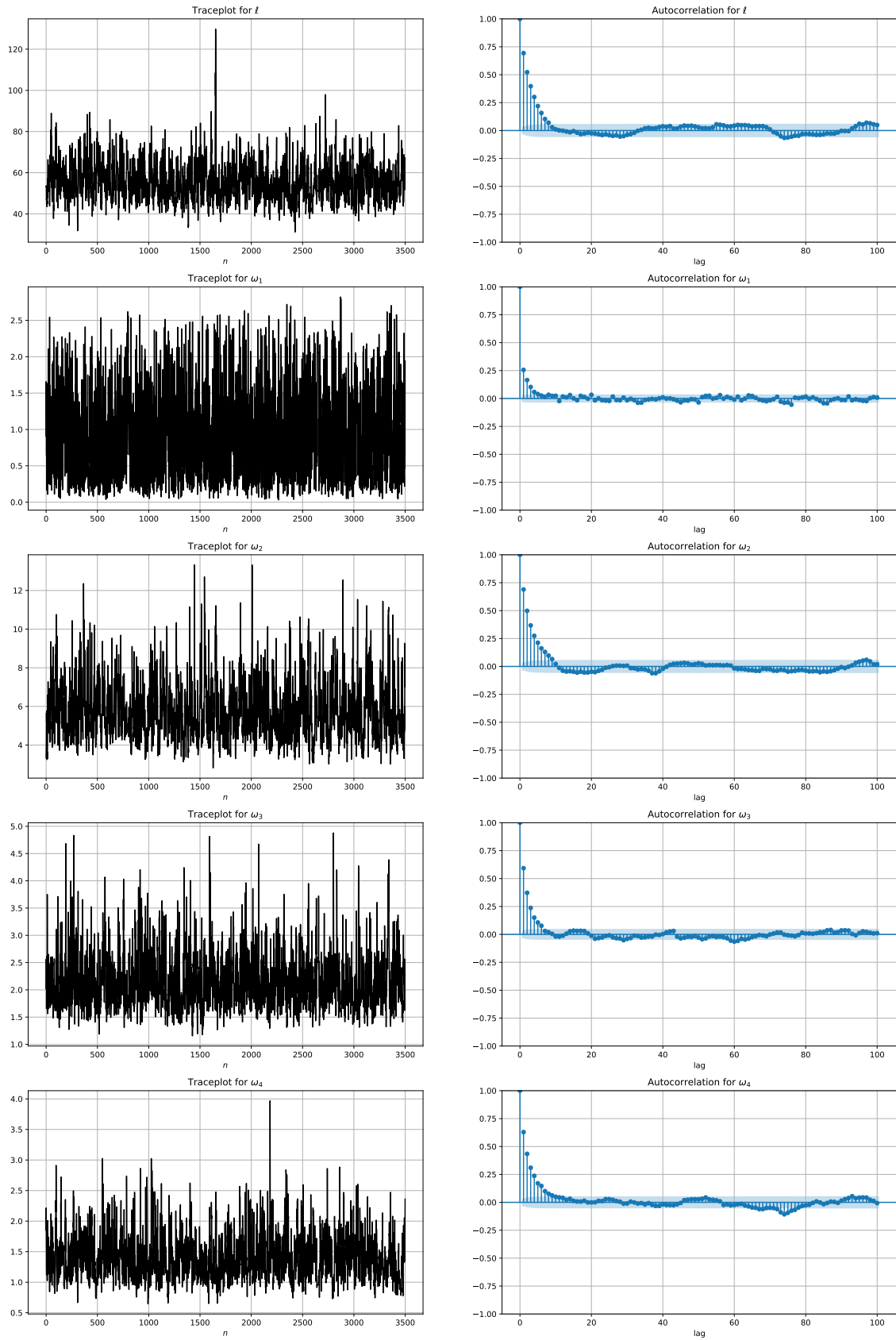


Figure 6.11: Diagnostic plots for HMC sampler (Section 6.7.2). *Left column:* Traceplots for the sampled hyper-parameters. *Right column:* Autocorrelation plots for the sampled hyper-parameters.

this distance for both the GP posterior predictive mean and the posterior predictions. Table 6.4 contains summary statistics for these two types of errors. From this table we again see that we achieve low errors overall indicating a high predictive accuracy.

Table 6.4: Statistics for predictive errors (Section 6.7.2).

	Errors (using mean)	Errors (using predictions)
mean	0.026329	0.026338
std	0.009333	0.009321
min	0.014288	0.014143
25%	0.018784	0.018884
50%	0.025607	0.025556
75%	0.030288	0.030296
max	0.054694	0.054689

6.7.3 Localised active subspaces for a parametrised PDE

Dataset Description

In this section we present a numerical experiment similar to the one in Section 6.7.2 but for the following parametrised elliptic PDE:

$$-\nabla \cdot (\kappa(\mathbf{x}; \mathbf{p}) \nabla u(\mathbf{x}; \mathbf{p})) = f(\mathbf{x}), \quad \mathbf{x} \in D := [0, 1]^2, \quad (6.80)$$

$$u(\mathbf{x}; \mathbf{p}) = 0, \quad \mathbf{x} \in \partial D. \quad (6.81)$$

Note that this is simply a 2-dimensional version of the parametrised second-order ODE Eq. (6.70). Here the logarithm of the coefficient $\kappa = \kappa(\mathbf{x}; \mathbf{p})$ of the differential operator in Eq. (6.80) is again given by the following truncated Karhunen-Lo  ve (KL) type expansion:

$$\log \kappa(\mathbf{x}; \mathbf{p}) := \sum_{k=1}^{d_p} p_k \gamma_k \varphi_k(\mathbf{x}), \quad (6.82)$$

where $\mathbf{p} = (p_1, \dots, p_{d_p})^T \in \mathbb{R}^{d_p}$ and the pairs $\{(\gamma_k, \varphi_k(\mathbf{x}))\}_{k=1}^{\infty}$ are the eigenpairs associated with the correlation operator corresponding to the 2-dimensional Mat  rn-1/2 covariance function:

$$c(\mathbf{x}, \mathbf{y}) := \exp(\beta^{-1} \|\mathbf{x} - \mathbf{y}\|). \quad (6.83)$$

We take the length scale in Eq. (6.83) to be $\beta = 1$. We choose to truncate the KL expansion of the field at $d_p = 100$ terms, which implies that our parameter space is again $\mathbb{R}^{100}$ with associated p.d.f. ρ being a standard multivariate normal density function, i.e. the parameter vector $\mathbf{p} \sim \mathcal{N}_{100}(\mathbf{0}, \mathbf{I}_{100})$.

As in the experiment in Section 6.7.2 the function of interest for which we want to apply the

localised active subspace methodology is $u(\mathbf{x}; \mathbf{p})$. Thus, the experimental setup is almost identical to that in Section 6.7.2 but now we are interested in learning a mapping from the unit-square to $\text{Gr}(100, 1)$, i.e. we wish to learn a mapping $P : [0, 1]^2 \rightarrow \text{Gr}(100, 1)$. Owing to the strong similarities between this experiment and the one in Section 6.7.2 we will simply highlight any differences in the modelling choices and then present the results.

Our GrassGP model will be utilised to learn this function given $N = 25$ observations of P , i.e. our training dataset is:

$$\mathcal{D}_{25} = \{(\mathbf{x}_i, P(\mathbf{x}_i))\}_{i=1}^{25}. \quad (6.84)$$

We augment the observed dataset $\mathcal{D}_{25}$ with an additional 75 points to yield a test dataset,

$$\mathcal{D}_{100}^* = \{(\mathbf{x}_i^*, P(\mathbf{x}_i^*))\}_{i=1}^{100}. \quad (6.85)$$

We note that we use the same Monte Carlo approach given in Algorithm 6 to compute these datasets as outlined in Remark 6.14. These train and test locations are illustrated in Fig. 6.12. Figure 6.13 illustrates how the localised active subspace drifts from the value at the bottom-left-

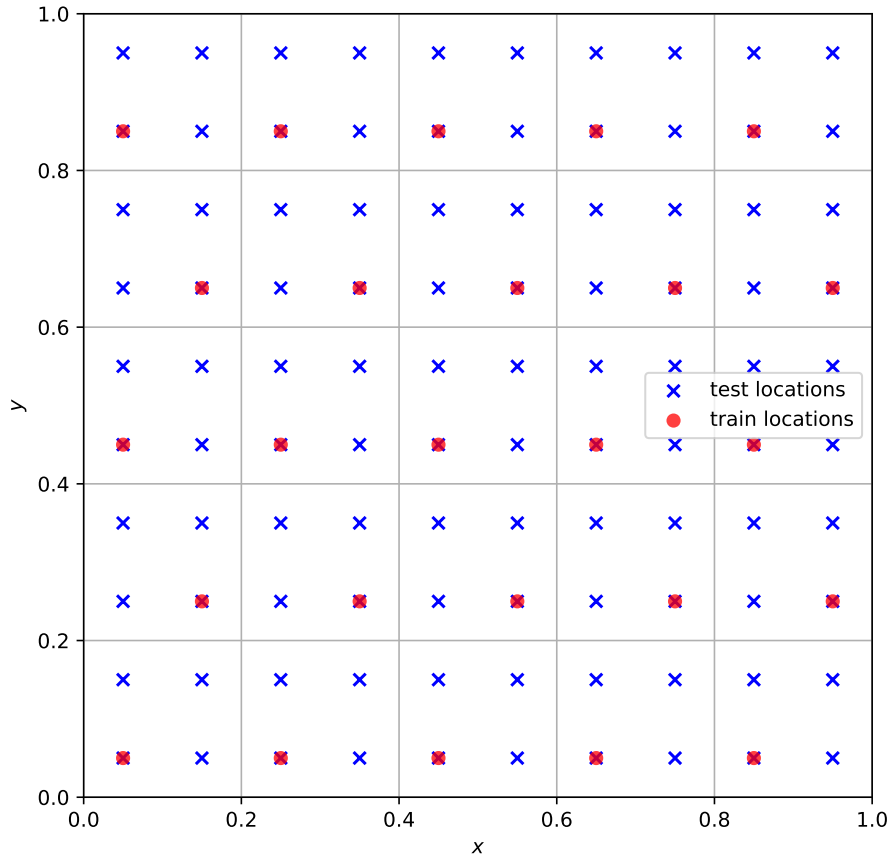


Figure 6.12: Plot illustrating the train and test locations for the numerical experiment in Section 6.7.3.

most test point $\mathbf{x}_0^*$ as we move along the PDE's domain $[0, 1]^2$. This plot shows the Riemannian

Grassmann distance between $P(\mathbf{x}_0^*)$ and the rest of the test data points as a contour plot. This plot corresponds to Fig. 6.10 for our 2-dimensional PDE. We again see that there is a substantial, yet smooth, change in the localised active subspace as we move along the unit square.

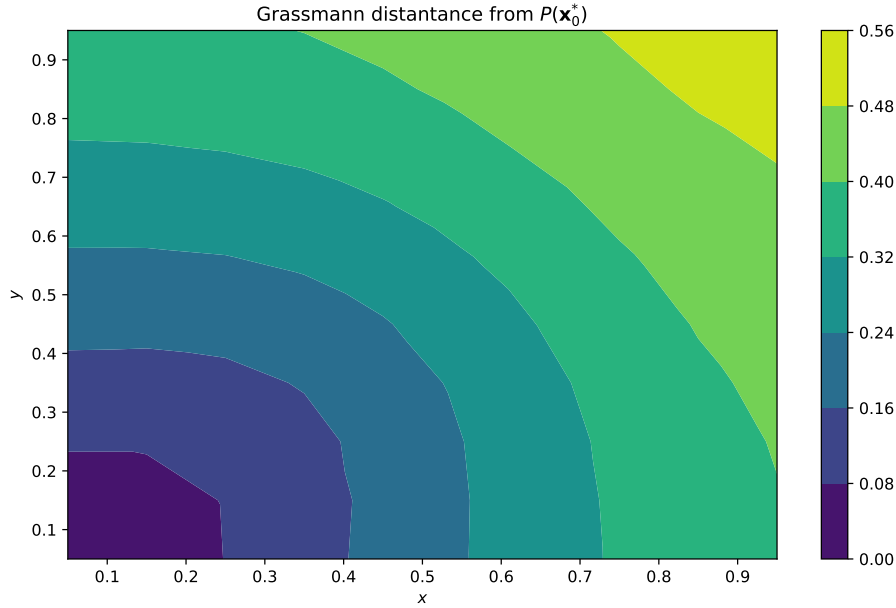


Figure 6.13: Plot illustrating the evolving localised active subspace by plotting the Riemannian Grassmann distance between $P(\mathbf{x}_0^*)$ and the test dataset.

Experiment Parameters

Due to the similarity with the experiment in Section 6.7.2 we make the same choices for the unknown parameters required for our GrassGP model. The only difference is that we use a larger likelihood noise level of $\sigma = 0.01$ for this example to help with numerical stability and convergence of the HMC chain.

Fitting the model

In order to fit the model we again utilise HMC to obtain samples from the hyper-parameter posterior distribution corresponding to the training dataset Eq. (6.84). We again initialise the HMC sampler with the MAP estimate which is obtained using NumPyro’s built-in SVI support. The HMC sampler is then run for $I = 7000$ iterations with an additional burn-in period of $I_b = 5000$ iterations. We again thin the resulting chain by 2 to obtain $\tilde{I} = 3500$ samples from the hyper-parameter posterior distribution. Figure 6.14 shows traceplots and autocorrelation plots for the resulting sample for $\ell, \omega_1, \omega_2, \omega_3$, and ω_4 . We exclude the plots for $\{\omega_i\}_{i=5}^{100}$ as they exhibited similar behaviour as the first four parameters. From the traceplots in Fig. 6.14 we observe good mixing of the chains while the autocorrelation plots show a drop to acceptable levels of autocorrelation

after a lag of around 10. The split Gelman-Rubin diagnostic was also computed to be 1.0 for all the hyper-parameters. These values give us further evidence for convergence of the HMC.

In-sample performance

To assess the in-sample performance of our model we again compute the Riemannian Grassmann distance between the training points and the Karcher mean of the resulting points in $\text{Gr}(100, 1)$ corresponding to the sampled hyper-parameters from the HMC as was computed for the examples in Sections 6.7.1 and 6.7.2. Table 6.5 presents summary statistics for the resulting in-sample errors (this is analogous to Tables 6.1 and 6.3 for the previous examples). From this table we can see that we achieve low errors overall indicating a high in-sample accuracy.

Table 6.5: Statistics for in-sample errors (Section 6.7.3).

	Errors
mean	0.020501
std	0.006538
min	0.012895
25%	0.015060
50%	0.017789
75%	0.024757
max	0.038437

Predictive performance

We now use our model to make predictions for the localised active subspaces contained in the test dataset Eq. (6.85) as described in Section 6.6.3. As in Sections 6.7.1 and 6.7.2 we again use the Riemannian Grassmann distance as a measure of predictive performance. As was done previously we compute this distance for both the GP posterior predictive mean and the posterior predictions. Table 6.6 contains summary statistics for these two types of errors. From this table we again see that we achieve low errors overall indicating a high predictive accuracy.

Table 6.6: Statistics for predictive errors (Section 6.7.3).

	Errors (using mean)	Errors (using predictions)
mean	0.055704	0.055698
std	0.016808	0.016811
min	0.037917	0.038006
25%	0.044399	0.044290
50%	0.049517	0.049505
75%	0.063978	0.064007
max	0.118090	0.118115

6.8 Discussion

In this chapter we have provided a review of the concepts of optimal ridge approximation and active subspaces and then motivated the idea of localised active subspaces where the dimension reducing subspace smoothly varies as a function of some of the input parameters. We then presented a probabilistic model based on GPs which can be used to interpolate a field of Grassmann-valued response variables and hence can be used to efficiently approximate the localised active subspace at novel locations. This model was then demonstrated using a simple numerical example involving low dimensions, while two higher-dimensional examples involving a parameterised ODE and PDE demonstrated the model's utility in estimating localised active subspaces.

Several possible avenues for future work present themselves. The first would be to consider other models for modelling Grassmann-valued response variables than the model considered here and apply these models to the estimation of localised active subspaces. Another avenue would be to extend this model to also be able to perform localised optimal ridge approximation, i.e. to estimate a function $f : \mathbb{R}^m \times \mathbb{R}^{d_p} \rightarrow \mathbb{R}$ by a ridge function with a parameter dependent ridge subspace, $f(\mathbf{s}, \mathbf{x}) \approx g(\mathbf{M}(\mathbf{s})^T \mathbf{x})$.

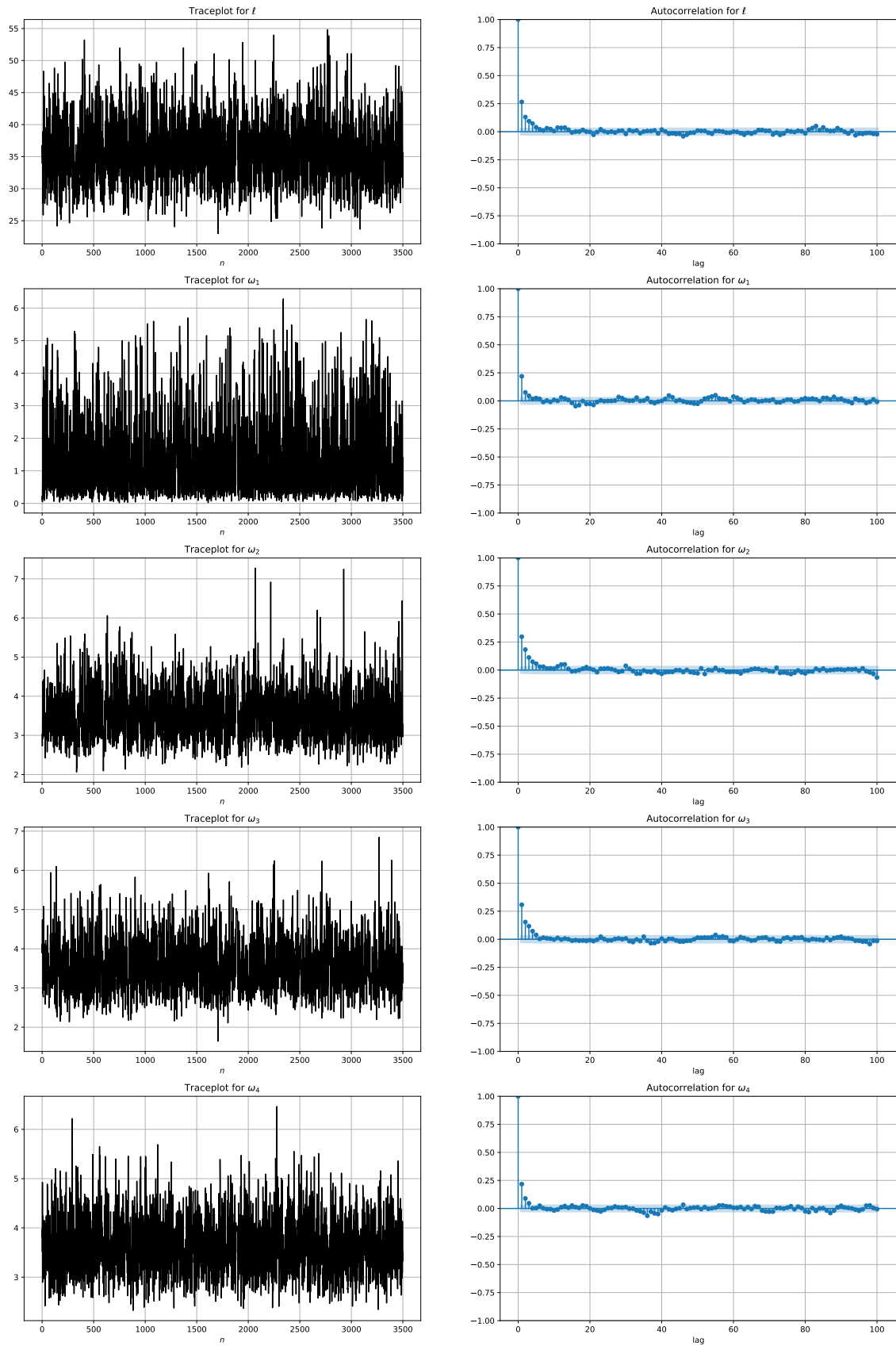


Figure 6.14: Diagnostic plots for HMC sampler (Section 6.7.3). *Left column*: Traceplots for the sampled hyper-parameters. *Right column*: Autocorrelation plots for the sampled hyper-parameters.

Chapter 7

Conclusion and Future Work

In this final chapter we summarize the contributions made in this thesis and also discuss potential avenues for future work.

7.1 Contributions

The novel contributions of this thesis can be separated into two areas which we now summarise.

In Chapters 3 and 4 we investigated the recently proposed StatFEM model and analysed its associated discretisation error yielding new perspectives on inference in the context of Bayesian UQ for PDE models. In particular, in the case of spatial elliptic BVPs we were able to prove in Section 3.4 that the StatFEM prior converges to the true prior as a function of the mesh width, with respect to the Wasserstein-2 distance, at a rate which matches the underlying deterministic FEM method. We were then able to extend this to show that given noisy observations of linear quantities of interest, the resulting posterior approximation is also controlled at the same rate. In both scenarios, the choice of the Wasserstein distance was vital in allowing us to compare the mutually singular true and StatFEM priors/posteriors respectively. As outlined both in Sections 2.3.6 and 3.4 this mutual singularity prohibits the use of the more commonly used probability discrepancies such as the KL divergence or Hellinger distance. We then presented our work towards extending the error analysis for the StatFEM method from elliptic to parabolic PDEs. This work, while still at an early stage, presents several avenues for future work which will be summarized below. We finally presented several numerical experiments on a number of illustrative FEM test problems demonstrating both the theoretical results for the spatial case, as well as possible extensions and scenarios which explore the limits of our theory where some of our assumptions no longer hold.

In Chapter 6 we reviewed the concepts of optimal ridge approximation and active subspaces which motivated the idea of localised active subspaces. We then introduced a fully Bayesian prob-

abilistic model based on GPs, called GrassGP, which can be used to estimate localised active subspaces by interpolating a field of Grassmann-valued random variables. The idea of lifting the response variables to a particular tangent space was crucial to the derivation of this model and is applicable to problems beyond the GrassGP framework where one might be interested in interpolating response variables on some other Riemannian manifold. Several numerical experiments were then presented, the first of which allows us to easily visualise the GrassGP model, while the others demonstrated its potential in estimating localised active subspaces.

7.2 Potential Future Research Directions

The work presented in this thesis suggests several avenues for potential future work which we now outline.

In Section 3.4 we proved in Theorem 3.4 that in the posterior case linear quantities of interest are also controlled at a rate of $\mathcal{O}(h^2)$, i.e. $W(\ell_{\#}\tilde{\nu}_{\star}, \ell_{\#}\tilde{\nu}_h) \leq \mathcal{O}(h^2)$, where we recall that $\tilde{\nu}_{\star}, \tilde{\nu}_h$ are the true and StatFEM posteriors respectively, and ℓ is a bounded linear functional on $L^2(D)$. It will be interesting to see if this theorem can be extended to the full posterior, something which the numerical experiments in Chapter 4 seem to suggest is possible. Another interesting avenue would be to see if similar error bounds can be proven for approximations other than FEM. This seems likely since much of the analysis in Section 3.4 relied on the existence of bounds on the underlying deterministic numerical method. In a similar vein, the generalisation of these results to other probability divergences such as Wasserstein- p distances, for $p > 2$, would have interesting implications such as allowing even more control over other features of the distribution such as quantiles and extrema. This would be especially important for applications involving risk-assessment.

Our efforts to extend the error analysis of the StatFEM method from elliptic to parabolic PDEs also yields several avenues for future work as outlined in Section 3.5.3. Chief among these potential directions would be extending the analysis to the error between the covariance terms which might allow an analogue of Theorem 3.3 to be proven for the time-dependent case. Further, in light of the increasing trend of incorporating noisy observational data into our statistical pipeline, it will be interesting to investigate if it is possible to derive error bounds for a Kalman-filter like scheme for the StatFEM framework as outlined in [50].

In Chapter 6 we motivated the concept of localised active subspaces and then suggested one methodology for estimating such subspaces via a Bayesian probabilistic model based on GPs. One important avenue of future work would be to extend this model to be able to perform localised optimal ridge approximation as well, i.e. to be able to approximate a function $f : \mathbb{R}^m \times \mathbb{R}^{d_p} \rightarrow \mathbb{R}$ by a ridge function with a parameter dependent ridge subspace, $f(\mathbf{s}, \mathbf{x}) \approx g(\mathbf{M}(\mathbf{s})^T \mathbf{x})$. Further, as

noted in our literature review for this project in Section 6.2, other models exist in the literature for modelling Grassmann-valued response variables. It will be interesting to consider the use of such models for the estimation of localised active subspaces and to compare their performance for this task. We also note that while the two projects presented in this thesis were separate in their focus it will be interesting to investigate the use of localised active subspaces in the context of StatFEM for PDEs with large numbers of parameters of interest.

7.3 Closing Remarks

Studying the effects of numerical discretisation error on inference as well as new methodologies for performing dimension reduction remain areas of great importance in the field of Bayesian Uncertainty Quantification, especially in relation to PDE-based models. This thesis has attempted to provide new perspectives on these two areas. Despite the focus on PDE/ODE-based models in this thesis it should be noted that the analysis used, and the methodologies presented in this thesis can extend well beyond such models and can help in the study of complex models from other domains.

Bibliography

- [1] P.-A. Absil, R. Mahony, and R. Sepulchre. Riemannian Geometry of Grassmann Manifolds with a View on Algorithmic Computation. *Acta Applicandae Mathematicae*, 80(2):199–220, Jan. 2004. ISSN 0167-8019. doi: 10.1023/B:ACAP.0000013855.14971.91.
- [2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, N.J. ; Woodstock, 2008. ISBN 978-0-691-13298-3.
- [3] Ö. D. Akyildiz, C. Duffin, S. Sabanis, and M. Girolami. Statistical finite elements via Langevin dynamics. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1560–1585, 2022.
- [4] M. Aldosary, J. Wang, and C. Li. Structural reliability and stochastic finite element methods: State-of-the-art review and evidence-based comparison. *EC*, 35(6):2165–2214, Oct. 2018. ISSN 0264-4401. doi: 10.1108/EC-04-2018-0157.
- [5] M. S. Alnaes, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS project version 1.5. *Archive of Numerical Software*, 3, 2015. doi: 10.11588/ans.2015.100.20553.
- [6] M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for Vector-Valued Functions: A Review. *arXiv:1106.6251 [cs, math, stat]*, Apr. 2012.
- [7] T. Bayes. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.
- [8] T. Bendokat, R. Zimmermann, and P.-A. Absil. A Grassmann Manifold Handbook: Basic Geometry and Computational Aspects. *arXiv:2011.13699 [cs, math]*, Dec. 2020.
- [9] M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

- [10] M. Betancourt, S. Byrne, S. Livingstone, and M. Girolami. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, Nov. 2017. ISSN 1350-7265. doi: 10.3150/16-BEJ810.
- [11] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 3rd ed edition, 1995. ISBN 978-0-471-00710-4.
- [12] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep Universal Probabilistic Programming, Oct. 2018.
- [13] A. Bobrowski. *Functional Analysis for Probability and Stochastic Processes: An Introduction*. Cambridge University Press, Cambridge, 2005. ISBN 978-0-521-83166-6. doi: 10.1017/CBO9780511614583.
- [14] V. I. Bogachev. *Gaussian Measures*. American Mathematical Soc., 1998. ISBN 978-0-8218-1054-5.
- [15] D. Bolin, K. Kirchner, and M. Kovács. Numerical solution of fractional elliptic stochastic PDEs with spatial white noise. *IMA Journal of Numerical Analysis*, 40(2):1051–1073, Apr. 2020. ISSN 0272-4979, 1464-3642. doi: 10.1093/imanum/dry091.
- [16] D. Bosq. *Linear Processes in Function Spaces*, volume 149 of *Lecture Notes in Statistics*. Springer, New York, NY, 2000. ISBN 978-0-387-95052-5 978-1-4612-1154-9. doi: 10.1007/978-1-4612-1154-9.
- [17] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 1 edition, Mar. 2023. ISBN 978-1-00-916616-4 978-1-00-916617-1 978-1-00-916615-7. doi: 10.1017/9781009166164.
- [18] P. Boyle and M. Frean. Dependent Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.
- [19] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: Composable transformations of Python+NumPy programs, 2018.
- [20] H. Brezis and F. Browder. Partial Differential Equations in the 20th Century. *Advances in Mathematics*, 135(1):76–144, Apr. 1998. ISSN 0001-8708. doi: 10.1006/aima.1997.1713.
- [21] S. Byrne and M. Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

- [22] G. Casella and E. I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, Aug. 1992. ISSN 0003-1305. doi: 10.1080/00031305.1992.10475878.
- [23] Z. Chen, J. Fan, and K. Wang. Multivariate Gaussian processes: Definitions, examples and applications. *METRON*, Jan. 2023. ISSN 2281-695X. doi: 10.1007/s40300-023-00238-3.
- [24] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, Nov. 1995. ISSN 0003-1305. doi: 10.1080/00031305.1995.10476177.
- [25] O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. A. Girolami. Bayesian Solution Uncertainty Quantification for Differential Equations. *Bayesian Analysis*, 11(4):1239–1267, Dec. 2016. ISSN 1936-0975, 1931-6690. doi: 10.1214/16-BA1017.
- [26] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Jan. 2002. ISBN 978-0-89871-514-9. doi: 10.1137/1.9780898719208.
- [27] J. Cockayne and A. B. Duncan. Probabilistic gradients for fast calibration of differential equation models. *arXiv*, 2020. ISSN 23318422.
- [28] J. Cockayne, C. Oates, T. Sullivan, and M. Girolami. Probabilistic numerical methods for partial differential equations and Bayesian inverse problems. *arXiv preprint arXiv:1605.07811*, 2016.
- [29] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian Probabilistic Numerical Methods. *SIAM Rev.*, 61(3):756–789, Jan. 2019. ISSN 0036-1445, 1095-7200. doi: 10.1137/17M1139357.
- [30] D. L. Cohn. *Measure theory*, volume 5. Springer, 2013.
- [31] P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart, and K. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Statistics and Computing*, 27:1065–1082, 2017.
- [32] P. Constantine and D. Gleich. Computing active subspaces with Monte Carlo, July 2015.
- [33] P. G. Constantine. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. Number 2 in SIAM Spotlights. Society for Industrial and Applied Mathematics, Philadelphia, 2015. ISBN 978-1-61197-385-3.
- [34] P. G. Constantine, E. Dow, and Q. Wang. Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces. *SIAM J. Sci. Comput.*, 36(4):A1500–A1524, Jan. 2014. ISSN 1064-8275. doi: 10.1137/130916138.

- [35] P. G. Constantine, A. Eftekhari, and M. B. Wakin. Computing active subspaces efficiently with gradient sketching. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 353–356, Dec. 2015. doi: 10.1109/CAMSAP.2015.7383809.
- [36] P. G. Constantine, A. Eftekhari, J. Hokanson, and R. A. Ward. A near-stationary subspace for ridge approximation. *Computer Methods in Applied Mechanics and Engineering*, 326: 402–421, Nov. 2017. ISSN 00457825. doi: 10.1016/j.cma.2017.07.038.
- [37] R. D. Cook and L. Ni. Sufficient Dimension Reduction via Inverse Regression. *Journal of the American Statistical Association*, 100(470):410–428, June 2005. ISSN 0162-1459. doi: 10.1198/016214504000001501.
- [38] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25(11):115008, Oct. 2009. ISSN 0266-5611. doi: 10.1088/0266-5611/25/11/115008.
- [39] S. G. Cox and K. Kirchner. Regularity and convergence analysis in Sobolev and Hölder spaces for generalized Whittle–Matérn fields. *Numerische Mathematik*, 146(4):819–873, Dec. 2020. ISSN 0029-599X, 0945-3245. doi: 10.1007/s00211-020-01151-x.
- [40] J. A. Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Diaz. On lower bounds for the L2-Wasserstein metric in a Hilbert space. *J Theor Probab*, 9(2):263–283, Apr. 1996. ISSN 1572-9230. doi: 10.1007/BF02214649.
- [41] G. Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer Science & Business Media, 2006.
- [42] M. Dashti and A. M. Stuart. The Bayesian Approach to Inverse Problems. In *Handbook of Uncertainty Quantification*, pages 311–428. Springer International Publishing, 2017. ISBN 978-3-319-12385-1. doi: 10.1007/978-3-319-12385-1_7.
- [43] F. Demengel and G. Demengel. *Functional Spaces for the Theory of Elliptic Partial Differential Equations*. Universitext. Springer London, London, 2012. ISBN 978-1-4471-2806-9 978-1-4471-2807-6. doi: 10.1007/978-1-4471-2807-6.
- [44] R. Dennis Cook. SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29(9-10):2109–2121, Jan. 2000. ISSN 0361-0926. doi: 10.1080/03610920008832598.

- [45] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker’s guide to the fractional Sobolev spaces. *Bulletin des Sciences Mathématiques*, 136(5):521–573, July 2012. ISSN 0007-4497. doi: 10.1016/j.bulsci.2011.12.004.
- [46] M. Diaz, A. Jaramillo, and J. C. Pardo. Fluctuations for matrix-valued Gaussian processes. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 58(4):2216 – 2249, 2022. doi: 10.1214/21-AIHP1238. URL <https://doi.org/10.1214/21-AIHP1238>.
- [47] D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, Sept. 1982. ISSN 0047-259X. doi: 10.1016/0047-259X(82)90077-X.
- [48] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, Sept. 1987. ISSN 0370-2693. doi: 10.1016/0370-2693(87)91197-X.
- [49] R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2 edition, 2002. ISBN 978-0-521-80972-6. doi: 10.1017/CBO9780511755347.
- [50] C. Duffin, E. Cripps, T. Stemler, and M. Girolami. Statistical finite elements for misspecified models. *Proceedings of the National Academy of Sciences*, 118(2):e2015006118, 2021.
- [51] C. Duffin, P. Branson, M. Rayson, M. Girolami, E. Cripps, and T. Stemler. Exploring Model Misspecification in Statistical Finite Elements via Shallow Water Equations, July 2023.
- [52] M. L. Eaton. *Multivariate Statistics: A Vector Space Approach*. Number 53 in Lecture Notes-Monograph Series / Institute of Mathematical Statistics. Inst. of Mathematical Statistics, Beachwood, Ohio, 2007. ISBN 978-0-940600-69-0.
- [53] A. Edelman, T. A. Arias, and S. T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM J. Matrix Anal. & Appl.*, 20(2):303–353, Jan. 1998. ISSN 0895-4798, 1095-7162. doi: 10.1137/S0895479895290954.
- [54] A. Eftekhari, M. B. Wakin, P. Li, P. G. Constantine, and R. A. Ward. Learning the second-moment matrix of a smooth function from point samples. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 671–675, Oct. 2017. doi: 10.1109/ACSSC.2017.8335427.
- [55] L. C. Evans. *Partial Differential Equations*. American Mathematical Soc., 2010. ISBN 978-0-8218-4974-3.
- [56] E. Febrianto, L. Butler, M. Girolami, and F. Cirak. Digital twinning of self-sensing structures using the statistical finite element method. *Data-Centric Engineering*, 3:e31, 2022.

- [57] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python Optimal Transport. page 8.
- [58] T. Fletcher. Statistics on manifolds. In *Riemannian Geometric Statistics in Medical Image Analysis*, pages 39–74. Elsevier, 2020. ISBN 978-0-12-814725-2. doi: 10.1016/B978-0-12-814725-2.00009-1.
- [59] M. Fornasier, K. Schnass, and J. Vybiral. Learning Functions of Few Arbitrary Linear Parameters in High Dimensions. *Found Comput Math*, 12(2):229–262, Apr. 2012. ISSN 1615-3383. doi: 10.1007/s10208-012-9115-y.
- [60] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l’institut Henri Poincaré*, 10(4):215–310, 1948. ISSN 2400-4855.
- [61] A. Garbuno-Inigo, T. Helin, F. Hoffmann, and B. Hosseini. Bayesian posterior perturbation analysis with integral probability metrics. *arXiv preprint arXiv:2303.01512*, 2023.
- [62] R. Gautier, P. Pandita, S. Ghosh, and D. Mavris. A fully Bayesian gradient-free supervised dimension reduction method using Gaussian processes. *arXiv preprint arXiv:2008.03534*, 2020.
- [63] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
- [64] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2010.00765.x.
- [65] M. Girolami, E. Febrianto, G. Yin, and F. Cirak. The Statistical Finite Element Method (statFEM) for coherent synthesis of observation data and model predictions. *Computer Methods in Applied Mechanics and Engineering*, 375:113533, Mar. 2021. ISSN 00457825. doi: 10.1016/j.cma.2020.113533.
- [66] I. Gohberg, S. Goldberg, and M. Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser, Dec. 2012. ISBN 978-3-0348-7980-4.
- [67] G. Golub and V. Pereyra. Separable Nonlinear Least Squares: The Variable Projection Method and its applications. *Inverse Problems*, 19(2):R1–R26, Apr. 2003. ISSN 0266-5611, 1361-6420. doi: 10.1088/0266-5611/19/2/201.

- [68] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Number 69 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, SIAM ed edition, 2011. ISBN 978-1-61197-202-3.
- [69] T. H. Gronwall. Note on the Derivatives with Respect to a Parameter of the Solutions of a System of Differential Equations. *Annals of Mathematics*, 20(4):292–296, 1919. ISSN 0003-486X. doi: 10.2307/1967124.
- [70] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.
- [71] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97.
- [72] P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, July 2015. doi: 10.1098/rspa.2015.0142.
- [73] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. ISSN 1533-7928.
- [74] J. M. Hokanson and P. G. Constantine. Data-Driven Polynomial Ridge Approximation Using Variable Projection. *SIAM J. Sci. Comput.*, 40(3):A1566–A1589, Jan. 2018. ISSN 1064-8275. doi: 10.1137/17M1117690.
- [75] B. Hosseini. Well-Posed Bayesian Inverse Problems with Infinitely Divisible and Heavy-Tailed Prior Measures. *SIAM/ASA J. Uncertainty Quantification*, 5(1):1024–1060, Jan. 2017. doi: 10.1137/16M1096372.
- [76] B. Hosseini and N. Nigam. Well-Posed Bayesian Inverse Problems: Priors with Exponential Tails. *SIAM/ASA J. Uncertainty Quantification*, 5(1):436–465, Jan. 2017. doi: 10.1137/16M1076824.
- [77] J. Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-21297-0 978-3-642-21298-7. doi: 10.1007/978-3-642-21298-7.
- [78] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

- [79] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977. ISSN 1097-0312. doi: 10.1002/cpa.3160300502.
- [80] K. Karhunen. *Über Lineare Methoden in Der Wahrscheinlichkeitsrechnung: Akademische Abhandlung*. Sana, 1947.
- [81] T. Karvonen, F. Cirak, and M. Girolami. Error analysis for a statistical finite element method. *arXiv preprint arXiv:2201.07543*, 2022.
- [82] S. Keiper. Analysis of generalized ridge functions in high dimensions. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 259–263, May 2015. doi: 10.1109/SAMPTA.2015.7148892.
- [83] W. S. Kendall. Probability, Convexity, and Harmonic Maps with Small Image I: Uniqueness and Fine Existence. *Proceedings of the London Mathematical Society*, s3-61(2):371–406, 1990. ISSN 1460-244X. doi: 10.1112/plms/s3-61.2.371.
- [84] H. J. Kim, N. Adluru, M. D. Collins, M. K. Chung, B. B. Bendlin, S. C. Johnson, R. J. Davidson, and V. Singh. Multivariate General Linear Models (MGLM) on Riemannian Manifolds with Applications to Statistical Analysis of Diffusion Weighted Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2705–2712, 2014.
- [85] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [86] G. Koop, R. León-González, and R. W. Strachan. Efficient posterior simulation for cointegrated models with priors on the cointegration space. *Econometric Reviews*, 29(2):224–242, 2009.
- [87] E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, New York, 1978. ISBN 978-0-471-50731-4.
- [88] Z. Lai, L.-H. Lim, and K. Ye. Simpler Grassmannian optimization. *arXiv preprint arXiv:2009.13502*, 2020.
- [89] S. Larsson and V. Thomée. *Partial Differential Equations with Numerical Methods*, volume 45. Springer Science & Business Media, 2008.
- [90] J. Latz. On the Well-posedness of Bayesian Inverse Problems. *SIAM/ASA J. Uncertainty Quantification*, 8(1):451–482, Jan. 2020. doi: 10.1137/19M1247176.

- [91] P. D. Lax and A. N. Milgram. IX. Parabolic Equations. In *IX. Parabolic Equations*, pages 167–190. Princeton University Press, Mar. 2016. ISBN 978-1-4008-8218-2. doi: 10.1515/9781400882182-010.
- [92] J. M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-91754-2 978-3-319-91755-9. doi: 10.1007/978-3-319-91755-9.
- [93] G. Leoni. *A First Course in Sobolev Spaces*. American Mathematical Soc., Nov. 2017. ISBN 978-1-4704-2921-8.
- [94] D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on Vines and Extended Onion Method. *Journal of Multivariate Analysis*, 100(9):1989–2001, Oct. 2009. ISSN 0047259X. doi: 10.1016/j.jmva.2009.04.008.
- [95] B. Li and S. Wang. On Directional Regression for Dimension Reduction. *Journal of the American Statistical Association*, 102(479):997–1008, Sept. 2007. ISSN 0162-1459. doi: 10.1198/016214507000000536.
- [96] K.-C. Li. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. ISSN 0162-1459. doi: 10.2307/2290563.
- [97] M. Lifshits and M. Lifshits. *Lectures on Gaussian Processes*. Springer, 2012.
- [98] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.00777.x.
- [99] M. Loeve. Probability theory II 1978 4 Berlin Springer 0385.60001 10.1007.
- [100] A. Logg, K.-A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012. doi: 10.1007/978-3-642-23099-8.
- [101] G. J. Lord, C. E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, Aug. 2014. ISBN 978-0-521-89990-1.
- [102] Y. Ma and L. Zhu. A Review on Dimension Reduction. *International Statistical Review / Revue Internationale de Statistique*, 81(1):134–150, 2013. ISSN 0306-7734.
- [103] A. Mallasto and A. Feragen. Wrapped Gaussian Process Regression on Riemannian Manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, 2018.

- [104] S. Martino and A. Riebler. Integrated Nested Laplace Approximations (INLA). *arXiv preprint arXiv:1907.01248*, 2019.
- [105] V. Masarotto, V. M. Panaretos, and Y. Zemel. Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes. *Sankhya A*, 81(1):172–213, Feb. 2019. ISSN 0976-8378. doi: 10.1007/s13171-018-0130-1.
- [106] H. G. Matthies, C. E. Brenner, C. G. Bucher, and C. Guedes Soares. Uncertainties in probabilistic numerical analysis of structures and solids-Stochastic finite elements. *Structural Safety*, 19(3):283–336, Jan. 1997. ISSN 01674730. doi: 10.1016/S0167-4730(97)00013-1.
- [107] U. Menzeffricke. Hierarchical modeling with Gaussian Processes. *Communications in Statistics - Simulation and Computation*, 29(4):1089–1108, Jan. 2000. ISSN 0361-0918. doi: 10.1080/03610910008813655.
- [108] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, Dec. 2004. ISSN 0021-9606. doi: 10.1063/1.1699114.
- [109] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Ltd, 1 edition, 1982. doi: 10.1002/9780470316559.
- [110] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.
- [111] S. A. Niederer, M. S. Sacks, M. Girolami, and K. Willcox. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1(5):313–320, 2021.
- [112] B. Øksendal. *Stochastic Differential Equations*. Universitext. Springer, Berlin, Heidelberg, 2003. ISBN 978-3-540-04758-2 978-3-642-14394-6. doi: 10.1007/978-3-642-14394-6.
- [113] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and Its Applications*, 1982. ISSN 00243795. doi: 10.1016/0024-3795(82)90112-4.
- [114] H. Owhadi. Bayesian Numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015.
- [115] H. Owhadi, C. Scovel, and T. Sullivan. On the brittleness of Bayesian Inference. *SIAM Review*, 57(4):566–582, 2015.
- [116] G. Parra and F. Tobar. Spectral Mixture Kernels for Multi-Output Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

- [117] G. Peyré, M. Cuturi, et al. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86), 2017.
- [118] D. Phan, N. Pradhan, and M. Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro, Dec. 2019.
- [119] D. Pigoli, J. A. D. Aston, I. L. Dryden, and P. Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014. ISSN 0006-3444.
- [120] D. Pigoli, A. Menafoglio, and P. Secchi. Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis*, 145:117–131, Mar. 2016. ISSN 0047259X. doi: 10.1016/j.jmva.2015.12.006.
- [121] A. Pinkus. *Ridge Functions*, volume 205. Cambridge University Press, 2015.
- [122] H. Pishro-Nik. Introduction to probability, statistics, and random processes. <https://www.probabilitycourse.com/>, 2016.
- [123] G. D. Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, Apr. 2014. ISBN 978-1-107-05584-1.
- [124] B. S. Rajput and S. Cambanis. Gaussian Processes and Gaussian Measures. *The Annals of Mathematical Statistics*, 43(6):1944–1952, 1972. ISSN 0003-4851.
- [125] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X.
- [126] Q. Rentmeesters and P.-A. Absil. Algorithm comparison for Karcher mean computation of rotation matrices and diffusion tensors. In *2011 19th European Signal Processing Conference*, pages 2229–2233, Aug. 2011.
- [127] C. Robert and G. Casella. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statist. Sci.*, 26(1), Feb. 2011. ISSN 0883-4237. doi: 10.1214/10-STS351.
- [128] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York, NY, 2004. ISBN 978-1-4419-1939-7 978-1-4757-4145-2. doi: 10.1007/978-1-4757-4145-2.
- [129] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(none):20–71, Jan. 2004. ISSN 1549-5787, 1549-5787. doi: 10.1214/1549578041000000024.

- [130] W. Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 3d ed edition, 1976. ISBN 978-0-07-054235-8.
- [131] O. Runborg. Verifying numerical convergence rates. <https://www.csc.kth.se/utbildning/kth/kurser/DN2255/ndiff13/ConvRate.pdf>.
- [132] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409–423, Nov. 1989. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1177012413.
- [133] A. M. Samarov. Exploring Regression Structure Using Nonparametric Functional Estimation. *Journal of the American Statistical Association*, 88(423):836–847, Sept. 1993. ISSN 0162-1459. doi: 10.1080/01621459.1993.10476348.
- [134] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20827-5 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2.
- [135] D. Sanz-Alonso and R. Yang. Finite element representations of Gaussian processes: Balancing numerical and statistical accuracy. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1323–1349, 2022.
- [136] M. Scheuerer. Regularity of the sample paths of a general second order random field. *Stochastic Processes and their Applications*, 120(10):1879–1897, Sept. 2010. ISSN 03044149. doi: 10.1016/j.spa.2010.05.009.
- [137] A. Scillitoe, P. Seshadri, C. Y. Wong, and A. Duncan. Polynomial ridge flowfield estimation. *Physics of Fluids*, 33(12):127110, Dec. 2021. ISSN 1070-6631, 1089-7666. doi: 10.1063/5.0064000.
- [138] P. Seshadri, S. Yuchi, and G. T. Parks. Dimension Reduction via Gaussian Ridge Functions. *SIAM/ASA J. Uncertainty Quantification*, 7(4):1301–1322, Jan. 2019. ISSN 2166-2525. doi: 10.1137/18M1168571.
- [139] X. Shi, H. Zhu, J. G. Ibrahim, F. Liang, J. Lieberman, and M. Styner. Intrinsic Regression Models for Medial Representation of Subcortical Structures. *Journal of the American Statistical Association*, 107(497):12–23, Mar. 2012. ISSN 0162-1459. doi: 10.1080/01621459.2011.643710.

- [140] F. Sigrist, H. R. Künsch, and W. A. Stahel. Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):3–33, 2015.
- [141] F. Simpson, A. Boukouvalas, V. Cadek, E. Sarkans, and N. Durrande. The Minecraft Kernel: Modelling correlated Gaussian Processes in the Fourier domain. In *International Conference on Artificial Intelligence and Statistics*, pages 1945–1953. PMLR, 2021.
- [142] B. Sprungk. On the local Lipschitz stability of Bayesian inverse problems. *Inverse Problems*, 36(5):1–28, 2020. ISSN 13616420. doi: 10.1088/1361-6420/ab6f43.
- [143] G. Stefanou. The stochastic finite element method: Past, present and future. *Computer Methods in Applied Mechanics and Engineering*, 198(9-12):1031–1051, 2009. ISSN 00457825. doi: 10.1016/j.cma.2008.11.007.
- [144] O. Stegle, C. Lippert, J. M. Mooij, N. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate Gaussian models with iid observation noise. *Advances in neural information processing systems*, 24, 2011.
- [145] G. Strang, G. J. Fix, and D. S. Griffin. An Analysis of the Finite-Element Method. *Journal of Applied Mechanics*, 41(1):62, Mar. 1974. ISSN 0021-8936. doi: 10.1115/1.3423272.
- [146] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.
- [147] B. Sudret and A. Der Kiureghian. *Stochastic finite element methods and reliability: a state-of-the-art report*. Department of Civil and Environmental Engineering, University of California, 2000.
- [148] T. Sullivan. *Introduction to Uncertainty Quantification*, volume 63 of *Texts in Applied Mathematics*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-23394-9 978-3-319-23395-6. doi: 10.1007/978-3-319-23395-6.
- [149] T. J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Problems & Imaging*, 11(5):857–874, 2017. ISSN 1930-8345. doi: 10.3934/ipi.2017040.
- [150] P. Thomas Fletcher. Geodesic Regression and the Theory of Least Squares on Riemannian Manifolds. *Int J Comput Vis*, 105(2):171–185, Nov. 2013. ISSN 1573-1405. doi: 10.1007/s11263-012-0591-y.
- [151] L. Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994. ISSN 0090-5364.

- [152] F. Tronarp, H. Kersting, S. Särkkä, and P. Hennig. Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: A new perspective. *Statistics and Computing*, 29:1297–1315, 2019.
- [153] L. W. Tu. *An Introduction to Manifolds*. Universitext. Springer, New York, NY, 2011. ISBN 978-1-4419-7399-3 978-1-4419-7400-6. doi: 10.1007/978-1-4419-7400-6.
- [154] H. Tyagi and V. Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Applied and Computational Harmonic Analysis*, 37(3):389–412, Nov. 2014. ISSN 1063-5203. doi: 10.1016/j.acha.2014.01.002.
- [155] K. R. Ulrich, D. E. Carlson, K. Dzirasa, and L. Carin. GP Kernels for Cross-Spectrum Analysis. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [156] C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren Der Mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.
- [157] M. Villani. Bayesian reference analysis of cointegration. *Econometric Theory*, 21(2):326–357, 2005.
- [158] C. Y. Wong, P. Seshadri, G. Parks, and M. Girolami. Embedded Ridge Approximations. *arXiv:1907.07037 [cs, math]*, Aug. 2020.
- [159] J. Xiong, X. Cai, and J. Li. Clustered active-subspace based local Gaussian Process emulator for high-dimensional and complex computer models. *Journal of Computational Physics*, 450: 110840, 2022.
- [160] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, July 2010. ISBN 978-1-4008-3534-8. doi: 10.1515/9781400835348.
- [161] A. Yagi. *Abstract Parabolic Evolution Equations and Their Applications*. Springer Monographs in Mathematics. Springer, Berlin, Heidelberg, 2010. ISBN 978-3-642-04630-8 978-3-642-04631-5. doi: 10.1007/978-3-642-04631-5.
- [162] Y. Yuan, H. Zhu, W. Lin, and J. S. Marron. Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):697–719, 2012. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.01022.x.
- [163] R. Zhang, S. Mak, and D. Dunson. Gaussian process subspace regression for model reduction. *arXiv preprint arXiv:2107.04668*, 2021.

- [164] Z. Zhang and G. E. Karniadakis. *Numerical Methods for Stochastic Partial Differential Equations with White Noise*, volume 196 of *Applied Mathematical Sciences*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-57510-0 978-3-319-57511-7. doi: 10.1007/978-3-319-57511-7.
- [165] R. Zimmermann. Manifold interpolation and model reduction. *arXiv preprint arXiv:1902.06502*, 2019.

List of Notation

Calculus

∂^α Classical partial derivative written in multi-index notation

$\mathcal{D}^\alpha$ Weak derivative written in multi-index notation

∇ Gradient operator

$\nabla \cdot \mathbf{v}$ Divergence of a vector field $\mathbf{v} : \mathbb{R}^d \rightarrow \mathbb{R}^d$

∂D Boundary of a bounded, connected, and open domain D

Differential Geometry

$T_p \mathcal{M}$ Tangent space of a differentiable manifold $\mathcal{M}$ at a point $p \in \mathcal{M}$

$(\langle \cdot, \cdot \rangle_p)_{p \in \mathcal{M}}$ Riemannian metric for a manifold $\mathcal{M}$

$d_{\mathcal{M}}(\cdot, \cdot)$ Riemannian distance for a Riemannian manifold $\mathcal{M}$

$\exp_p^{\mathcal{M}}$ Riemannian Exponential map

$\log_p^{\mathcal{M}}$ Riemannian Logarithm map

$\text{lift}_p(\xi)$ Horizontal lift of ξ at p

$\text{St}(d, n)$ (compact) Stiefel manifold of $d \times n$ orthonormal matrices

$\exp_{\mathbf{X}}^{\text{St}}$ Riemannian Stiefel Exponential map

$\log_{\mathbf{X}}^{\text{St}}$ Riemannian Stiefel Logarithm map

$\text{Gr}(d, n)$ Grassmann manifold of all n -dimensional (linear) subspaces of $\mathbb{R}^d$

$\exp_{\mathcal{X}}^{\text{Gr}}$ Riemannian Grassmann Exponential map

$\log_{\mathcal{X}}^{\text{Gr}}$ Riemannian Grassmann Logarithm map

Functional Analysis

$\langle \cdot, \cdot \rangle_{\mathcal{X}}$ Inner product on a vector space $\mathcal{X}$

$\mathcal{X}, \mathcal{Y}, \mathcal{E}$ Calligraphic upper-case Latin letters denoting normed spaces and/or Hilbert spaces

$L(\mathcal{X}, \mathcal{Y})$ Set of all bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$

$\mathcal{X}'$ (Topological) dual of the normed space $\mathcal{X}$

$A^\dagger$ Adjoint of a linear operator $A \in L(\mathcal{X}, \mathcal{Y})$

$\text{tr}(A)$ Trace of a linear operator A

$L^p(D)$ Space of p -integrable measurable functions on D

$L^\infty(D)$ Space of (essentially) bounded functions on D

$C(D)$ Space of continuous functions on D

$C^n(D)$ Space of continuous functions on D with n continuous derivatives

$C^\infty(D)$ Space of infinitely differentiable functions on D

$C_c^\infty(D)$ Set of functions $u \in C^\infty(D)$ with compact support

$C_b^n(D)$ Subset of $C^n(D)$ containing functions with bounded and uniformly continuous partial derivatives up to order n

$W^{k,p}(D)$ Sobolev space of integral or fractional order k

$H^k(D)$ Sobolev space $W^{k,2}(D)$ of integral or fractional order k (with $p = 2$)

$H_0^1(D)$ Space of function in $H^1(D)$ which are zero on the boundary ∂D

$H^{1/2}(\partial D)$ Trace space

$H_g^1(D)$ Space of functions in $H^1(D)$ whose *trace* on the boundary ∂D is the function g in the trace space

$H^{-1}(D)$ Dual space of $H_0^1(D)$

Linear Algebra

$\mathbf{x}, \mathbf{y}$ Boldfaced lower-case Latin letters denoting vectors in $\mathbb{R}^d$

$\mathbf{A}, \mathbf{M}$ Boldfaced upper-case Latin letters denoting matrices of arbitrary, but finite, dimension

$\mathbf{A}^-$ Generalized (pseudo) inverse of a matrix $\mathbf{A}$

$\det(\mathbf{A})$ Determinant of a square matrix $\mathbf{A}$

$\text{tr}(\mathbf{A})$	Trace of a square matrix $\mathbf{A}$
$A(U)$	Image of a subset $U \subset \mathcal{X}$ under an operator $A : \mathcal{X} \rightarrow \mathcal{Y}$
$\text{im}(A)$	Image of a linear operator $A : \mathcal{X} \rightarrow \mathcal{Y}$ between two vector spaces
$\ker(A)$	Kernel/null-space of a linear operator $A : \mathcal{X} \rightarrow \mathcal{Y}$ between two vector spaces
$\otimes$	Outer product for vectors/Kroenecker product for matrices
vec	Vectorisation operator
unvec	Inverse of the vectorisation operator
$O(d)$	Orthogonal group of $d \times d$ orthogonal matrices

Sets of Numbers

$\mathbb{N}$	The set of natural numbers, $\mathbb{N} = \{1, 2, \dots\}$
$\mathbb{Z}$	The set of integers
$\mathbb{N}_0$	The set of non-negative integers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$
$\mathbb{R}$	The set of real numbers

Probability

$\mathbb{1}_A$	Indicator function of a set A
μ, ν	Lower-case Greek letters denoting (probability) measures
$\mathcal{B}(\mathcal{E})$	Borel sigma-algebra of $\mathcal{E}$
$\mathcal{P}(\mathcal{E})$	Set of Borel probability measures on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$
$X_{\#}\mu$	Pushforward of the measure μ by the measurable function X
X, Y	Upper-case Latin letters denoting random variables
$\frac{d\mu}{d\nu}$	Radon-Nikodym derivative of μ w.r.t. ν
$\mathbb{E}$	Expectation
$\mathbb{E}_{\mu}$	Expectation with respect to a measure μ
$\mathbb{V}$	Variance
$\mathbb{V}_{\mu}$	Variance of a random variable with respect to a measure μ
$\Gamma(\mu, \nu)$	Set of couplings between two probability measures μ, ν

μ^* Centered measure associated with a probability measure μ

$\mathcal{N}(\mu, \sigma^2)$ Univariate normal distribution with mean μ and variance σ^2

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$\mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ Matrix Normal Distribution with parameters $\mathbf{M}, \mathbf{U}, \mathbf{V}$

$\mathcal{GP}(m, k)$ Gaussian process with mean function m and covariance function k

$\text{Mat } \mathcal{GP}(\boldsymbol{\mu}, k, \boldsymbol{\Omega})$ Matrix-valued Gaussian Process with mean function $\boldsymbol{\mu}$, kernel function k , and parameter covariance matrix $\boldsymbol{\Omega}$

$\text{KL}(\mu \parallel \nu)$ Kullback-Leibler divergence from the probability measure μ to ν

$W_2(\mu, \nu), W(\mu, \nu)$ Wasserstein-2 distance between two probability measures μ, ν

$d_P(\cdot, \cdot)$ Procrustes distance between covariance operators

Appendix A

Additional Background Material

A.1 Analysis

In this section we introduce and discuss any additional concepts in functional analysis which will be helpful for this thesis. We will focus on concepts which are especially relevant for the field of Functional Data Analysis (FDA) and for the study of PDEs and FEM. For a more in-depth treatment of functional analysis the reader is referred to [87, 43, 13].

A.1.1 Basics of Functional Analysis

Recall that a *metric* on a set V can be defined as follows:

Definition A.1 (*Metric/Metric Space*). A *metric* on a set V is a function $d : V \times V \rightarrow \mathbb{R}$ satisfying the following properties for all $x, y, z \in V$:

1. **Positivity:** $d(x, y) \geq 0$.
2. **Symmetry:** $d(x, y) = d(y, x)$.
3. **Triangle Inequality:** $d(x, z) \leq d(x, y) + d(y, z)$.
4. **Separating:** $d(x, y) = 0 \iff x = y$.

If d is a metric we call the pair (V, d) a *metric space*. A function d is called a *pseudometric* if it satisfies all of the above properties except Item 4. ♦

Recall that a *norm* on a (real) vector space V can be defined as follows:

Definition A.2 (*Norm/Normed Vector Space*). A *norm* on a (real) vector space V is a function $\|\cdot\| : V \rightarrow \mathbb{R}$ satisfying the following properties for all $x, y \in V$ and for all scalars $\alpha \in \mathbb{R}$:

1. **Triangle Inequality:** $\|x + y\| \leq \|x\| + \|y\|$.

2. **Absolute Homogeneity:** $\|\alpha x\| = |\alpha|\|x\|$.

3. **Positive Definiteness/Separating:** $\|x\| = 0 \implies x = 0$.

A vector space equipped with a norm, $(V, \|\cdot\|)$, is called a *normed vector space*. ♦

Remark A.1. A norm provides a notion of **distance** turning the vector space V into a *metric space*. In particular, the function $d : V \times V \rightarrow \mathbb{R}$ defined via $d(x, y) := \|x - y\|$ can be shown to be a metric. ♦

Recall that a *bounded linear operator* $A : \mathcal{X} \rightarrow \mathcal{Y}$ between two normed spaces $\mathcal{X}, \mathcal{Y}$ can be defined as follows:

Definition A.3 (Bounded linear operator). A linear operator $A : \mathcal{X} \rightarrow \mathcal{Y}$ between two normed spaces $\mathcal{X}, \mathcal{Y}$ is said to be *bounded* if there is some constant $C > 0$ such that for all $x \in \mathcal{X}$ we have:

$$\|Ax\|_{\mathcal{Y}} \leq C\|x\|_{\mathcal{X}}. \quad (\text{A.1})$$

♦

Recall that an *inner product* on a (real) vector space V can be defined as follows:

Definition A.4 (Inner Product/Inner Product Space). An *inner product* on a (real) vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ satisfying the following properties for all $x, y, w \in V$ and for all scalars $\alpha, \beta \in \mathbb{R}$:

1. **Symmetry:** $\langle x, y \rangle = \langle y, x \rangle$.

2. **Linearity:** $\langle \alpha x + \beta y, w \rangle = \alpha \langle x, w \rangle + \beta \langle y, w \rangle$.

3. **Positive Definiteness:** $\langle x, x \rangle \geq 0$ with equality if and only if $x = 0 \in V$.

A vector space equipped with an inner product is called an *inner product space*. ♦

Note that the definition of norms and inner products can be generalised to a vector space over a general field $\mathbb{F}$; a commonly used generalization being for complex vector spaces ($\mathbb{F} = \mathbb{C}$). In this thesis we will exclusively be concerned with the case of real vector spaces.

Remark A.2. An inner product defines a norm on V via $\|x\| := \sqrt{\langle x, x \rangle}$. Thus, any inner product space can be trivially made into a normed space. A norm defined via an inner product in this way satisfies the important and well-known **Cauchy-Schwarz Inequality**:

$$|\langle x, y \rangle| \leq \|x\| \|y\|, \quad \forall x, y \in V, \quad (\text{A.2})$$

with equality if and only if x and y are linearly dependent. ♦

Remark A.3. It is important to note that not every normed space is an inner product space. A simple way to see this is to recall that the *parallelogram law* holds in any inner product space, $(V, \langle \cdot, \cdot \rangle)$, with the associated norm $\|x\| = \sqrt{\langle x, x \rangle}$, i.e. the following holds for all $x, y \in V$:

$$2\|x\|^2 + 2\|y\|^2 = \|x + y\|^2 + \|x - y\|^2. \quad (\text{A.3})$$

We can use this fact to find a normed space which is not an inner product space. For example, consider the plane, $\mathbb{R}^2$, equipped with the 1-norm:

$$\|x\|_1 := \sum_{i=1}^2 |x_i|. \quad (\text{A.4})$$

This is not an inner product space since the parallelogram law does not hold: take $x = (-1, -1)^T$ and $y = (-1, 1)^T$ in Eq. (A.3). We obtain:

$$2\|x\|_1^2 + 2\|y\|_1^2 = 16, \quad (\text{A.5})$$

$$\|x + y\|_1^2 + \|x - y\|_1^2 = 8, \quad (\text{A.6})$$

and so the parallelogram law does not hold for the normed space $(\mathbb{R}^2, \|\cdot\|_1)$. Thus, no inner product gives rise to the $\|\cdot\|_1$ norm. ♦

Two important definitions are those of *Banach and Hilbert spaces*. These can be defined as follows:

Definition A.5 (*Banach Space*). A normed vector space $(\mathcal{X}, \|\cdot\|)$ is called a *Banach space* if it is complete with respect to its norm, i.e. if every Cauchy sequence in $\mathcal{X}$ converges to some point in $\mathcal{X}$. ♦

Definition A.6 (*Hilbert Space*). An inner product space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is called a *Hilbert space* if it is complete with respect to the norm induced by its inner product. ♦

Note that any Hilbert space is automatically a Banach space.

A.1.2 Fractional Sobolev Spaces

In this section we define fractional Sobolev spaces. We follow the presentation in [45].

We start off by defining the fractional Sobolev spaces, $W^{s,p}(D)$, for fractional exponents, s , restricted to the interval $s \in (0, 1)$. We then extend the definition to all exponents $s > 1$.

Definition A.7 (*Fractional Sobolev Space, case $s \in (0, 1)$*). For any $p \in [1, \infty)$ and for

$s \in (0, 1)$, we define the *fractional Sobolev space*, $W^{s,p}(D)$, as follows:

$$W^{s,p}(D) := \left\{ u \in L^p(D) \left| \left((\mathbf{x}, \mathbf{y}) \mapsto \frac{|u(\mathbf{x}) - u(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^{\frac{d}{p} + s}} \right) \in L^p(D \times D) \right. \right\}, \quad (\text{A.7})$$

i.e. $W^{s,p}(D)$ is an interpolated Banach space “between” the spaces $L^p(D)$ and $W^{1,p}(D)$, endowed with the following natural norm:

$$\|u\|_{W^{s,p}(D)} := \left(\|u\|_{L^p(D)}^p + \int_D \int_D \frac{|u(\mathbf{x}) - u(\mathbf{y})|^p}{\|\mathbf{x} - \mathbf{y}\|^{d+sp}} d\mathbf{x} d\mathbf{y} \right)^{\frac{1}{p}}, \quad (\text{A.8})$$

where the term,

$$[u]_{W^{s,p}(D)} := \left(\int_D \int_D \frac{|u(\mathbf{x}) - u(\mathbf{y})|^p}{\|\mathbf{x} - \mathbf{y}\|^{d+sp}} d\mathbf{x} d\mathbf{y} \right)^{\frac{1}{p}}, \quad (\text{A.9})$$

is the so-called *Gagliardo semi-norm* of u . ♦

This definition cannot be extended to the case $s > 1$ (for details on this see the discussion following Proposition 2.2 in [45]) and we instead have to consider the following modification to define fractional Sobolev spaces for exponents $s > 1$:

Definition A.8 (*Fractional Sobolev Space, case $s > 1$*). Let $p \in [1, \infty)$ and $s > 1$ with $s \notin \mathbb{N}$. We write $s = m + \sigma$, where $m \in \mathbb{N}$ and $\sigma \in (0, 1)$. We define the *fractional Sobolev space*, $W^{s,p}(D)$, as the space consisting of equivalence classes of functions in the integral order Sobolev space (see Definition 2.2), $u \in W^{m,p}(D)$, whose weak derivatives, $\mathcal{D}^\alpha u$, with $|\alpha| = m$, belong to the fractional Sobolev space $W^{\sigma,p}(D)$ defined in Definition A.7, i.e.

$$W^{s,p}(D) := \{u \in W^{m,p}(D) \mid \mathcal{D}^\alpha u \in W^{\sigma,p}(D) \text{ for any multi-index } \alpha \text{ such that } |\alpha| = m\}. \quad (\text{A.10})$$

This space is also a Banach space equipped with the following norm:

$$\|u\|_{W^{s,p}(D)} := \left(\|u\|_{W^{m,p}(D)}^p + \sum_{|\alpha|=m} \|\mathcal{D}^\alpha u\|_{W^{\sigma,p}(D)}^p \right)^{\frac{1}{p}}. \quad (\text{A.11})$$

Note that if $s = m$ is an integer then the space $W^{s,p}(D)$ coincides with the usual integral order Sobolev space $W^{m,p}(D)$ (see Definition 2.2). ♦

Note that Definitions A.7 and A.8 were both adopted from Section 2 of [45]. As with integral order Sobolev spaces, in the special case of $p = 2$ the fractional Sobolev spaces turn out to also be Hilbert spaces and we denote them via the notation $H^s(D) := W^{s,2}(D)$. We will also abbreviate the norms and Gagliardo semi-norm of the spaces $H^s(D)$ to $\|\cdot\|_s, [\cdot]_s$ respectively.

Remark A.4. There is an alternative formulation of fractional Sobolev spaces via Fourier transforms. The two notions can be shown to be equivalent. For more details see Section 3 of [45] and

Sections 1.11.3-1.11.5 of [161]. ♦

A.2 PDEs and FEM

In this section we include any additional relevant concepts from PDE and FEM theory which are helpful for this thesis.

A.2.1 Additional Concepts from PDE Theory

The Riesz Representation Theorem

The Riesz Representation Theorem is an important theoretical tool in functional analysis and for completeness we include the theorem below (adopted from Theorem 1.57 in [101]).

Theorem A.1 (*Riesz Representation*). *Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and let $\ell \in H'$ be a bounded linear functional on H . There exists a unique $u_\ell \in H$ which represents ℓ , i.e.*

$$\langle u_\ell, x \rangle = \ell(x), \quad \forall x \in H. \quad (\text{A.12})$$

♦

The Lax-Milgram Lemma

The Lax-Milgram Lemma [91] is an important theoretical tool in the existence and uniqueness theory for PDEs. Before presenting the lemma we will first include, for completeness, definitions of the terms *bounded* and *coercive* for a bilinear form. These definitions are adopted from Definition 1.58 in [101].

Definition A.9 (*Bounded bilinear form*). Let X, Y, Z be three real vector spaces. A bilinear form $a : X \times Y \rightarrow Z$ is said to be *bounded* if there exists some $\alpha > 0$ such that,

$$\|a(x, y)\|_Z \leq \alpha \|x\|_X \|y\|_Y, \quad \forall x \in X, y \in Y. \quad (\text{A.13})$$

♦

Definition A.10 (*Coercive bilinear form*). Let X be a real vector space. A bilinear form $a : X \times X \rightarrow \mathbb{R}$ is said to be *coercive* if there exists some $\beta > 0$ such that,

$$a(x, x) \geq \beta \|x\|_X^2, \quad \forall x \in X. \quad (\text{A.14})$$

♦

We can now present the Lax-Milgram Lemma (adopted from Lemma 1.59 in [101]):

Lemma A.1 (*Lax-Milgram*). Let H be a real Hilbert space with norm $\|\cdot\|$ and let $\ell : H \rightarrow \mathbb{R}$ be a bounded linear functional on H . Let $a : H \times H \rightarrow \mathbb{R}$ be a bilinear form that is bounded and coercive. There exists a unique $u_\ell \in H$ such that $a(u_\ell, x) = \ell(x)$ for all $x \in H$. $\blacklozenge$

Weyl's Law

A well known-result regarding the asymptotic properties of the eigenvalues of the elliptic differential operator $\mathcal{L}$ given in Eq. (2.27) is Weyl's law which is presented below (adopted from Lemma 1 in [39]):

Lemma A.2 (*Weyl's Law*). Assuming Assumptions 2.1 and 2.4 hold we have that the eigenvalues $\{\lambda_i\}_{i=1}^\infty$ (in non-decreasing order) satisfy,

$$C_l i^{2/d} \leq \lambda_i \leq C_u i^{2/d}, \quad \forall i \in \mathbb{N}, \quad (\text{A.15})$$

where $C_l, C_u > 0$ are constants (which depend on κ and D). $\blacklozenge$

Fractional powers of elliptic differential operators

In the additional details for the proof of Theorem 3.3 we will require the notion of fractional powers of the elliptic differential operator, $\mathcal{L}$, given in Eq. (2.27). We follow the presentation in [39]. Denoting by $\{e_i\}_{i=1}^\infty$ the set of orthonormal eigenfunctions of the operator $\mathcal{L}$ with corresponding (in nondecreasing order) eigenvalues $\{\lambda_i\}_{i=1}^\infty$ we can define the fractional power operator $\mathcal{L}^s$ for $s > 0$ via the spectral representation:

$$\mathcal{L}^s \varphi := \sum_{i=1}^{\infty} \lambda_i^s \langle \varphi, e_i \rangle_{L^2(D)} e_i, \quad \forall \varphi \in \mathcal{D}(\mathcal{L}^s), \quad (\text{A.16})$$

where $\mathcal{D}(\mathcal{L}^s)$ is the domain of $\mathcal{L}^s$ defined as:

$$\mathcal{D}(\mathcal{L}^s) := \left\{ \varphi \in L^2(D) \left| \sum_{i=1}^{\infty} \lambda_i^{2s} \langle \varphi, e_i \rangle_{L^2(D)}^2 < \infty \right. \right\}. \quad (\text{A.17})$$

We will denote by $\dot{H}_{\mathcal{L}}^s := \mathcal{D}(\mathcal{L}^{s/2}) \subset L^2(D)$. This is itself a Hilbert space with inner product given by $\langle \varphi, \psi \rangle_{s, \mathcal{L}} := \langle \mathcal{L}^{s/2} \varphi, \mathcal{L}^{s/2} \psi \rangle_{L^2(D)}$ [15, 39]. We denote the corresponding norm by $\|\cdot\|_{s, \mathcal{L}}$.

Note that due to the spectral representation Eq. (A.16) we have:

$$\langle \varphi, \psi \rangle_{s, \mathcal{L}} = \sum_{i=1}^{\infty} \lambda_i^s \langle \varphi, e_i \rangle_{L^2(D)} \langle \psi, e_i \rangle_{L^2(D)}, \quad \forall \varphi, \psi \in \dot{H}_{\mathcal{L}}^s. \quad (\text{A.18})$$

We will further denote by $\dot{H}_{\mathcal{L}}^0 := L^2(D)$ and for $s > 0$, we will denote the dual space of $\dot{H}_{\mathcal{L}}^s$ by

$\dot{H}_{\mathcal{L}}^{-s} := (\dot{H}_{\mathcal{L}}^s)'$ after identification via the inner product on $L^2(D)$ which is continuously extended to a duality pairing, yielding the following norm on $\dot{H}_{\mathcal{L}}^{-s}$:

$$\|\varphi\|_{-s, \mathcal{L}} := \left(\sum_{i=1}^{\infty} \lambda_i^{-s} \langle \varphi, e_i \rangle_{L^2(D)}^2 \right)^{\frac{1}{2}}. \quad (\text{A.19})$$

An important fact regarding these Hilbert spaces is that under our Assumptions 2.1 and 2.4 we have that $\dot{H}_{\mathcal{L}}^1 = H_0^1(D)$ and $\dot{H}_{\mathcal{L}}^s \subset H^s(D)$ for all $s \geq 0$ [39, Lemma 2].

$H^{1+\alpha}(\mathbf{D})$ -regular

In the additional details for the proof of Theorem 3.3 we will require the notion of $H^{1+\alpha}(D)$ -regularity of the differential operator, $\mathcal{L}$, associated with the elliptic BVP from Eqs. (2.27) and (2.28), where we recall that we are considering $g = 0$ for the boundary condition. This definition is adopted from Definition 4 in [39]:

Definition A.11 ($H^{1+\alpha}(\mathbf{D})$ -regular). Assume that Assumption 2.4 holds and that D has a Lipschitz continuous boundary ∂D . Let $0 < \alpha \leq 1$ and $\mathcal{L}$ be the second order differential operator in Eq. (2.27). The elliptic problem associated with $\mathcal{L}$ is said to be $H^{1+\alpha}(D)$ -regular if the restriction of $\mathcal{L} : H_0^1(D) \rightarrow H_0^1(D)'$ to $H_0^1(D) \cap H^{1+\alpha}(D)$ is a continuous map to $\dot{H}_{\mathcal{L}}^{-1+\alpha} = (\dot{H}_{\mathcal{L}}^{1-\alpha})'$, and if further the data-to-solution map $\mathcal{L}^{-1} : f \mapsto \mathcal{L}^{-1}f$ is a bounded linear operator as a mapping from $\dot{H}_{\mathcal{L}}^{-1+\alpha}$ to $H_0^1(D) \cap H^{1+\alpha}(D)$. ♦

A.2.2 Additional Concepts from FEM Theory

Orthogonal projection onto V_h

We will have need for the $L^2(D)$ -orthogonal projection onto the finite dimensional space V_h from Section 2.2.2. Following the notation in [39] we denote this by

$$\Pi_h : L^2(D) \rightarrow V_h. \quad (\text{A.20})$$

As the authors Cox and Kirchner note we can uniquely extend this orthogonal projection to a bounded linear operator $\Pi_h : H^{-1} \rightarrow V_h$.

Galerkin discretisation of $\mathcal{L}$

Following the notation in [39] we denote by $L_h : V_h \rightarrow V_h$ the Galerkin discretisation of the elliptic differential operator $\mathcal{L}$ given in Eq. (2.27). This is defined as follows:

Definition A.12 (*Galerkin discretisation of $\mathcal{L}$*). The *Galerkin discretisation* of $\mathcal{L}$ is the operator $L_h : V_h \rightarrow V_h$ defined via the following equation:

$$\langle L_h \varphi_h, \psi_h \rangle_{L^2(D)} = \langle \mathcal{L} \varphi_h, \psi_h \rangle_{H^{-1}(D) \times H_0^1(D)}, \quad \forall \varphi_h, \psi_h \in V_h, \quad (\text{A.21})$$

where $\langle \cdot, \cdot \rangle_{H^{-1}(D) \times H_0^1(D)}$ is the duality pairing between $H_0^1(D)$ and $H^{-1}(D)$. $\blacklozenge$

Remark A.5. Note that one can also define fractional powers of L_h via a spectral approach similar to the case for $\mathcal{L}$ (see Eq. (A.16)). Since L_h is a self-adjoint linear operator between finite dimensional vector spaces this is done in the same way as defining fractional powers of symmetric matrices: there exists an orthonormal eigenbasis of V_h consisting of eigenvectors $\{e_{i,h}\}_{i=1}^{n_u}$ of L_h together with their corresponding eigenvalues $\{\lambda_{i,h}\}_{i=1}^{n_u}$. We can then define L_h^s for $s \in (0, \infty)$ via:

$$L_h^s \varphi_h := \sum_{i=1}^{n_u} \lambda_{i,h}^s \langle \varphi_h, e_{i,h} \rangle_{L^2(D)} e_{i,h}, \quad \forall \varphi_h \in V_h, \quad (\text{A.22})$$

where we note that due to the finite-dimensional nature of V_h there is no restriction on the domain of L_h^s as opposed to the case of $\mathcal{L}^s$. $\blacklozenge$

$H^1(D)$ -stable

The following definition (adopted from [39]) will be useful in Appendix B.2:

Definition A.13 ($H^1(D)$ -stable). We say that Π_h is $H^1(D)$ -stable if there exists a constant $C_\Pi > 0$ such that,

$$\|\Pi_h\|_{H^1(D) \rightarrow H^1(D)} \leq C_\Pi, \quad (\text{A.23})$$

for all $h > 0$ sufficiently small. $\blacklozenge$

A.3 Probability Theory

In this section we include any additional relevant concepts from probability theory and inference which are helpful for this thesis and not included in the main text.

A.3.1 Additional Background Material

Probability space

A probability space is a triplet $(\Omega, \mathcal{F}, \mu)$ consisting of:

- a *sample space*, Ω , which is the set of all possible outcomes,
- a *sigma algebra*, $\mathcal{F}$, which is the set of all allowable events (subsets of Ω),

- a *probability measure*, $\mu : \mathcal{F} \rightarrow [0, 1]$, which quantifies the probability (measure) of events $E \in \mathcal{F}$.

An event $E \in \mathcal{F}$ will also be referred to as a measurable set.

Conditional probability

Given a probability space $(\Omega, \mathcal{F}, \mu)$ and an event $B \in \mathcal{F}$ with $\mu(B) > 0$ we define the conditional probability measure given the event B , $\mu(\cdot|B)$, as follows:

$$\mu(A|B) := \frac{\mu(A \cap B)}{\mu(B)}, \quad \forall A \in \mathcal{F}. \quad (\text{A.24})$$

This function, $\mu(\cdot|B) : \mathcal{F} \rightarrow [0, 1]$, is itself a probability measure and it can be thought of as quantifying how the probability of events change once we know the event B has happened. This definition, and its consequences will be expanded on in Section 2.3.3.

Mean and Covariance of Random Variables

The *expectation* of X with respect to the probability measure μ , assuming it exists, will be denoted by $\mathbb{E}_\mu(X)$ and is given by:

$$\mathbb{E}_\mu(X) := \int_{\Omega} X(\omega) d\mu(\omega), \quad (\text{A.25})$$

where the integral above is to be interpreted appropriately according to the nature of the space $\mathcal{E}$ in which the random variable takes values (see the discussion on Bochner and Pettis integrals in Section 2.3.4).

When the probability measure with which we are taking the expectation with respect to is clear we will drop the μ and just write $\mathbb{E}(X)$.

The *mean* of a random variable, assuming it exists, is the expectation of the random variable with respect to its law, i.e. the mean of an $\mathcal{E}$ -valued random variable, X , is the element $m \in \mathcal{E}$ such that $m = \mathbb{E}(X)$, where the expectation is with respect to the law μ_X (see Remark 2.8).

When $\mathcal{E}$ is a finite dimensional vector space the *covariance* of an $\mathcal{E}$ -valued random variable X , assuming it exists, is the matrix $\text{cov}(X) \in \mathcal{E} \otimes \mathcal{E}$ given by:

$$\text{cov}(X) := \mathbb{E}[(X - \mathbb{E}(X)) \otimes (X - \mathbb{E}(X))]. \quad (\text{A.26})$$

If Y is another $\mathcal{E}$ -valued random variable then we denote the covariance between X and Y by the following matrix:

$$\text{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}(X)) \otimes (Y - \mathbb{E}(Y))]. \quad (\text{A.27})$$

Remark A.6. In the common case where $\mathcal{E} = \mathbb{R}^d$, for some $d \in \mathbb{N}$, it is common to denote the random variables above as random vectors, $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$. In this common scenario the mean of $\mathbf{X}$ is the vector $\mathbf{m} \in \mathbb{R}^d$ whose i -th entry is the mean of the i -th component of $\mathbf{X}$, i.e. $m_i := \mathbb{E}(X_i)$. Further, the covariance between $\mathbf{X}, \mathbf{Y}$, corresponding to Eq. (A.27) is the $d \times d$ matrix given by:

$$\text{cov}(\mathbf{X}, \mathbf{Y}) := \mathbb{E}[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))^T] \quad (\text{A.28})$$

and thus the ij -th entry of $\text{cov}(\mathbf{X}, \mathbf{Y})$ is simply the covariance between the i -th entry of $\mathbf{X}$, and the j -th entry of $\mathbf{Y}$, i.e. $\text{cov}(\mathbf{X}, \mathbf{Y})_{ij} = \text{cov}(X_i, Y_j)$. Similarly the covariance of $\mathbf{X}$ is simply $\text{cov}(\mathbf{X}) = \text{cov}(\mathbf{X}, \mathbf{X})$. ♦

Remark A.7. If $\mathbf{X}, \mathbf{Y}$ are two **independent** random vectors then $\text{cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$. The reverse implication is **not** true in general. ♦

We will sometimes also refer to the moments (both centered and uncentered) of a probability measure ν . This is to be interpreted as the corresponding moment of a random variable X distributed according to the law ν , i.e. $X \sim \nu$.

Properties of the multivariate normal distribution

We now collate several useful properties of the multivariate normal distribution which will be useful throughout the thesis.

Proposition A.1 (Adopted from Theorem 1.2.6 from [109]). *If $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a fixed deterministic matrix, and $\mathbf{b} \in \mathbb{R}^m$ is a fixed deterministic vector then,*

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b} \sim \mathcal{N}_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T), \quad (\text{A.29})$$

i.e. an affine transformation of a multivariate normal random vector remains multivariate normal.

♦

Proof of Proposition A.1.

The fact that $\mathbf{Y}$ is multivariate normal follows directly from Definition 2.7 and the following computation of the inner product of $\mathbf{Y}$ with an arbitrary fixed deterministic vector $\mathbf{a} \in \mathbb{R}^m$:

$$\langle \mathbf{a}, \mathbf{Y} \rangle = \langle \mathbf{a}, \mathbf{A}\mathbf{X} + \mathbf{b} \rangle \quad (\text{A.30})$$

$$= \langle \mathbf{a}, \mathbf{A}\mathbf{X} \rangle + \langle \mathbf{a}, \mathbf{b} \rangle \quad (\text{A.31})$$

$$= \langle \mathbf{A}^T \mathbf{a}, \mathbf{X} \rangle + \langle \mathbf{a}, \mathbf{b} \rangle. \quad (\text{A.32})$$

Since $\mathbf{X}$ is a multivariate normal random vector and $\mathbf{A}$ is deterministic it follows that $\langle \mathbf{A}^T \mathbf{a}, \mathbf{X} \rangle$ is

a univariate normal random variable with mean $\mathbf{a}^T \mathbf{A} \boldsymbol{\mu}$ and variance $\mathbf{a}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{A}^T \mathbf{a}$. Thus, using the well known transformation properties of the univariate normal distribution we have:

$$\langle \mathbf{a}, \mathbf{Y} \rangle \sim \mathcal{N}(\mathbf{a}^T (\mathbf{A} \boldsymbol{\mu} + \mathbf{b}), \mathbf{a}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{A}^T \mathbf{a}). \quad (\text{A.33})$$

Since $\mathbf{a}$ is arbitrary we have $\mathbf{Y} \sim \mathcal{N}_m(\mathbf{A} \boldsymbol{\mu} + \mathbf{b}, \mathbf{A}^T \boldsymbol{\Sigma} \mathbf{A})$ as required. $\blacksquare$

Proposition A.2 (Adopted from Theorem 1.2.7 from [109]). *If $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the marginal distribution of any subset of $k < d$ elements of $\mathbf{X}$ is itself a k -dimensional multivariate normal distribution. In particular, if $\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ are partitioned as follows:*

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad (\text{A.34})$$

where $\mathbf{X}_1, \boldsymbol{\mu}_1 \in \mathbb{R}^k$, and $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{k \times k}$ then $\mathbf{X}_1 \sim \mathcal{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. $\blacklozenge$

Proof of Proposition A.2.

Follows directly from Definition 2.7. For more details see proof in [109, Theorem 1.2.7]. $\blacksquare$

Remark A.8. Note that Proposition A.2 implies that if $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ for $i = 1, \dots, d$, i.e. the components of a multivariate normal random vector are each marginally normally distributed with mean and variance given by the appropriate entries of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Note: the reverse implication is **not** true in general. For a counter example please see below. $\blacklozenge$

Counter example for Remark A.8.

We now present a counter example adopted from [122] which demonstrates that the reverse implication in Remark A.8 is **not** true in general, i.e. that a collection of univariate normal random variables need not be jointly multivariate normal. Let $X \sim \mathcal{N}(0, 1)$ and $W \sim \text{Bernoulli}(\frac{1}{2})$ be independent random variables. Let Y be a random variable defined as follows:

$$Y := \begin{cases} X, & \text{if } W = 0, \\ -X, & \text{if } W = 1. \end{cases} \quad (\text{A.35})$$

Due to the symmetry of the standard normal distribution we have that $-X \sim \mathcal{N}(0, 1)$. This allows

us to compute the cumulative distribution function of Y as follows:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(Y \leq y | W = 0)\mathbb{P}(W = 0) + \mathbb{P}(Y \leq y | W = 1)\mathbb{P}(W = 1) \quad (\text{A.36})$$

$$= \frac{1}{2}\mathbb{P}(X \leq y | W = 0) + \frac{1}{2}\mathbb{P}(-X \leq y | W = 1) \quad (\text{A.37})$$

$$= \frac{1}{2}\mathbb{P}(X \leq y) + \frac{1}{2}\mathbb{P}(-X \leq y) \quad (\text{since } X \perp\!\!\!\perp W) \quad (\text{A.38})$$

$$= \frac{1}{2}\Phi(y) + \frac{1}{2}\Phi(y) \quad (\text{A.39})$$

$$= \Phi(y), \quad (\text{A.40})$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Thus, $Y \sim \mathcal{N}(0, 1)$. Now note that $V = X + Y$ has the following form:

$$V := \begin{cases} 2X, & \text{if } W = 0, \\ 0, & \text{if } W = 1, \end{cases} \quad (\text{A.41})$$

i.e. V is an equal mixture of a point mass at 0 and a $\mathcal{N}(0, 4)$ random variable. Thus, the vector $\mathbf{X} := (X, Y)^T$ has components which are both normally distributed while the vector is **not** multivariate-normal. ■

As is discussed in *Remark A.7* if two random variables are independent then the covariance between them, provided it exists, is zero and the converse is not generally true. The converse however holds for multivariate-normal random vectors as the following proposition (adopted from [109]) shows:

Proposition A.3. *If $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ are again partitioned as in Eq. (A.34) then the subvectors $\mathbf{X}_1, \mathbf{X}_2$ are independent if and only if the off-diagonal term in the covariance of $\mathbf{X}$ is zero, i.e. $\boldsymbol{\Sigma}_{12} = \mathbf{0}$. Note that $\boldsymbol{\Sigma}_{12} = \text{cov}(\mathbf{X}_1, \mathbf{X}_2)$ so this is equivalent to saying that the subvectors are independent if and only if the covariance between them is zero. ♦*

Proof of Proposition A.3.

This is a well known result and one can consult the proof of Theorem 1.2.8 in [109]. ■

A key property of the multivariate normal distribution which will be of great use in this thesis is the existence of closed-form expressions for its conditional distributions. This is presented in the following classical result, which can be found in [109, Theorem 1.2.11] as well as in [52, Proposition 3.13]:

Proposition A.4. *Let $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and partition $\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ as in Eq. (A.34). Let $\boldsymbol{\Sigma}_{22}^-$ be a*

generalized (pseudo) inverse of Σ_{22} , i.e. a matrix satisfying

$$\Sigma_{22}\Sigma_{22}^-\Sigma_{22} = \Sigma_{22}. \quad (\text{A.42})$$

Then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is:

$$\mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}_k(\boldsymbol{\mu}_{1\cdot 2}, \Sigma_{11\cdot 2}), \quad (\text{A.43})$$

where,

$$\boldsymbol{\mu}_{1\cdot 2} := \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^-(\mathbf{X}_2 - \boldsymbol{\mu}_2), \quad (\text{A.44})$$

$$\Sigma_{11\cdot 2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21}. \quad (\text{A.45})$$

◆

The proof of this result is presented below after a few remarks.

Remark A.9. In the special case of the submatrix Σ_{22} being non-singular, i.e. $\Sigma_{22}^- = \Sigma_{22}^{-1}$ one can replace the pseudo-inverse in Eqs. (A.44) and (A.45) with the regular inverse. This special case is more widely known, and its proof is slightly more straightforward. In Chapter 6 we will have need for the more general result presented in Proposition A.4 since we will be dealing with the case of degenerate multivariate and matrix normal distributions (see Section 5.1). ◆

Remark A.10. Since the matrix normal distribution (Definition 5.1) is defined in terms of the multivariate normal distribution we can utilise Proposition A.4 to condition normally distributed matrices on their sub-matrices. This observation will be useful in Chapter 6. ◆

Proposition A.4 will allow us to discuss how we can condition GPs (see Section 2.3.5) on observational data as well as make predictions with GPs, both of which will be essential tools in Chapters 3 and 6 and Section 3.5.

The proof of Proposition A.4 will require the following supporting result adopted from [109, Lemma 1.2.10] and [52, Proposition 2.16]:

Proposition A.5. *Let the $d \times d$ covariance matrix, Σ , which is non-negative definite, be partitioned as in Eq. (A.34), then the following holds:*

$$(a) \quad \ker(\Sigma_{22}) \subset \ker(\Sigma_{12}), \quad (\text{A.46})$$

$$(b) \quad \text{im}(\Sigma_{21}) \subset \text{im}(\Sigma_{22}). \quad (\text{A.47})$$

◆

Proof of Proposition A.5.

The following proof is adopted from the proof of Lemma 1.2.10 in [109]. We start with showing (a). Let $\mathbf{z} \in \ker(\Sigma_{22})$. Since Σ is non-negative definite we have, for all $\mathbf{y} \in \mathbb{R}^k$ and for all $\alpha \in \mathbb{R}$ that $\mathbf{w}^T \Sigma \mathbf{w} \geq 0$ where $\mathbf{w} \in \mathbb{R}^d$ is defined as follows:

$$\mathbf{w} := \begin{pmatrix} \mathbf{y} \\ \alpha \mathbf{z} \end{pmatrix}. \quad (\text{A.48})$$

Expanding this into the partitions yields:

$$\begin{pmatrix} \mathbf{y}^T & \alpha \mathbf{z}^T \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \alpha \mathbf{z} \end{pmatrix} = \mathbf{y}^T \Sigma_{11} \mathbf{y} + 2\alpha \mathbf{y}^T \Sigma_{12} \mathbf{z} + \alpha^2 \mathbf{z}^T \Sigma_{22} \mathbf{z} \quad (\text{A.49})$$

$$= \mathbf{y}^T \Sigma_{11} \mathbf{y} + 2\alpha \mathbf{y}^T \Sigma_{12} \mathbf{z} \quad (\text{since } \Sigma_{22} \mathbf{z} = \mathbf{0}) \quad (\text{A.50})$$

$$\geq 0. \quad (\text{A.51})$$

Since this holds for all $\mathbf{y} \in \mathbb{R}^k$ we can choose to take $\mathbf{y} = \Sigma_{12} \mathbf{z}$ which yields:

$$\mathbf{z}^T \Sigma_{21} \Sigma_{11} \Sigma_{12} \mathbf{z} + 2\alpha (\Sigma_{12} \mathbf{z})^T (\Sigma_{12} \mathbf{z}) \geq 0 \quad (\text{A.52})$$

$$(\Sigma_{12} \mathbf{z})^T \Sigma_{11} (\Sigma_{12} \mathbf{z}) + 2\alpha (\Sigma_{12} \mathbf{z})^T (\Sigma_{12} \mathbf{z}) \geq 0 \quad (\text{A.53})$$

$$(\Sigma_{12} \mathbf{z})^T (\Sigma_{11} + 2\alpha \mathbf{I}_k) (\Sigma_{12} \mathbf{z}) \geq 0. \quad (\text{A.54})$$

Note that we have utilized the fact that $\Sigma_{21} = \Sigma_{12}^T$ in the above. As this holds for all $\alpha \in \mathbb{R}$ we can deduce that $\Sigma_{12} \mathbf{z} = \mathbf{0}$, i.e. $\mathbf{z} \in \ker(\Sigma_{12})$. Thus, (a) holds as required.

Part (b) follows from noting that $\ker(\Sigma_{12})^\perp \subset \ker(\Sigma_{22})^\perp$, where $W^\perp$ denotes the orthogonal complement of W , together with the well known fact:

$$\ker(\mathbf{M})^\perp = \text{im}(\mathbf{M}^T), \quad (\text{A.55})$$

where we have also used the fact that Σ and hence Σ_{22} are symmetric. ■

We can now prove Proposition A.4.

Proof of Proposition A.4.

The following proof is adopted from the proof of Theorem 1.2.11 in [109]. From Proposition A.5 we have $\text{im}(\Sigma_{21}) \subset \text{im}(\Sigma_{22})$ and so there exists a matrix $\mathbf{B} \in \mathbb{R}^{(d-k) \times k}$ such that:

$$\Sigma_{21} = \Sigma_{22} \mathbf{B}, \quad (\text{A.56})$$

i.e. we have $\Sigma_{12} = \mathbf{B}^T \Sigma_{22}$. Thus, we can compute:

$$\Sigma_{12} \Sigma_{22}^- \Sigma_{22} = \mathbf{B}^T \Sigma_{22} \Sigma_{22}^- \Sigma_{22} = \mathbf{B}^T \Sigma_{22} = \Sigma_{12}. \quad (\text{A.57})$$

If we now let,

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_k & -\Sigma_{12} \Sigma_{22}^- \\ \mathbf{0} & \mathbf{I}_{d-k}, \end{pmatrix} \quad (\text{A.58})$$

we have from Proposition A.1 that $\mathbf{Y} = \mathbf{A}\mathbf{X}$ is distributed as follows:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \sim \mathcal{N}_d(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\Sigma\mathbf{A}^T). \quad (\text{A.59})$$

We now simplify the mean and covariance terms above:

$$\mathbf{A}\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 - \Sigma_{12} \Sigma_{22}^- \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad (\text{A.60})$$

$$\mathbf{A}\Sigma\mathbf{A}^T = \begin{pmatrix} \mathbf{I}_k & -\Sigma_{12} \Sigma_{22}^- \\ \mathbf{0} & \mathbf{I}_{d-k} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \mathbf{A}^T \quad (\text{A.61})$$

$$= \begin{pmatrix} \Sigma_{11 \cdot 2} & \Sigma_{12} - \Sigma_{12} \Sigma_{22}^- \Sigma_{22} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \mathbf{A}^T \quad (\text{A.62})$$

$$= \begin{pmatrix} \Sigma_{11 \cdot 2} & \mathbf{0} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{0} \\ -\Sigma_{22}^- \Sigma_{21} & \mathbf{I}_{d-k} \end{pmatrix} \quad (\text{A.63})$$

$$= \begin{pmatrix} \Sigma_{11 \cdot 2} & \mathbf{0} \\ \Sigma_{21} - \Sigma_{22} \Sigma_{22}^- \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (\text{A.64})$$

$$= \begin{pmatrix} \Sigma_{11 \cdot 2} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}, \quad (\text{A.65})$$

where in the covariance computation we have used Eq. (A.57) together with its transpose to simplify the off-diagonal terms in the intermediate steps to zero. Noting that $\mathbf{A}\mathbf{X}$ is equal to

$$\mathbf{A}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 - \Sigma_{12} \Sigma_{22}^- \mathbf{X}_2 \\ \mathbf{X}_2, \end{pmatrix} \quad (\text{A.66})$$

and utilising Proposition A.3 we can deduce that $\mathbf{X}_1 - \Sigma_{12} \Sigma_{22}^- \mathbf{X}_2$ is independent from $\mathbf{X}_2$. Thus, the conditional distribution of $\mathbf{X}_1 - \Sigma_{12} \Sigma_{22}^- \mathbf{X}_2$ given $\mathbf{X}_2$ is the same as that of $\mathbf{X}_1 - \Sigma_{12} \Sigma_{22}^- \mathbf{X}_2$

which implies, using Proposition A.2, that we have:

$$(\mathbf{X}_1 - \Sigma_{12}\Sigma_{22}^-\mathbf{X}_2) | \mathbf{X}_2 \sim \mathcal{N}_k(\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^-\boldsymbol{\mu}_2, \Sigma_{11.2}), \quad (\text{A.67})$$

rearranging yields:

$$\mathbf{X}_1 | \mathbf{X}_2 \sim \mathcal{N}_k(\boldsymbol{\mu} + \Sigma_{12}\Sigma_{22}^-(\mathbf{X}_2 - \boldsymbol{\mu}_2), \Sigma_{11.2}), \quad (\text{A.68})$$

i.e.

$$\mathbf{X}_1 | \mathbf{X}_2 \sim \mathcal{N}_k(\boldsymbol{\mu}_{1.2}, \Sigma_{11.2}), \quad (\text{A.69})$$

as required. ■

Conditioning of Gaussian Measures

As mentioned in Section 2.3.5 it is possible to obtain the conditional distribution of a Gaussian measure on a Hilbert space in a similar manner to Proposition 2.4 for GPs. This is presented in the following result taken from [146, see Theorem 6.20].

Theorem A.2 (Conditioning of Gaussian measures). *Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ be a separable Hilbert space with projectors $\Pi_i : \mathcal{H} \rightarrow \mathcal{H}_i$. Let $(x_1, x_2) \in \mathcal{H}_1 \oplus \mathcal{H}_2$ be an $\mathcal{H}$ -valued Gaussian random variable with mean $m = (m_1, m_2)$ and positive definite covariance operator $\mathcal{C}$. Define*

$$C_{ij} = \mathbb{E}(x_i - m_i) \otimes (x_j - m_j). \quad (\text{A.70})$$

Then the conditional distribution of x_1 given x_2 is Gaussian with mean

$$m' = m_1 + C_{12}C_{22}^{-1}(x_2 - m_2), \quad (\text{A.71})$$

and covariance operator

$$\mathcal{C}' = C_{11} - C_{12}C_{22}^{-1}C_{21}. \quad (\text{A.72})$$

◆

Note that the posterior mean and covariance operator Eqs. (A.71) and (A.72) are the analogues of the posterior mean and covariance function Eqs. (2.76) and (2.77) respectively, and they share a similar structure due to the link between GPs and Gaussian measures discussed in Section 2.3.5.

A.3.2 Additional Concepts from Probability Theory and Inference

The Feldman-Hájek Theorem

An important result in the theory of Gaussian measures is the Feldman-Hájek Theorem. For completeness we include this theorem below (adopted from Theorem 5.25 in [123]):

Theorem A.3 (Feldman-Hájek). *Let $\mu = \mathcal{N}(\mu_1, \Sigma_1)$, $\nu = \mathcal{N}(\mu_2, \Sigma_2)$ be two Gaussian measures on a Hilbert space H . The following statements hold:*

1. *The Gaussian measures μ and ν are either singular or equivalent.*
2. *They are equivalent if and only if the following conditions hold:*

- (a) $\text{im}(\Sigma_1)^{1/2} = \text{im}(\Sigma_2^{1/2}) = H_0$,
- (b) $\mu_1 - \mu_2 \in H_0$,
- (c) *The operator $T := \left(\Sigma_1^{-1/2}\Sigma_2^{1/2}\right) \left(\Sigma_1^{-1/2}\Sigma_2^{1/2}\right)^\dagger - I$, where I is the identity on H , is a Hilbert-Schmidt operator on $\overline{H_0}$.*

◆

Remark A.11. Note that the spaces $\text{im}(\Sigma_i^{1/2})$ in Theorem A.3 are the *Cameron-Martin spaces* associated with the Gaussian measures $\mathcal{N}(\mu_i, \Sigma_i)$ on the Hilbert space H . The Cameron-Martin space is an object of deep significance to the theory of Gaussian measures. For more details on such spaces the interested reader is directed to [146, 123]. ◆

The Karhunen-Loève Theorem

One very useful representation of stochastic processes is the so-called Karhunen-Loève expansion which is given by the theorem of the same name. This theorem allows certain stochastic processes to be expanded via an eigendecomposition of their covariance function and was independently discovered in [80, 99]. For completeness, we include the theorem below adopting the presentation from [148, Theorem 11.4]:

Theorem A.4 (Karhunen-Loève). *Let $U : D \rightarrow \mathbb{R}$ be a zero-mean, square-integrable stochastic process with a continuous and square-integrable covariance function, $k(\mathbf{x}, \mathbf{y}) := \text{cov}(U(\mathbf{x}), U(\mathbf{y}))$. Then we have the following expansion in $L^2(D)$,*

$$U = \sum_{n=1}^{\infty} Z_n \psi_n, \tag{A.73}$$

where the $\{\psi_n\}_{n=1}^{\infty}$ are a set of orthonormal eigenfunctions for the covariance operator K_U associated with the covariance function via the relation Eq. (2.69), while the $\{Z_n\}_{n=1}^{\infty}$ are random

variables given by the equation,

$$Z_n := \int_D U(\mathbf{x}) \psi_n(\mathbf{x}) d\mathbf{x}. \quad (\text{A.74})$$

Furthermore, the random variable $\{Z_n\}_{n=1}^\infty$ are mean zero and have the following covariance structure,

$$\mathbb{E}(Z_m Z_n) = \lambda_m \delta_{mn}, \quad (\text{A.75})$$

where $\{\lambda_n\}_{n=1}^\infty$ are the corresponding non-negative eigenvalues associated with the eigenfunctions $\{\psi_n\}_{n=1}^\infty$ and δ_{mn} is the Kronecker delta. $\blacklozenge$

MCMC algorithms

For completeness we include here in Algorithms 8 and 9 pseudo-code for the MH and HMC algorithms described in Section 5.3.

Algorithm 8: Metropolis-Hastings

Input: unnormalised target density $\pi^\dagger(\boldsymbol{\theta})$, proposal density $q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$, initialisation for $\boldsymbol{\theta}_0$,
number of MCMC iterations N

Output: samples $\{\boldsymbol{\theta}_n\}_{n=1}^N$

```

1 for  $n \leftarrow 0$  to  $N$  do
2   Sample  $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}_n, \cdot)$  ; // proposal
3   Sample  $V \sim U(0, 1)$  ;
4   if  $V \leq \min \left[ 1, \frac{\pi^\dagger(\boldsymbol{\theta}') q(\boldsymbol{\theta}', \boldsymbol{\theta}_n)}{\pi^\dagger(\boldsymbol{\theta}_n) q(\boldsymbol{\theta}_n, \boldsymbol{\theta}')} \right]$  then
5      $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}'$  ; // accept
6   else
7      $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n$  ; // reject
8   end
9 end
10 return  $\{\boldsymbol{\theta}_n\}_{n=1}^N$ 

```

Algorithm 9: Hamiltonian Monte Carlo

Input: unnormalised target density $\pi^\dagger(\boldsymbol{\theta})$, together with its gradients (either analytically or via auto-diff), momentum density $\pi(\boldsymbol{\varphi})$, number of leapfrog steps L , leapfrog step-size ϵ , initialisation for $\boldsymbol{\theta}_0$, number of MCMC iterations N

Output: samples $\{\boldsymbol{\theta}_n\}_{n=1}^N$

```

1 for  $n \leftarrow 0$  to  $N$  do
2   Sample momentum  $\boldsymbol{\varphi}_{n+1} \sim \pi(\cdot)$  ; // random momentum
3    $(\boldsymbol{\theta}', \boldsymbol{\varphi}') \leftarrow (\boldsymbol{\theta}_n, \boldsymbol{\varphi}_{n+1})$  ;
4   for  $\ell \leftarrow 1$  to  $L$  do
5      $\boldsymbol{\varphi}' \leftarrow \boldsymbol{\varphi}' + \frac{\epsilon}{2} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi^\dagger(\boldsymbol{\theta}')$  ;
6      $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}' + \epsilon \frac{\partial}{\partial \boldsymbol{\varphi}} K(\boldsymbol{\varphi}')$  ;
7      $\boldsymbol{\varphi}' \leftarrow \boldsymbol{\varphi}' + \frac{\epsilon}{2} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi^\dagger(\boldsymbol{\theta}')$  ;
8   end
9    $\boldsymbol{\varphi}' \leftarrow (-\boldsymbol{\varphi}')$  ;
10  Sample  $V \sim U(0, 1)$  ;
11  if  $V \leq \min \left[ 1, \frac{\pi^\dagger(\boldsymbol{\theta}')\pi(\boldsymbol{\varphi}')}{\pi^\dagger(\boldsymbol{\theta}_n)\pi(\boldsymbol{\varphi}_{n+1})} \right]$  then
12     $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}'$  ; // accept
13  else
14     $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n$  ; // reject
15  end
16 end
17 return  $\{\boldsymbol{\theta}_n\}_{n=1}^N$ 

```

A.4 Differential Geometry and Manifolds

In this section we include any additional relevant concepts from differential geometry which are helpful for this thesis and not included in the main text.

A.4.1 Additional Background Material

As mentioned in Section 5.4 an n -dimensional differentiable manifold $\mathcal{M}$ can intuitively be thought of as a topological space which locally resembles $\mathbb{R}^n$ and which satisfies certain smoothness properties. This intuition is made precise by stipulating that for every point $p \in \mathcal{M}$, there exists a *coordinate chart*, $x : \mathcal{D}_p \rightarrow \mathbb{R}^n$, which bijectively maps an open neighbourhood around p , $\mathcal{D}_p \subset \mathcal{M}$, to an open neighbourhood around $x(p)$, $D_{x(p)}$, together with the important extra property that charts are *compatible* in the sense that the *coordinate change* between two charts $x, \tilde{x}$,

$$x \circ \tilde{x}^{-1} : \tilde{x}(\mathcal{D}_p \cap \tilde{\mathcal{D}}_p) \rightarrow x(\mathcal{D}_p \cap \tilde{\mathcal{D}}_p), \quad (\text{A.76})$$

is a diffeomorphism provided their domains of definition overlap.

Remark A.12. Standard concepts of multivariable calculus apply to the coordinate change mapping, $x \circ \tilde{x}^{-1}$, since it maps between open domains of $\mathbb{R}^n$. The reader is referred to [2, Section 3.1.1] for further details. ♦

Embedded Submanifolds

For completeness we include a definition of n -dimensional embedded submanifolds of $\mathbb{R}^{n+d}$, adopting Definition 2.1 from [165]:

Definition A.14 (*Embedded Submanifolds of $\mathbb{R}^{n+d}$*). Let $D \subset \mathbb{R}^n$ be an open domain in $\mathbb{R}^n$. A *parameterisation* is a bijective differentiable function $f : D \rightarrow f(D) \subset \mathbb{R}^{n+d}$ with continuous inverse such that its Jacobi matrix $Df_x \in \mathbb{R}^{(n+d) \times n}$ has full rank n at every point $x \in D$. A subset $\mathcal{M} \subset \mathbb{R}^{n+d}$ is called an *n -dimensional embedded submanifold of $\mathbb{R}^{n+d}$* , if for every $p \in \mathcal{M}$, there exists an open neighbourhood $\Omega \subset \mathbb{R}^{n+d}$ such that $\mathcal{D}_p := \mathcal{M} \cap \Omega \subset \mathbb{R}^{n+d}$ is the image of a parametrisation from some open domain $D_x \subset \mathbb{R}^n$,

$$f : D_x \rightarrow f(D_x) = \mathcal{D}_p. \quad (\text{A.77})$$

♦

Remark A.13. If $f : D \rightarrow \mathcal{M} \cap \Omega$ and $\tilde{f} : \tilde{D} \rightarrow \mathcal{M} \cap \tilde{\Omega}$ are two parametrisations with $f(x_0) = \tilde{f}(\tilde{x}_0) = p \in \mathcal{M} \cap \Omega \cap \tilde{\Omega}$ then one can show that $(f^{-1} \circ \tilde{f}) : \tilde{f}^{-1}(\Omega \cap \tilde{\Omega}) \rightarrow f^{-1}(\Omega \cap \tilde{\Omega})$ is a

diffeomorphism between open sets in $\mathbb{R}^n$. Thus, one can view parametrisations f as the inverses of coordinate charts x . ♦

Embedded submanifolds can alternatively be characterised via equality constraints as the following theorem from [165, see Theorem 2.2] describes:

Theorem A.5 (*Embedded submanifolds via equality constraints*). *Let Ω be an open subset of $\mathbb{R}^{n+d}$. Let $h : \Omega \rightarrow \mathbb{R}^d$ be a differentiable map and let $c_0 \in \mathbb{R}^d$ be defined such that the differential $Dh_p \in \mathbb{R}^{d \times (n+d)}$ has full rank d at every point $p \in \Omega$ with $h(p) = c_0$. Then, the preimage,*

$$h^{-1}(c_0) = \{p \in \Omega | h(p) = c_0\}, \quad (\text{A.78})$$

is an n -dimensional submanifold of $\mathbb{R}^{n+d}$. ♦

Note that this fact follows from the well known Inverse Function Theorem from classical multivariate calculus. A simple application of the above theorem is to apply it with the function $h : \mathbb{R}^3 \rightarrow \mathbb{R}$, given by $h(\mathbf{x}) = \|\mathbf{x}\|_2^2 - 1$, to show that the unit sphere $S^2 = h^{-1}(0)$ is a 2-dimensional embedded submanifold of $\mathbb{R}^3 = \mathbb{R}^{2+1}$. Here $\|\cdot\|_2$ is the standard Euclidean norm in $\mathbb{R}^3$. We will see a further use of this theorem in Section 5.4.3.

Tangent spaces

The tangent space of an embedded submanifold can also be characterised with respect to both intrinsic and extrinsic coordinates on the manifold as the following two theorems from [165, Theorems 2.4, 2.5] describe:

Theorem A.6 (*Intrinsic Characterisation of Tangent Space*). *Let $\mathcal{M} \subset \mathbb{R}^{n+d}$ be an n -dimensional submanifold of $\mathbb{R}^{n+d}$, $D \subset \mathbb{R}^n$ be an open domain and $f : D \rightarrow f(D) \subseteq \mathcal{M}$ be a parametrisation of $\mathcal{M}$. Then, for $x \in D$ with $p = f(x) \in \mathcal{M}$ one has that the tangent space at p is given by the image of the differential of f at x i.e.*

$$T_p \mathcal{M} = \text{im}(Df_x). \quad (\text{A.79})$$

♦

Theorem A.7 (*Extrinsic Characterisation of Tangent Space*). *Let $\Omega \subset \mathbb{R}^{n+d}$, $h : \Omega \rightarrow \mathbb{R}^d$ and $c_0 \in \mathbb{R}^d$ be as in Theorem A.5, and let $\mathcal{M} := h^{-1}(c_0) \subset \mathbb{R}^{n+d}$. Then, the tangent space at a point $p \in \mathcal{M}$ is given by the kernel/null space of the differential of h at p i.e.*

$$T_p \mathcal{M} = \ker(Dh_p). \quad (\text{A.80})$$

◆

Remark A.14. It is an immediate consequence of Theorems A.6 and A.7 that the dimension of the tangent space $T_p\mathcal{M}$ as a vector space is equal to the dimension of the manifold, i.e.

$$\dim T_p\mathcal{M} = \dim \mathcal{M} = n, \quad (\text{A.81})$$

for all $p \in \mathcal{M}$. ◆

Stiefel Manifold

As mentioned in Section 5.4 the Stiefel manifold is an embedded submanifold of $\mathbb{R}^{n \times d}$. This can be shown by applying Theorem A.5 with the function $h : \mathbb{R}^{d \times n} \rightarrow \text{Sym}(n)$ defined via $h(\mathbf{X}) = \mathbf{X}^T \mathbf{X} - \mathbf{I}_n$, where $\text{Sym}(n) = \{\mathbf{A} \in \mathbb{R}^{n \times n} | \mathbf{A}^T = \mathbf{A}\}$ is the linear space of symmetric $n \times n$ matrices. This theorem also justifies the dimension of $\text{St}(d, n)$ given in Eq. (5.67).

Appendix B

Proofs and Additional Details

In this chapter we include any proofs not included in the main thesis body together with any additional details required for the proofs of the thesis.

B.1 Additional Proofs for Chapter 2

In this section we include any proofs of results from Section 2.3 not included in the main body of the thesis.

Proof of Proposition 2.4.

This proof is adopted from the proof of Theorem 3.1 in [78]. The fact that the posterior is given by Eq. (2.75) can be established by first working out the distribution of a finite-dimensional approximation of the posterior via evaluating the posterior on a set of $m \in \mathbb{N}$ test locations $\mathcal{D}_m^* := \{\mathbf{x}_1^*, \dots, \mathbf{x}_m^*\} \subset D$. This can be obtained by first noting that based on Definition 2.12 the vector of function evaluations at the test and observation locations, $(f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_m^*), f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^T \in \mathbb{R}^{m+n}$ has the following multivariate normal distribution under the GP prior Eq. (2.73):

$$\mathcal{N}_{m+n} \left(\begin{pmatrix} \mathbf{m}^* \\ \mathbf{m} \end{pmatrix}, \begin{pmatrix} \mathbf{K}^{**} & \mathbf{K}^* \\ (\mathbf{K}^*)^T & \mathbf{K} \end{pmatrix} \right), \quad (\text{B.1})$$

where,

$$\mathbf{m} := (m(\mathbf{x}_1), \dots, m(\mathbf{x}_n))^T \in \mathbb{R}^n, \quad (\text{B.2})$$

$$\mathbf{m}^* := (m(\mathbf{x}_1^*), \dots, m(\mathbf{x}_m^*))^T \in \mathbb{R}^m, \quad (\text{B.3})$$

$$\mathbf{K} := \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n \in \mathbb{R}^{n \times n}, \quad (\text{B.4})$$

$$\mathbf{K}^* := \{k(\mathbf{x}_i^*, \mathbf{x}_j^*)\}_{i,j=1}^{i=m, j=n} \in \mathbb{R}^{m \times n}, \quad (\text{B.5})$$

$$\mathbf{K}^{**} := \{k(\mathbf{x}_i^*, \mathbf{x}_j^*)\}_{i,j=1}^m \in \mathbb{R}^{m \times m}. \quad (\text{B.6})$$

Since the observational noise, ϵ , is independent from the GP prior we have that the joint distribution of the observations, $\mathbf{y}$, with the function evaluations at the test locations has the following multivariate normal distribution:

$$\mathcal{N}_{m+n} \left(\begin{pmatrix} \mathbf{m}^* \\ \mathbf{m} \end{pmatrix}, \begin{pmatrix} \mathbf{K}^{**} & \mathbf{K}^* \\ (\mathbf{K}^*)^T & \mathbf{K} + \sigma^2 \mathbf{I}_n \end{pmatrix} \right). \quad (\text{B.7})$$

We can now utilize Proposition A.4 to deduce that the conditional distribution of $(f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_m^*))$ given $\mathbf{y}$ is:

$$\mathcal{N}_m(\tilde{\mathbf{m}}, \tilde{\Sigma}), \quad (\text{B.8})$$

where the posterior mean vector and covariance matrix are given by:

$$\tilde{\mathbf{m}} := \mathbf{m}^* + \mathbf{K}^* (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} (\mathbf{y} - \mathbf{m}), \quad (\text{B.9})$$

$$\tilde{\Sigma} := \mathbf{K}^{**} - \mathbf{K}^* (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} (\mathbf{K}^*)^T. \quad (\text{B.10})$$

Note that since $\sigma > 0$ the matrix $\mathbf{K} + \sigma^2 \mathbf{I}_n$ is non-singular so we can replace the pseudo-inverse in Proposition A.4 with the regular inverse.

Since Eq. (B.8) holds for any arbitrary number of test locations it follows from the Kolmogorov Extension Theorem [49, Theorem 12.1.2] together with Definition 2.12 that the full posterior distribution of $f|\mathbf{y}$ is indeed given by Eq. (2.75) as required. $\blacksquare$

B.2 Additional Proof Details for Chapter 3

In this section we include additional details for the proofs of the main theoretical results of Chapter 3 not included in the main thesis body.

We will require the following additional notation, adopted from [39], for the proofs below. Let Θ be a parameter set and consider mappings $A, B : \Theta \rightarrow \mathbb{R}$. We will write $A(\theta) \lesssim B(\theta)$ to denote the relation that there exists some constant $C > 0$, independent of $\theta \in \Theta$, such that $A(\theta) \leq CB(\theta)$

for all $\theta \in \Theta$. Given another parameter set Φ and mappings $A, B : \Theta \times \Phi \rightarrow \mathbb{R}$ we will denote by $A(\theta, \varphi) \lesssim_{\varphi} B(\theta, \varphi)$ the relation that for all $\varphi \in \Phi$ there exists a constant $C_{\varphi} > 0$, independent of $\theta \in \Theta$, such that $A(\theta, \varphi) \leq C_{\varphi} B(\theta, \varphi)$ for all $\theta \in \Theta$ and $\varphi \in \Phi$. Finally, we will use the notation $A(\theta) \approx B(\theta)$ when both $A(\theta) \lesssim B(\theta)$ and $B(\theta) \lesssim A(\theta)$ hold. The notation $A(\theta, \varphi) \approx_{\varphi} B(\theta, \varphi)$ is similarly defined.

Additional details for proof of Theorem 3.3.

We rewrite the FEM solution operator $\hat{L}_h^{-1}$ following the notation in [39] as $\hat{L}_h^{-1} := L_h^{-1} \Pi_h$, where $L_h : V_h \rightarrow V_h$ is the Galerkin discretisation of the differential operator $\mathcal{L}$ with respect to V_h (see Definition A.12), and $\Pi_h : H^{-1}(D) \rightarrow V_h$ is the unique extension to $H^{-1}(D)$ of the $L^2(D)$ -orthogonal projection onto V_h (see Eq. (A.20)).

We start by verifying the statement given in the proof of Theorem 3.3 in the main thesis body that our setting and assumptions satisfy the assumptions required for the use of part of Proposition 4 in [39].

- The form of the differential operator $\mathcal{L}$, together with Assumption 2.4 satisfy Assumptions 1.I-II from [39].
- Assumption 2.1 implies Assumption 2.I from [39] (see comment following Assumption 2 in [39]).
- Assumption 2.5 guarantees Assumptions 3.II-III from [39] (see comment after Assumption 4 in [39]).
- That Π_h is $H^1(D)$ -stable (see Definition A.13) follows from Assumption 2.5 (see comment in proof of Theorem 3 from [39]).
- We have that Assumption 3.I from [39] holds with $\theta_0 = 1/2, \theta_1 = 2$. This follows from Assumptions 2.1, 2.4, and 2.5 together with the choice of FE basis functions (using Lemma 11 from [39]).
- Finally, Assumptions 2.1 and 2.4 show that $\mathcal{L}^{-1}$ is continuous as a mapping from $L^2(D)$ to $H^2(D) \cap H_0^1(D)$. This means that $\mathcal{L}$ is $H^2(D)$ -regular, see Definition A.11 (i.e. with $\alpha = 1$). Thus, we can take α as in (62) from [39] to be $\alpha = 1$ (this follows from using Lemma 10 from [39]).

For completeness, we will now provide additional, explicit, details of how to obtain the $\mathcal{O}(h^{2-d/2-\epsilon'})$ bound for the covariance contribution here. These steps closely follow the steps taken in the proof of Proposition 4 from [39].

Let $E := \mathcal{L}^{-1} - \hat{L}_h^{-1} = \mathcal{L}^{-1} - L_h^{-1}\Pi_h$ be the difference of solution operators. Thus, we need to bound $\|E\|_{HS}$. To this end, let $\eta \in (0, 1)$ and rewrite E as follows (as is done in the proof of [39, Proposition 4]):

$$\begin{aligned} E &= \mathcal{L}^{-1} - L_h^{-1}\Pi_h \\ &= \left(\mathcal{L}^{-(1-\eta)} - L_h^{-(1-\eta)}\Pi_h \right) L_h^{-\eta}\Pi_h + \mathcal{L}^{-(1-\eta)} (\mathcal{L}^{-\eta} - L_h^{-\eta}\Pi_h). \end{aligned}$$

In rewriting E as above we have used the property that $\Pi_h L_h^{-\beta} = L_h^{-\beta}$ for all $\beta \geq 0$ which holds since $L_h^{-\beta}$ maps into V_h .

Thus, we can bound $\|E\|_{HS}$ as:

$$\begin{aligned} \|E\|_{HS} &\leq \left\| \left(\mathcal{L}^{-(1-\eta)} - L_h^{-(1-\eta)}\Pi_h \right) L_h^{-\eta}\Pi_h \right\|_{HS} + \left\| \mathcal{L}^{-(1-\eta)} (\mathcal{L}^{-\eta} - L_h^{-\eta}\Pi_h) \right\|_{HS} \\ &:= E_1 + E_2. \end{aligned}$$

We bound each of E_1, E_2 separately, starting with E_1 . To this end, let $\epsilon_0 > 0$ be chosen to be sufficiently small so that $2 - d/2 > 4\epsilon_0$.

$$\begin{aligned} E_1 &= \left\| \left(\mathcal{L}^{-(1-\eta)} - L_h^{-(1-\eta)}\Pi_h \right) L_h^{-\eta}\Pi_h \right\|_{HS} \\ &\leq \left\| \mathcal{L}^{-(1-\eta)} - L_h^{-(1-\eta)}\Pi_h \right\|_{L^2} \|L_h^{-\eta}\Pi_h\|_{HS}. \end{aligned}$$

We start with bounding $\|L_h^{-\eta}\Pi_h\|_{HS}$. Now, Assumptions 2.1 and 2.4, together with the fact that we are using a conforming FE method (see Remark 2.7) allow us to use (81) from [39] to obtain:

$$\|L_h^{-\eta}\Pi_h\|_{HS} \lesssim_{(\eta, \kappa, D)} \max\{h^{2\eta-d/2}, 1\}, \quad (\text{B.11})$$

provided that $\eta \neq d/4$. **Note:** since we are interested in obtaining bounds for sufficiently small h it would be ideal if $2\eta - d/2 > 0 \implies \eta > d/4$. Since $\eta \neq d/4$ is required we can set $\eta = d/4 + \epsilon_0$, where ϵ_0 is as above. (*Note: since $4\epsilon_0 < 2 - d/2$ we have $d < 4\eta < 2 + d/2$. This, together with $d \in \{1, 2\}$, implies that $\eta \in (1/4, 3/4) \subseteq (0, 1)$ as required.*)

Remark B.1. In order to use (81) from [39] we need to check that all the assumptions of Remark 6, Lemma 1 and Assumption 3.II from that paper are satisfied. The use of a conforming FE method deals with Remark 6, Assumptions 2.1 and 2.4 deal with Assumptions 1.I-II, and Assumption 2.5 implies Assumption 3.III holds (see comments after Assumption 4 from [39]). ♦

We now bound the term $\|\mathcal{L}^{-(1-\eta)} - L_h^{-(1-\eta)}\Pi_h\|_{L^2}$. To bound this term we will use Theorem 1 from [39]. In order to use this theorem we must check various assumptions and requirements from

that paper are satisfied. For completeness and clarity we do that here:

- Assumptions 1.I-II are covered by the form of the differential operator $\mathcal{L}$ together with Assumption 2.4.
- Assumption 2.I is covered by Assumption 2.1 (comments after Assumption 2 in [39]).
- Assumption 3.II is covered by Assumption 2.5 (comments after Assumption 4 in [39]).
- Π_h is $H^1(D)$ -stable follows from Assumption 2.5 (comment in proof of Theorem 3 in [39]).
- That Assumption 3.I holds with $\theta_0 = 1/2, \theta_1 = 2$ follows from Assumptions 2.1, 2.4, and 2.5 together with the choice of FE basis functions (using Lemma 11 from [39]).
- Finally, Assumptions 2.1 and 2.4 gives us $\mathcal{L}^{-1}$ is continuous as a mapping from $L^2(D)$ to $H^2(D) \cap H_0^1(D)$. This means that $\mathcal{L}$ is $H^2(D)$ -regular as in Definition 4 from [39] (i.e. with $\alpha = 1$). Thus, we can take α as in (62) from [39] to be $\alpha = 1$ (this follows using Lemma 10 from [39]).

We thus obtain:

$$\begin{aligned} \|\mathcal{L}^{-(1-\eta)} - L_h^{-(1-\eta)} \Pi_h\|_{L^2} &= \|\mathcal{L}^{-(1-d/4-\epsilon_0)} - L_h^{-(1-d/4-\epsilon_0)} \Pi_h\|_{L^2} \\ &= \|\mathcal{L}^{-\beta'} - L_h^{-\beta'} \Pi_h\|_{L^2}, \end{aligned}$$

where $\beta' := 1 - d/4 - \epsilon_0$. We can now use Theorem 1 from [39] with parameters $\beta = \beta', \sigma = \sigma' := 0, \delta = \delta' := 0$ (*note: since $\beta' = 1 - \eta$ and $\eta \in (0, 1)$ we have $\beta' > 0$ and so all the inequalities on the parameters are satisfied*) to obtain:

$$\|\mathcal{L}^{-\beta'} - L_h^{-\beta'} \Pi_h\|_{L^2} \lesssim_{(\epsilon', \epsilon_0, \kappa, D)} h^{\min\{2\beta' - \epsilon', 2\}}, \quad (\text{B.12})$$

for arbitrary $\epsilon' > 0$ and all $h > 0$ sufficiently small.

Thus, for sufficiently small $h > 0$, we can combine Eqs. (B.11) and (B.12) to obtain:

$$E_1 \lesssim_{(\epsilon', \epsilon_0, \kappa, D)} h^{\min\{2-d/2-2\epsilon_0-\epsilon', 2\}}, \quad (\text{B.13})$$

for arbitrary $\epsilon' > 0$.

We now move on to bounding E_2 . After rewriting E_2 (as is done in the proof of Proposition 4 from [39]) we have:

$$\begin{aligned} E_2 &= \left\| \mathcal{L}^{-(1-\eta)} (\mathcal{L}^{-\eta} - L_h^{-\eta} \Pi_h) \right\|_{HS} \\ &\leq \left\| (\mathcal{L}^{-\eta} - L_h^{-\eta} \Pi_h) \mathcal{L}^{-(1-2\eta)} \right\|_{L^2} \|\mathcal{L}^{-\eta}\|_{HS}. \end{aligned}$$

Note that, due to the spectral properties of $\mathcal{L}$, we have that $\|\mathcal{L}^{-m}\|_{HS} < \infty$ for any $m > d/4$. Since $\eta = d/4 + \epsilon_0 > d/4$ we thus have that $\|\mathcal{L}^{-\eta}\|_{HS} < \infty$ (*and is independent of h*).

We can now again use Theorem 1 from [39] with parameters $\beta = \beta'' := \eta$, $\sigma = \sigma'' := 0$, $\delta = \delta'' := \min\{2 - 4\eta, 1 + \alpha\}$ in order to bound the remaining term. We must check that these parameters satisfy the requirements of Theorem 1. We check only the non-trivial ones, i.e. that $\delta'' \geq -1$, and $2\beta'' + \delta'' > 0$.

To start, note that $d < 4\eta < 2 + d/2 \implies -d/2 < 2 - 4\eta < 2 - d$. Since $d \in \{1, 2\}$ we thus have $2 - 4\eta < 2 - d \leq 1 < 1 + \alpha = 2$ and so $\delta'' = 2 - 4\eta$. This immediately shows that $\delta'' \geq -1$. We now move on and compute:

$$\begin{aligned} 2\beta'' + \delta'' &= 2\eta + 2 - 4\eta \\ &= 2 - 2\eta > 0, \end{aligned}$$

since $\eta \in (0, 1)$. The parameters thus satisfy all the requirements and so we have, for arbitrary $\epsilon'' > 0$ that:

$$E_2 \lesssim_{(\epsilon'', \epsilon_0, \kappa, D)} h^{\min\{2\beta'' + \delta'' - \epsilon'', 2, 2 + \delta''\}}. \quad (\text{B.14})$$

Now note that since $\delta'' = 2 - 4\eta$ we can compute:

$$\begin{aligned} 1 + \alpha + \delta'' &= 4 - 4\eta \\ &= 4 - d - 4\epsilon_0 \\ &> 4 - d + \frac{d}{2} - 2 \\ &= 2 - \frac{d}{2} \\ &> 2 - \frac{d}{2} - 4\epsilon_0 \\ &= \min\{2 - d/2 - 4\epsilon_0, 1 + \alpha\}. \end{aligned} \quad (\text{B.15})$$

Recalling the definitions of β'', δ'', η we have:

$$E_2 \lesssim_{(\epsilon'', \epsilon_0, \kappa, D)} h^{\min\{2 - d/2 - 2\epsilon_0 - \epsilon'', 2, 2 + \delta''\}}. \quad (\text{B.16})$$

Recalling Eq. (B.13) and using Eq. (B.15) we can conclude that:

$$\|E\|_{HS} \lesssim_{(\epsilon, \alpha, \kappa, D)} h^{\min\{2 - d/2 - \epsilon, 2\}} = h^{2 - d/2 - \epsilon}, \quad (\text{B.17})$$

for sufficiently small $h > 0$ and any $\epsilon > 0$ (by adjusting $\epsilon_0, \epsilon', \epsilon'' > 0$).

The mean contribution to the Wasserstein-2 distance will still be bounded by $\mathcal{O}(h^2)$ and so

we conclude that the bound on the Wasserstein-2 distance is of order $\mathcal{O}(h^{2-d/2-\epsilon})$ for sufficiently small $h > 0$ and any $\epsilon > 0$. Relabelling the dummy variable ϵ with ϵ' gives us what we require. $\blacksquare$

Additional details for Theorem 3.2. We now provide additional details regarding Theorem 3.2. To start we will show that in the case of $\lambda \in [0, 1]$ solutions drawn from the true prior lie almost surely in the fractional Sobolev Space $H^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ for all $\epsilon' > 0$. We will require the following supporting lemma:

Lemma B.1. If $f \sim \mathcal{N}(0, K)$, where K satisfies Assumption 3.1, then it holds that $f \in \dot{H}_{\mathcal{L}}^{-(1-\lambda)\frac{d}{2}-\epsilon'}$, $\mathbb{P}$ -a.s., for all $\epsilon' > 0$. (Note: we are assuming that $\lambda \in [0, 1]$ and so K is also assumed to share the same eigenfunctions as $\mathcal{L}$ as described in Assumption 3.1.) $\blacklozenge$

Proof. The proof of this is similar to the proof of Proposition 2.3 in [15]. Under Assumption 3.1 we can use the Karhunen-Lo  ve theorem (see Theorem A.4) to formally represent f as follows:

$$f = \sum_{j \in \mathbb{N}} \zeta_j e_j, \quad (\text{B.18})$$

where the $\{\zeta_j\}_{j \in \mathbb{N}}$ are mean-zero, uncorrelated random variables with covariance structure given by $\mathbb{E}[\zeta_i \zeta_j] = \mu_j \delta_{ij}$.

Remark B.2. Note that this implicitly uses the fact that the eigenfunctions of K are the same as those of $\mathcal{L}$ and that $Ke_j = \mu_j e_j$. $\blacklozenge$

We now introduce, for any $N \in \mathbb{N}$, the truncated random forcing:

$$f_N := \sum_{j=1}^N \zeta_j e_j. \quad (\text{B.19})$$

We can now compute, for any $s \geq 0$:

$$\begin{aligned} \mathbb{E} [\|f_N\|_{-s, \mathcal{L}}^2] &= \mathbb{E} \left[\sum_{j \in \mathbb{N}} \lambda_j^{-s} \langle f_N, e_j \rangle_{L^2(D)}^2 \right] \\ &= \mathbb{E} \left[\sum_{j=1}^N \lambda_j^{-s} \zeta_j^2 \right] \\ &= \sum_{j=1}^N \lambda_j^{-s} \mu_j. \end{aligned}$$

Now, using the spectral properties of the eigenvalues of $\mathcal{L}$, (see Weyl's Law Lemma A.2) together

with Assumption 3.1 we have:

$$\begin{aligned}\mathbb{E} [\|f_N\|_{-s, \mathcal{L}}^2] &= \sum_{j=1}^N \lambda_j^{-s} \mu_j \\ &\lesssim \tilde{C} \sum_{j=1}^N j^{-\lambda - \frac{2s}{d}} \\ &\leq \tilde{C} \left(1 + \int_1^N x^{-\lambda - \frac{2s}{d}} dx \right).\end{aligned}$$

If we now let $s = \frac{d}{2}(1 - \lambda) + \epsilon'$, for $\epsilon' > 0$, we can compute:

$$\begin{aligned}\mathbb{E} [\|f_N\|_{-s, \mathcal{L}}^2] &\lesssim \left(1 + \int_1^N x^{-\lambda - \frac{2s}{d}} dx \right) \\ &= \left(1 + \int_1^N x^{-1 - \frac{2\epsilon'}{d}} dx \right) \\ &= \left(1 + \frac{d}{2\epsilon'} - \frac{d}{2\epsilon'} N^{-\frac{2\epsilon'}{d}} \right).\end{aligned}$$

By taking $N \rightarrow \infty$, we can conclude that, for all $\epsilon' > 0$ we have $f \in L^2 \left(\Omega; \dot{H}_{\mathcal{L}}^{-(1-\lambda)\frac{d}{2}-\epsilon'} \right)$ and so $f \in \dot{H}_{\mathcal{L}}^{-(1-\lambda)\frac{d}{2}-\epsilon'}$, $\mathbb{P}$ -a.s., as required. $\blacksquare$

Using this lemma gives us that draws from the true prior, which can be expressed as $\mathcal{L}^{-1}f$, lie almost surely in $\dot{H}_{\mathcal{L}}^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ for all $\epsilon' > 0$. We now note that under Assumptions 2.1 and 2.4 we have that the Hilbert space $\dot{H}_{\mathcal{L}}^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ is actually a subset of the fractional Sobolev space $H^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ as discussed in Appendix A.2. Thus, solutions drawn from the true prior lie almost surely in $H^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ as required.

We now outline how the Embedding Theorem for fractional Sobolev spaces (see Theorem 2.1) together with Corollary 2.1 can be used to argue that $H^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ is a subset of $C(D)$ for an appropriate choice of ϵ' in the three cases:

- $d = 1$ and $\lambda \in [0, 1]$,
- $d = 2$ and $\lambda \in (0, 1]$,
- $d = 3$ and $\lambda \in (2/3, 1]$.

Corollary 2.1 gives us that $H^{2-(1-\lambda)\frac{d}{2}-\epsilon'}$ is a subset of $C(D)$ provided $2s > d$ where $s := 2 - (1 -$

$\lambda)^{\frac{d}{2}} - \epsilon'$. This is equivalent to:

$$2s > d$$

$$4 - (1 - \lambda)d - 2\epsilon' > d$$

$$2\epsilon' < 4 - (1 - \lambda)d - d$$

$$2\epsilon' < 4 - (2 - \lambda)d.$$

We now consider the cases $d = 1, 2, 3$ separately. In the first case of $d = 1$ note that for any $\lambda \in [0, 1]$ it is possible to choose $\epsilon' > 0$ such that $2\epsilon' < 2 + \lambda$ and so no further restriction on λ is required.

In the second case of $d = 2$ we will show that it is necessary to restrict λ to the interval $(0, 1]$. To see this note that:

$$2s > d \stackrel{d=2}{\iff} 2\epsilon' < 2\lambda.$$

Thus, in order to be able to choose $\epsilon' > 0$ such that $2\epsilon' < 2\lambda$ we will require that $2\lambda > 0$ and so we must exclude $\lambda = 0$.

In the third case of $d = 3$ we will show that it is necessary to restrict λ to $(2/3, 1]$. To see this note that:

$$2s > d \stackrel{d=3}{\iff} 2\epsilon' < 3\lambda - 2.$$

Thus, in order to be able to choose $\epsilon' > 0$ such that $2\epsilon' < 3\lambda - 2$ we will require that $3\lambda - 2 > 0$ i.e. $\lambda > 2/3$. Thus, we must restrict λ to the interval $(2/3, 1]$ in order to insure that draws from the true prior lie almost-surely in $C(D)$. ■

B.3 Additional Proofs for Chapter 5

In this section we include any proofs of results from Chapter 5 not included in the main body of the thesis.

Proof of Proposition 5.1.

This follows from Definitions 2.7 and 5.1. We can compute:

$$\text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{AXB} + \mathbf{C}) \tag{B.20}$$

$$= \text{vec}(\mathbf{AXB}) + \text{vec}(\mathbf{C}) \tag{B.21}$$

$$= (\mathbf{B}^T \otimes \mathbf{A}) \text{vec}(\mathbf{X}) + \text{vec}(\mathbf{C}), \tag{B.22}$$

where we have used standard properties of the vectorisation operator, vec , and the Kronecker

product, $\otimes$. Since $\mathbf{X} \sim \mathcal{MN}_{d,n}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ we have via Definition 5.1 that

$$\text{vec}(\mathbf{X}) \sim \mathcal{N}_{dn}(\text{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U}). \quad (\text{B.23})$$

Thus, using Proposition A.1 together with the above computation yields:

$$\text{vec}(\mathbf{Y}) \sim \mathcal{N}_{mk}((\mathbf{B}^T \otimes \mathbf{A}) \text{vec}(\mathbf{M}) + \text{vec}(\mathbf{C}), (\mathbf{B}^T \otimes \mathbf{A})(\mathbf{V} \otimes \mathbf{U})(\mathbf{B}^T \otimes \mathbf{A})^T) \quad (\text{B.24})$$

$$\equiv \mathcal{N}_{mk}(\text{vec}(\mathbf{AMB} + \mathbf{C}), (\mathbf{B}^T \mathbf{VB}) \otimes (\mathbf{AUA}^T)), \quad (\text{B.25})$$

where we have simplified the mean and covariance terms using standard properties of the vectorisation operator and Kroenecker product. Thus, $\mathbf{Y} \sim \mathcal{MN}_{m,k}(\mathbf{AMB} + \mathbf{C}, \mathbf{AUA}^T, \mathbf{B}^T \mathbf{VB})$ as required. ■

Appendix C

Additional Implementation Details for Experiments

In this chapter any additional implementation details for the numerical experiments in this thesis which are not included in the main thesis body are presented.

C.1 Additional Details for Experiments in Chapter 4

Remark C.1. In both the StatFEM prior and posterior the $n_u \times n_u$ matrix $\mathbf{K}_{\mathcal{F}_h} = \mathcal{F}_h K \mathcal{F}_h^\dagger$ appears. This is the covariance matrix for the random load vector $\mathbf{f} = \mathcal{F}_h f$. The ij -th element of this covariance matrix is given by:

$$(\mathbf{K}_{\mathcal{F}_h})_{ij} = \int_D \int_D \varphi_i(\mathbf{x}) k_f(\mathbf{x}, \mathbf{y}) \varphi_j(\mathbf{y}) d\mathbf{x} d\mathbf{y}. \quad (\text{C.1})$$

For a covariance function k_f which is globally supported, the covariance matrix, $\mathbf{K}_{\mathcal{F}_h}$, is dense as can be seen from expression Eq. (C.1) above. As such, it can be very difficult to efficiently compute and assemble the entries of this matrix. We thus employ an approximation of this matrix, by interpolating the covariance function k_f with the finite element basis functions, as suggested by [65], i.e. denoting the non-boundary nodes of the FE triangulation by $\{\mathbf{x}_i\}_{i=1}^{n_u}$, we approximate k_f by,

$$k_f(\mathbf{x}, \mathbf{y}) \approx \sum_{k,l=1}^{n_u} \varphi_k(\mathbf{x}) k_f(\mathbf{x}_k, \mathbf{x}_l) \varphi_l(\mathbf{y}). \quad (\text{C.2})$$

Using this interpolation of k_f we can approximate the entries of $\mathbf{K}_{\mathcal{F}_h}$ by,

$$\begin{aligned}
 (\mathbf{K}_{\mathcal{F}_h})_{ij} &\approx \sum_{k,l=1}^{n_u} k_f(\mathbf{x}_k, \mathbf{x}_l) \int_D \int_D \varphi_i(\mathbf{x}) \varphi_k(\mathbf{x}) \varphi_j(\mathbf{y}) \varphi_l(\mathbf{y}) d\mathbf{x} d\mathbf{y} \\
 &= \sum_{k,l=1}^{n_u} k_f(\mathbf{x}_k, \mathbf{x}_l) \int_D \varphi_i(\mathbf{x}) \varphi_k(\mathbf{x}) d\mathbf{x} \int_D \varphi_j(\mathbf{y}) \varphi_l(\mathbf{y}) d\mathbf{y} \\
 &= \sum_{k,l=1}^{n_u} k_f(\mathbf{x}_k, \mathbf{x}_l) M_{ik} M_{jl} \\
 &= (\mathbf{M} \mathbf{C}_f \mathbf{M}^T)_{ij},
 \end{aligned}$$

i.e. we can approximate:

$$\mathbf{K}_{\mathcal{F}_h} \approx \mathbf{M} \mathbf{C}_f \mathbf{M}^T, \quad (\text{C.3})$$

where $\mathbf{C}_f \in \mathbb{R}^{n_u \times n_u}$ is the covariance matrix corresponding to evaluating k_f on the non-boundary nodes, $\{\mathbf{x}_i\}_{i=1}^{n_u}$, of the FEM triangulation, and $\mathbf{M}$ is the mass matrix introduced in Section 2.2.2. This approximation will improve as h gets smaller. $\blacklozenge$

Remark C.2. Since all of the true and StatFEM priors and posteriors are Gaussian measures, and since we are interested in demonstrating a bound on the error between the respective priors and posteriors for the Wasserstein-2 distance the expression Eq. (2.86) will be useful. Focussing on the covariance contribution in Eq. (2.86), note that this is difficult to compute explicitly for the measures in this thesis. For two GPs $\nu_i \sim \mathcal{N}(0, \Sigma_i)$, $i \in \{1, 2\}$ we approximate the covariance contribution by fixing a fine reference grid $\{\tilde{\mathbf{x}}_i\}_{i=1}^N$, computing the covariance matrices, $\mathbf{C}_i = \{k_i(\tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_l)\}_{k,l=1}^N$, corresponding to the covariance operators, Σ_i , evaluated on this reference grid, where k_i is the covariance function associated with Σ_i respectively, and finally approximating the Wasserstein distance as follows:

$$\begin{aligned}
 W^2(\mathcal{N}(0, \Sigma_1), \mathcal{N}(0, \Sigma_2)) &= \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2 \text{tr} \sqrt{\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}} \\
 &\approx \text{tr}(\mathbf{C}_1) + \text{tr}(\mathbf{C}_2) - 2 \text{tr} \sqrt{\mathbf{C}_1^{1/2} \mathbf{C}_2 \mathbf{C}_1^{1/2}}.
 \end{aligned} \quad (\text{C.4})$$

$\blacklozenge$

Remark C.3. Let $m_\star, m_h$ denote the means of the true and StatFEM prior (or posterior) respectively. The mean contribution to the Wasserstein-2 distance is $\|m_\star - m_h\|_{L^2(D)}$. This contribution can be easily approximated by utilising a modern FEM package such as FEniCS [5, 100], which we use in this thesis. $\blacklozenge$