Statistical Computation with Kernels

by

François-Xavier Briol

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

University of Warwick,
Department of Statistics

September 2018

Contents

List of Figures

Acknowledgments

I am first and foremost grateful to my advisor, Mark Girolami, for giving me opportunities which are usually reserved to researchers in later stages of their careers. He introduced me to exciting areas of research across the fields of statistics, applied mathematics and machine learning, and to researchers doing interesting work in these areas. Mark also gave me the opportunity to attend conferences which greatly enriched my PhD experience and broadened my overview of statistics research. Finally, he has always been a great mentor, and regularly taken the time to make sure my PhD was running smoothly.

Of course, most of my work would not have been possible without the support and guidance of Chris Oates, who I consider as a great mentor and unofficial second advisor. Chris has been very patient in answering many of my mathematical questions, and taught me how to structure my research in an effective way.

I would also like to express my gratitude to many of the faculty members, students and visitors at the various institutions I have attended during my PhD, including the University of Oxford, University of Warwick, Imperial College London and The Alan Turing Institute. I am in particular grateful to all of the lecturers and students in the first year of the Oxford-Warwick Statistics programme which I have greatly enjoyed. I am also thankful to Michael Osborne and Dino Sejdinovic who have initiated me into the world of machine learning, Andrew Duncan and Alessandro Barp who have greatly widened my understanding of mathematics, and Jon Cockayne, Louis Ellam and Thibaut Lienart who have been of great help in answering many of my questions about statistics, machine learning and programming.

I am also grateful to the Society for Industrial and Applied Mathematics

(SIAM), the American Statistical Association (ASA), the Statistical and Applied Mathematical Sciences Institute (SAMSI) and the International Society for Bayesian Analysis (ISBA) for several travel awards which funded some of my travels to international conferences. I have also been fortunate to have the opportunity to visit several universities across the world, including a week-long visit at the University of Technology Sydney hosted by Chris Oates and Matt Wand, a month-long visit at the California Institute of Technology hosted by Andrew Stuart and Houman Owhadi, and a three week visit at the Isaac Newton Institute at the University of Cambridge, where I was hosted by Catherine Powell. To all my hosts: thank you.

Finally, I would also like to thank Wilfrid Kendall and Arthur Gretton for carefully examining this thesis, for their many suggestions, and an enjoyable viva.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below. Parts of this thesis have been published by the author:

1. F-X. Briol, C. J. Oates, M. Girolami, and M. A. Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances In Neural Information Processing Systems 28*, pages 1162–1170, 2015a

   ▷ The article was mostly written by Chris Oates and myself, and we both contributed equally to the methodology and theory sections. The numerical experiments were done by myself, with some code from a previous publication contributed by Chris Oates. Mark Girolami and Michael Osborne provided helpful suggestions to improve the manuscript. The paper was awarded a "spotlight presentation" at NIPS 2015, which was only awarded to the top 4.5% of submitted papers.

2. F-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *To appear in "Statistical Science" with discussion and rejoinder, arXiv:1512.00933*, 2015b

   ▷ The article was mostly written by Chris Oates and myself, and we both contributed equally to the methodology and theory sections. The numerical experiments were done by myself, with some code from previous publications also contributed by Chris Oates, Shiwei Lan and Ricardo Marques.

The paper was awarded a "Best Student Paper" award in 2016 by the section on Bayesian Statistical Science of the American Statistical Association, and was reviewed in a series of blog posts by eminent statisticians including Andrew Gelman: `http://andrewgelman.com/2015/12/07/28279/` and Christian Robert `https://xianblog.wordpress.com/2015/12/17/`. It has been accepted to "Statistical Science", and will appear with discussions from leading statisticians (including Michael L. Stein, Ying Hung, Art Owen, Fred Hickernell and R. Jagadeeswaran) and a rejoinder from the authors [Briol et al., 2018].

3. F-X. Briol, J. Cockayne, and O. Teymur. Comments on "Bayesian solution uncertainty quantification for differential equations" by Chkrebtii, Campbell, Calderhead & Girolami. *Bayesian Analysis*, 11(4):1285–1293, 2016

   ▷ This paper is a contributed discussion of Chkrebtii et al. [2016], and is the outcome of a series of discussions between Jon Cockayne, Onur Teymur and myself. The paper was mostly written by myself.

4. F-X. Briol, C. J. Oates, J. Cockayne, W. Y. Chen, and M. Girolami. On the sampling problem for kernel quadrature. In *Proceedings of the International Conference on Machine Learning*, pages 586–595, 2017

   ▷ This article is the result of collaborative work between Chris Oates and myself. The numerical experiments were all performed by myself.

5. C. J. Oates, S. Niederer, A. Lee, F-X. Briol, and M. Girolami. Probabilistic models for integration error in the assessment of functional cardiac models. *Advances in Neural Information Processing*, 2017d

   ▷ This articles originates from discussions between Chris Oates and myself. The articles and numerical experiments are Chris Oates' work. Angela Lee and Steven Niederer provided the application for the paper.

6. F.-X. Briol and M. Girolami. Bayesian numerical methods as a case study

for statistical data science. In *Statistical Data Science*, pages 99–110. World Scientific, 2018

▷ This article is a chapter in the book “Statistical Data Science” edited by Niall Adams and Edward Cohen, and was written independently by myself.

7. A. Barp, F.-X. Briol, A. D. Kennedy, and M. Girolami. Geometry and dynamics for Markov chain Monte Carlo. *Annual Reviews in Statistics and Its Applications*, 5, 2018

   ▷ The article was written by Alessandro Barp and myself, with suggestions and revisions by Anthony Kennedy and Mark Girolami.

8. C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on Stein's identity. *Bernoulli*, 2018

   ▷ The article was mostly written by Chris Oates, who also developed most of the theory. Jon Cockayne contributed numerical simulations, and I helped generalise the proofs.

9. X. Xi, F-X. Briol, and M. Girolami. Bayesian quadrature for multiple related integrals. *International Conference on Machine Learning, PMLR 80:5369-5378*, 2018

   ▷ This article emanates from Xiaoyue Xi's thesis project for the MSc in Statistics at Imperial College London, which was supervised by myself. The project, methodology and theory were by all done by myself. Xiaoyue Xi was in charge of numerical simulations. The article was accepted for a “long talk” at ICML 2018, which was awarded to the top 8% of submitted papers.

10. W. Y. Chen, L. Mackey, J. Gorham, F-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the International Conference on Machine Learning, PMLR 80:843-852*, 2018

    ▷ The idea was originally proposed by myself. Chris Oates wrote most of the paper, Wilson Chen performed the numerical experiments and Lester Mackey and Jackson Gorham developed the theory.

11. F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Rejoinder for “Probabilistic integration: a role in statistical computation?”. *Statistical Science (to appear), arXiv:1811.10275*, 2018

    ▷ The article was written myself, with comments and feedback from all other authors. It will appear as the rejoinder for Briol et al. [2015b], which will appear with discussions at Statistical Science.

Chapter 5 is also part of ongoing work with Andrew Duncan, Alessandro Barp, Lester Mackey & Chris Oates.

Abstract

Modern statistical inference has seen a tremendous increase in the size and complexity of models and datasets. As such, it has become reliant on advanced computational tools for implementation. A first canonical problem in this area is the numerical approximation of integrals of complex and expensive functions. Numerical integration is required for a variety of tasks, including prediction, model comparison and model choice. A second canonical problem is that of statistical inference for models with intractable likelihoods. These include models with intractable normalisation constants, or models which are so complex that their likelihood cannot be evaluated, but from which data can be generated. Examples include large graphical models, as well as many models in imaging or spatial statistics.

This thesis proposes to tackle these two problems using tools from the kernel methods and Bayesian non-parametrics literature. First, we analyse a well-known algorithm for numerical integration called Bayesian quadrature, and provide consistency and contraction rates. The algorithm is then assessed on a variety of statistical inference problems, and extended in several directions in order to reduce its computational requirements. We then demonstrate how the combination of reproducing kernels with Stein's method can lead to computational tools which can be used with unnormalised densities, including numerical integration and approximation of probability measures. We conclude by studying two minimum distance estimators derived from kernel-based statistical divergences which can be used for unnormalised and generative models.

In each instance, the tractability provided by reproducing kernels and their properties allows us to provide easily-implementable algorithms whose theoretical foundations can be studied in depth.

Abbreviations

BIS	Bayesian importance sampling
BMC, BMCMC	Bayesian Monte Carlo, Bayesian Markov chain Monte Carlo
BQ	Bayesian quadrature
BQMC	Bayesian quasi-Monte Carlo
FW, FWLS	Frank-Wolfe, Frank-Wolfe with line search
FWBQ	Frank-Wolfe Bayesian quadrature
FWLSBQ	Frank-Wolfe with line search Bayesian quadrature
GP	Gaussian process
IID	Identically and independently distributed
IS	Importance sampling
KL	Kullback-Leibler
KSD	Kernel Stein discrepancy
MC, MCMC	Monte Carlo, Markov chain Monte Carlo
MMD	Maximum mean discrepancy
QMC	Quasi-Monte Carlo
RBF	Radial basis function
RKHS	Reproducing kernel Hilbert space
SM	Score matching
SMC	Sequential Monte Carlo
SMC-BQ	Sequential Monte Carlo Bayesian quadrature
SMC-BQ-KL	Sequential Monte Carlo Bayesian quadrature with kernel learning
WCE	Worst-case error

Chapter 1

Challenges for Statistical Computation

> "Computations are an issue in statistics whenever processing a dataset becomes a difficulty, a liability, or even an impossibility."
>
> Green et al. [2015]

As illustrated by Green et al. [2015], computation has always been an issue for large-scale statistical inference. Recently, computational issues have been exacerbated by increases in computing resources and the availability of larger datasets, which has encouraged scientists to fit ever-more complex models. Keeping up with these changes is a constant challenge for researchers in computational statistics. In this thesis, we review some of the main problems in this area and contribute novel methodology to two of them: (i) the problem of numerical integration of complex and expensive functions, and (ii) the problem of statistical inference for models with intractable likelihoods.

1.1 Challenge I: Numerical Integration and Sampling

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space[1]. A major issue preventing the application of many complex statistical methodologies is the need to compute the Lebesgue integral of

[1] We assume the reader is familiar with notions of measure and probability theory. If this is not the case, see Appendix A.2 for a brief introduction.

some integrable functions $f : \mathcal{X} \to \mathbb{R}$:

$$\Pi[f] \;\; := \;\; \int_{\mathcal{X}} f(\mathbf{x})\Pi(\mathrm{d}\mathbf{x}), \tag{1.1}$$

where Π is some probability measure on $(\mathcal{X}, \mathcal{F})$ assumed to admit some probability density function π with respect to some underlying reference measure μ on the space $\mathcal{X}$. The space $\mathcal{X}$ is called the state space and is usually a subspace of $\mathbb{R}^d$ or some manifold embedded in $\mathbb{R}^d$ for some $d \in \mathbb{N}$ (where we adopt the convention that $\mathbb{N}$ does not include 0).

From the point of view of statistical computation, the main issue arises when these integrals cannot be evaluated in closed form and have to be estimated numerically. Historically, classical quadrature rules such as Gaussian quadratures have been used extensively [Naylor and Smith, 1982; Smith et al., 1985]. These are however only suitable for low-dimensional integrals with a smooth integrand. Nowadays, it is common to use Monte Carlo (MC) methods [Meyn and Tweedie, 1993; Liu, 2001; Robert and Casella, 2004] to approximate the integral by taking an average of function values at samples from Π (either identically and independently distributed (IID) or approximately IID).

In both of the cases above, we obtain an approximation of the form:

$$\hat{\Pi}[f] \;\; := \;\; \sum_{i=1}^{n} w_i f(\mathbf{x}_i), \tag{1.2}$$

called quadrature (or cubature) rule, based on point sets (also called samples) $\mathbf{x}_i \in \mathcal{X}$ and weights $w_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Under certain regularity conditions, this estimator converges to the solution of the integral as $n \to \infty$. For finite but large sample sizes n, the estimator reasonably approximates the truth. However, these estimators will have (potentially very) large errors whenever π is highly multimodal, the state-space $\mathcal{X}$ is high-dimensional or the integrand f is computationally expensive to evaluate. Adapting numerical integration methods to each of these scenarios is one of the main tasks in computational statistics. We now highlight several applications of numerical integration in statistics.

1.1.1 Applications in Bayesian Statistics

In Bayesian statistics [Robert, 1994; Gelman et al., 2013], once a model and a prior have been specified, all that remains to be done is to repeatedly apply Bayes' theorem until we obtain a distribution on the variables of interest conditioned on every other observable variable. Denote by $\mathbf{X}$ the matrix whose rows are data points $\{\mathbf{x}_i\}_{i=1}^n$

from some data space denoted $\mathcal{D}$ and by $\theta \in \Theta$ the parameters of a statistical model. For simplicity, we assume that $\mathcal{D}$ and Θ are both Euclidean spaces. The simplest formulation of Bayesian inference (assuming the existence of all densities) is the following equation[2]:

$$p(\theta|\mathbf{X}) = \frac{p_0(\theta)p(\mathbf{X}|\theta)}{p(\mathbf{X})}, \tag{1.3}$$

where $p(\mathbf{X}|\theta)$ denotes the likelihood, or statistical model, and describes the plausibility of the parameter taking value θ when $\mathbf{X}$ is observed. Furthermore, $p_0(\theta)$ is the prior density on the unknown model parameters θ and $p(\theta|\mathbf{X})$ denotes the posterior density (after having observed $\mathbf{X}$) on these same parameters. The quantity in the denominator, $p(\mathbf{X})$ is called the model evidence or marginal likelihood, and can be expressed as

$$p(\mathbf{X}) = \int_\Theta p(\mathbf{X}|\theta)p_0(\theta)\mathrm{d}\theta. \tag{1.4}$$

The model evidence is an example of an integral that almost always needs to be computed, in this particular case in order to be able to evaluate our posterior on parameters θ. This is not possible in all but special cases, in which case we call this Bayesian approach a conjugate analysis.

Integrals are also required when predicting new data values $\mathbf{x}' \in \mathcal{D}$. This can be done by computing the posterior predictive distribution

$$p(\mathbf{x}'|\mathbf{X}) = \int_\Theta p(\mathbf{x}'|\theta)p(\theta|\mathbf{X})\mathrm{d}\theta, \tag{1.5}$$

which allows us to propagate the uncertainty in our posterior through to predictions. Similar integrals are also required to do model selection with Bayes factors [Kass and Raftery, 1995] or for Bayesian model averaging [Hoeting et al., 1999].

Clearly, Bayesian inference would be restricted to very simple models without numerical integration. This explains why Bayesian methods only became widely popular across the sciences in the 1990s, at which point the statistics community had been introduced to Markov Chain Monte Carlo (MCMC) methods [Robert and Casella, 2011].

[2] There is a clear abuse of notation in this equation since p is used for different densities. However, this is common in practice for ease of exposition.

1.1.2 Applications in Frequentist Statistics

Challenging integrals are also ubiquitous in frequentist statistics, and are often required for maximum likelihood estimation. Suppose we have IID realisations $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{D}$ from a probability measure $\mathbb{P}_{\theta^*}$ from some parametric family of Borel probability measures $\mathcal{P}_\Theta(\mathcal{D}) = \{\mathbb{P}_\theta : \theta \in \Theta\}$ defined on $\mathcal{D}$. Once again assume $\mathcal{D}$ and Θ are Euclidean spaces and denote by $p(\cdot|\theta)$ the Lebesgue density of $\mathbb{P}_\theta$. We are interested in finding the "true" parameter $\theta^* \in \Theta$ which generated these samples, and the maximum likelihood approach proposes to do so by maximising the expected log-likelihood under the data-generating process:

$$\arg\max_{\theta \in \Theta} \int_{\mathcal{D}} \log p(\mathbf{x}|\theta) p(\mathbf{x}|\theta^*) \mathrm{d}\mathbf{x}. \tag{1.6}$$

In practice, the integral is usually approximated using a MC estimate with the samples $\{\mathbf{x}_i\}_{i=1}^n$ that are readily available, and we get the following optimisation problem:

$$\arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p(\mathbf{x}_i|\theta). \tag{1.7}$$

Numerical integration is therefore clearly fundamental here, and we may wish to use more efficient methods to approximate the integral. Note that the approach is of course only feasible if the likelihood can be evaluated in closed form. In the case of latent variable models, this does not necessarily hold. Indeed, assume we have a set of unobserved variables $\mathbf{y}$ (called nuisance parameters) in some space $\mathcal{Y}$. In this case, we would usually have access to a conditional likelihood $p(\mathbf{x}|\mathbf{y}, \theta)$ and therefore need to integrate out all possible values of the latent variable $\mathbf{y}$ to get a marginal likelihood:

$$p(\mathbf{x}|\theta) \quad = \quad \int_{\mathcal{Y}} p(\mathbf{x}|\mathbf{y}, \theta) p(\mathbf{y}|\theta) \mathrm{d}\mathbf{y}. \tag{1.8}$$

which we can then use for maximum likelihood estimation. This will be infeasible for most models and will once again require numerical integration; see Diggle et al. [2013] for an example with log Gaussian Cox models or Grazzini et al. [2017] for a example with agent-based models.

1.1.3 Existing Methodology

The ubiquity of integration across statistics should now be clear to the reader. We now move on to discuss how the problem of numerical integration can be tackled in practice. In this section, we briefly review existing methodology, then discuss some of their shortcomings. Recall that we assume throughout this chapter that $(\mathcal{X}, \mathcal{F}, \mu)$ is a measure space and Π is a probability measure with density (with respect to μ) denoted π.

Monte Carlo Integration and Importance Sampling

Monte Carlo methods are quadrature rules based on uniform weights. The simplest of those methods, which is usually simply referred to as "Monte Carlo" [Robert and Casella, 2004; Glasserman, 2004], consists of obtaining IID realisations $\{\mathbf{x}_i^{\text{MC}}\}_{i=1}^n$ from the measure Π and approximating $\Pi[f]$ as:

$$\hat{\Pi}_{\text{MC}}[f] \quad := \quad \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{x}_i^{\text{MC}}\right).$$

An illustration of such a point set is available in Figure 1.1 for the case where Π is a uniform measure on the unit cube $\mathcal{X} = [0,1]^2$. MC estimators are popular in statistics owing to their wide applicability and their well-known properties. For instance, under regularity conditions (omitted for brevity), the central limit theorem gives that

$$\sqrt{n}\left(\hat{\Pi}_{\text{MC}}[f] - \Pi[f]\right) \quad \xrightarrow{D} \quad \mathcal{N}(0, \text{Var}_\pi[f]), \tag{1.9}$$

where $\xrightarrow{D}$ denotes convergence in distribution. We use the notation $\mathcal{N}(m,c)$ to denote a normal distribution with mean m and covariance c, and $\text{Var}_\pi[f] = \Pi[f^2] - \Pi[f]^2$ is the variance of f under Π. MC is well-suited to numerical integration problems since it provides a dimension-independent convergence rate of $O_P(n^{-1/2})$[3].

A major limitation with MC is the need to sample IID realisations from Π, which is only possible for a limited set of distributions. An alternative estimator with weighted point sets is called importance sampling (IS) and is of the form:

$$\hat{\Pi}_{\text{IS}}[f] \quad := \quad \sum_{i=1}^{n} w_i^{\text{IS}} f\left(\mathbf{x}_i^{\text{IS}}\right), \tag{1.10}$$

[3] We write that some function $f(\mathbf{x})$ is $O(g(\mathbf{x}))$ if the statement "$\exists M, x_0 > 0$ such that $|f(\mathbf{x})| \leq Mg(\mathbf{x})$ whenever $\mathbf{x} \geq \mathbf{x}_0$" holds. Furthermore, we write $f(\mathbf{x})$ is $O_P(g(\mathbf{x}))$ if the statement holds with high probability.

where $\{\mathbf{x}_i^{\text{IS}}\}_{i=1}^n$ are IID realisations from another probability measure Π' called importance measure. This importance measure is defined on $(\mathcal{X}, \mathcal{F})$ and is specified a-priori by the user. Its density with respect to μ satisfies $\pi'(\mathbf{x}) > 0$ whenever $\pi(\mathbf{x})f(\mathbf{x}) \neq 0$, and the IS weights are given by:

$$w_i^{\text{IS}} \quad := \quad \frac{\pi(\mathbf{x}_i)}{n\pi'(\mathbf{x}_i)}, \tag{1.11}$$

for $i = 1, \ldots, n$. The IS estimator in Equation 1.10 can be be seen as an MC estimator where the function $f'(\mathbf{x}) = f(\mathbf{x})\pi(\mathbf{x})/\pi'(\mathbf{x})$ is integrated with respect to the measure Π'. IS is most often used when IID sampling from Π is not feasible, or because clever choices of importance distribution Π' can lead to significant variance reduction in the corresponding central limit theorem. However, IS tends to become inefficient in high dimensions when most samples will have near zero weight. This is due to the fact that, in high dimensions, regions of high probability will tend to be concentrated on small subsets of the sample space $\mathcal{X}$ (a phenomenon known as the curse of dimensionality; see [MacKay, 2003; Betancourt, 2017]).

An illustration of IS is given in Figure 1.1 (middle left), where IID realisations $\{\mathbf{x}_i^{\text{IS}}\}_{i=1}^n$ are obtained from some importance measure Π' which is a truncated Gaussian centred at the origin. The size of the samples is plotted proportional to their weight (as given by Equation 1.11). As observed, there are fewer realisations in the top right corner, but these have larger weights. This compensates for the fact that Π' has very low mass in that part of the domain. The choice of importance distribution will be particularly efficient if the integrand f is such that $\pi'(\mathbf{x}) \propto f(\mathbf{x})$. In this case, Π' would be the optimal importance sampling distribution and the IS estimator would have lower asymptotic variance than the MC estimator.

Markov Chain Monte Carlo Integration

Often we only know π, the density of Π, up to a multiplicative constant. That is, we are able to evaluate $\tilde{\pi}$ where $\pi(\mathbf{x}) = \tilde{\pi}(\mathbf{x})/Z$ for some unknown $Z \in \mathbb{R}^+$. This is for example the case in Bayesian statistics where the probability measure Π is a posterior measure and the normalisation constant Z is the model evidence in Equation 1.4. In this case, neither MC or IS can be used, but MCMC methods [Meyn and Tweedie, 1993; Robert and Casella, 2004] can be a useful alternative. The idea behind MCMC is to generate correlated samples $\{\mathbf{x}_i^{\text{MCMC}}\}_{i=1}^n$ which, marginally, are approximately IID realisations from the target measure Π by obtaining a realisation from a Markov chain whose stationary measure is Π. The estimator of the integral

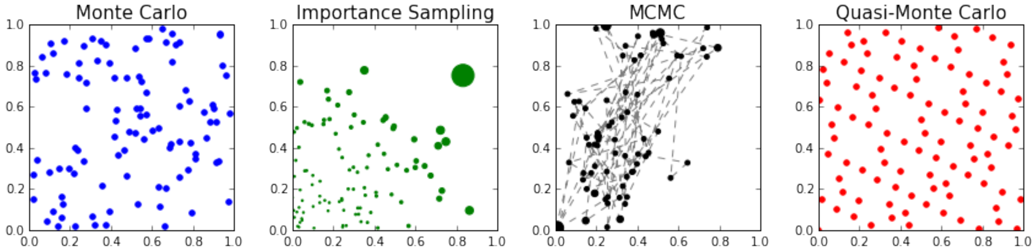

Figure 1.1: *Monte Carlo, importance sampling, Markov chain Monte Carlo and quasi-Monte Carlo (Halton sequence) point sets.* Plot of $n = 100$ points for each algorithm for integration against a uniform distribution on $[0,1]^2$.

then becomes:

$$\hat{\Pi}_{\text{MCMC}}[f] \quad := \quad \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{x}_i^{\text{MCMC}}\right). \tag{1.12}$$

Recall that a Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that the distribution of X_i is only conditional on X_{i-1}. A Markov chain may be specified by an initial measure H_0 (with density h_0) for X_0 and a transition measure T, (with density $t(\cdot|\mathbf{x}) : \mathcal{X} \to \mathbb{R}_+$) from which we can sample. X_i is then a realisation from a measure H_i with density given by $h_i(\mathbf{x}') = \int_{\mathcal{X}} t(\mathbf{x}'|\mathbf{x}) h_{i-1}(\mathbf{x}) \mathrm{d}\mathbf{x}$. The measure Π is called a stationary measure of the Markov chain if whenever X_i is a realisation from Π, then X_{i+1} is also a realisation from Π. This can be summarised succinctly with the following condition: $\pi(\mathbf{x}') = \int_{\mathcal{X}} t(\mathbf{x}'|\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x}$. If the Markov chain is ergodic, it will converge to its stationary measure independently of its initialisation.

The Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970] is the most widely used example of MCMC. It aims to construct a Markov chain converging to the desired target measure Π by the means of a proposal kernel $K : \mathcal{X} \times \mathcal{X} \to [0,1]$, where for each $\mathbf{x} \in \mathcal{X}$, $K(\cdot, \mathbf{x})$ is a probability measure with density $\kappa(\cdot, \mathbf{x}) : \mathcal{X} \to \mathbb{R}$. The algorithm proceeds as follows. First, draw a realisation $\mathbf{x}_0 \in \mathcal{X}$ from H_0. Then, at each iteration, given the current state $\mathbf{x}_i \in \mathcal{X}$:

1. Propose a new state $\tilde{\mathbf{x}}$ by obtaining a realisation from $K(\cdot, \mathbf{x}_i)$.

2. Accept the proposed state (i.e. set $\mathbf{x}_{i+1} = \tilde{\mathbf{x}}$) with probability $A(\tilde{\mathbf{x}}|\mathbf{x}_i) := \min\left\{1, \frac{\pi(\tilde{\mathbf{x}})\kappa(\mathbf{x}_i,\tilde{\mathbf{x}})}{\pi(\mathbf{x}_i)\kappa(\tilde{\mathbf{x}},\mathbf{x}_i)}\right\}$, else keep the previous state (i.e. set $\mathbf{x}_{i+1} = \mathbf{x}_i$).

This induces a transition kernel $T : \mathcal{X} \times \mathcal{X} \to [0,1]$ where for fixed $\mathbf{x} \in \mathcal{X}$, induces a distribution $T(\cdot, \mathbf{x})$. When it exists, the density of $T(\cdot, \mathbf{x})$ is given by:

$$t(\mathbf{x}'|\mathbf{x}) \quad := \quad \kappa(\mathbf{x}', \mathbf{x})A(\mathbf{x}'|\mathbf{x}) + 1_{\{\mathbf{x}'=\mathbf{x}\}} \left(1 - A(\mathbf{x}'|\mathbf{x})\right),$$

where $1_{\{\mathbf{x}'=\mathbf{x}\}}$ takes value 1 when $\mathbf{x}' = \mathbf{x}$ and 0 otherwise. Note that the computation of $A(\mathbf{x}'|\mathbf{x})$ does not rely on the constant Z, due to a cancellation in the ratio.

In principle, there are only mild requirements on the proposal kernel required to obtain an asymptotically correct algorithm. The choice of proposal kernel will however have a high influence on the performance of the algorithm. Intuitively, the aim is to choose a proposal kernel which will favour values with high probability of acceptance. Concurrently, we would also like the proposal kernel to be designed so that chain explores the state space well in a small number of iterations, so that the realisations are as close to IID as possible.

A common choice is a symmetric distribution centred on the current state of the chain, which gives the well-known random-walk Metropolis algorithm. This algorithm is illustrated in Figure 1.1 (middle right) where we have used a Gaussian proposal centred at the current state and with variance 0.1. The black dots give the samples of the Markov chain and their size depends on the number of time they are repeated in the chain. On the other hand, the dotted lines indicate the path of the chain. This example can highlight the difficulty of tuning Markov chains; the chain is not very efficient at covering the whole space and this could most likely be resolved by increasing the variance of the proposal. It is also often possible to use more efficient transition kernels.

A more advanced algorithm is the Metropolis-adjusted Langevin algorithm [Rossky et al., 1978; Scalettar et al., 1986; Roberts and Rosenthal, 1998], which exploits gradients by approximating the path of a diffusion which is invariant to the target distribution. Duane et al. [1987] also later proposed a method based on approximating Hamiltonian dynamics with potential energy given by the log target density. This method was originally named Hybrid Monte Carlo, but is also commonly known as Hamiltonian Monte Carlo [Neal, 2011; Betancourt, 2017; Barp et al., 2018]. Informally, these two methods have the advantage of using transition kernels directing the Markov chain towards areas of high probability and are hence preferable to the random-walk Metropolis algorithm above.

Another alternative to these algorithms is to restrict our proposal measure to a parametric family of transition kernels. We then assume that a member of this family is a good choice, and attempt to learn the corresponding parameter on the fly. Algorithms of this form are called adaptive MCMC algorithms [Gilks et al., 1994;

Haario et al., 2001; Andrieu and Thoms, 2008]. Adaptive MCMC algorithms can be very efficient, but proving their correctness is difficult since the Markov property no longer holds since we are allowing the process to depend on more than the current state.

Sequential Monte Carlo samplers

An approach which combines ideas from IS and MCMC are sequential Monte Carlo (SMC) methods. SMC (and other particle-based schemes) have had an enormous influence in signal processing, and more generally filtering and smoothing in a Bayesian context [Doucet and Johansen, 2011; Särkkä, 2013]. More recently SMC algorithms have been proposed to sample from complex distributions, and these can be particularly efficient for multimodal target distributions.

SMC samplers [Chopin, 2002; Del Moral et al., 2006] start by defining a sequence of probability measures $\Pi_0, \Pi_1, \ldots, \Pi_T$ where $\Pi_T = \Pi$ is the measure we would like to integrate against. The main idea behind SMC is to sequentially obtain realisations from this sequence of measures by moving a set of particles. If Π_0 is simple to sample from, and moving particles across consecutive measures in this sequence is also relatively easy, then SMC samplers can render the task of sampling from Π manageable even when sampling from Π directly (e.g. using MCMC) would be difficult of even infeasible. The algorithm follows the following step. First, we start by obtaining particles $\{\mathbf{x}_i\}_{i=1}^n$ as IID realisations from Π_0, then at each iteration of the algorithm:

1. Update the weights using the formula for IS weights in Equation 1.11 with importance distribution Π_{t-1} and target Π_t.

2. If some resampling criterion (described below) is satisfied, do a resampling step. This means sampling (with replacement) from our current set of particles according to their respective weights and setting the weights of the resampled particle to $1/n$.

3. Update the particles using MCMC step(s) with invariant measure Π_t.

Once iteration T is attained, a last resampling step is used to obtain a final set of equally-weighted particles which we will denote $\{\mathbf{x}_i^{\text{SMC}}\}_{i=1}^n$. When this procedure is completed, we end up with an estimator:

$$\hat{\Pi}_{\text{SMC}}[f] \quad := \quad \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{x}_i^{\text{SMC}}\right).$$

The resampling strategy can be useful to avoid the degeneracy of particle weights which is common with IS methods (i.e. most particles end up having near zero weight). The most common resampling approach is the multinomial sampling described above, but alternatives can be more efficient [Douc et al., 2005].

Note that to avoid resampling at every step, it is common to use a criterion based on the variability of the current samples such as the effective sample size. Another example, called conditional effective sample size, was proposed by Zhou et al. [2016]. Given a set of weighted particles $\{\mathbf{x}_i, w_i\}_{i=1}^n$ at iteration j, it can be computed as:

$$\text{CESS} \;=\; \frac{n\left(\sum_{i=1}^n w_i z_i\right)^2}{\left(\sum_{i=1}^n w_i z_i^2\right)},$$

where $z_i = (\pi(\mathbf{x}_i)/\pi_0(\mathbf{x}_i))^{(t_j - t_{j-1})}$ for $i = 1, \ldots, n$ and π_0 is the density of Π_0.

Quasi-Monte Carlo Integration

All of the methods we have seen so far focus on approximating the target measure Π. When Π is simple, it is common to exploit properties of the integrands instead. Quasi-Monte Carlo (QMC) methods [Dick and Pillichshammer, 2010; Dick et al., 2013] are estimators based on point sets with grid-like structures and uniform weights, usually defined on some domain $\mathcal{X}$ which is the unit cube and for integration against a measure Π which is the uniform measure on this cube:

$$\hat{\Pi}_{\text{QMC}}[f] \;\;:=\;\; \frac{1}{n}\sum_{i=1}^n f\left(\mathbf{x}_i^{\text{QMC}}\right).$$

The point set $\{\mathbf{x}_i^{\text{QMC}}\}_{i=1}^n$ is chosen to minimise some notion of discrepancy between an empirical measure and the measure Π. For this reason, they are often referred to as "space-filling designs", and different notions of discrepancy lead to different QMC rules. These designs are either nested (i.e. the point set with $n+1$ points can be obtained by adding one point to the point set of size n) in which case they are called "open", or non-nested (the point set of size $n+1$ needs to be recomputed from scratch) in which case they are called "closed".

An example of an open QMC point set is the Halton sequence, which is given in the case $d = 2$ in red in Figure 1.1 (right). It can be observed that the sequence fills the space in a much more uniform way than the plot of MC points (in blue).

This space filling property means that the QMC rules can usually attain faster convergence than MC methods. Under mild conditions on f, the error decreases at the asymptotic rate $O(n^{-1+\epsilon})$ where ϵ denotes log terms. Specific methods have also been used to obtain fast convergence rates of $O(n^{-\frac{\alpha}{d}+\epsilon})$ in the classical Sobolev spaces[4] $W_2^{\alpha}(\mathcal{X})$, where α denotes the number of weak derivatives of functions in the space.

It is also well-known [Sloan and Woźniakowski, 1998] that QMC can perform particularly well in high dimensions; for example, a dimension-independent convergence rate of $O(n^{-\alpha+\epsilon})$ can be proved for Sobolev spaces of mixed dominating smoothness (usually denoted $\mathcal{S}_2^{\alpha}(\mathcal{X})$).

Another direction of research has been randomised quasi-Monte Carlo which proposes to randomise QMC point sets in a way which preserves their space-filling properties. A particular example is the scrambling method of Owen [1997], which can also be shown to converge fast for smooth functions (in mean-squared error).

Although QMC methods can be used to obtain fast convergence rates, they tend to be impractical for many applications due to their restriction to the cube and the uniform measure. Several ways of avoiding this issue have been proposed in the literature, mostly focusing on transforming alternative problems to fit in this setup, but these tend to be impractical.

Classical Deterministic Quadrature Rules

As already pointed out, another alternative which has historically been popular but is now rarely used in modern statistical inference problems are classical deterministic quadrature rules [Davis and Rabinowitz, 2007]. These rules are usually designed to integrate functions on some interval $(a,b) \subset \mathbb{R}$, and the weights and points are often chosen so as to integrate any polynomial up to a certain degree exactly.

The simplest examples include the midpoint rule, consisting of one point $x_1 = (b-a)/2$ and weight $w_1 = |b-a|$, and the trapezoidal rule, consisting of the two points $x_1 = a, x_2 = b$ and weights $w_1 = w_2 = |b-a|/2$. The two rules integrate exactly all polynomials of degree 0 and 1 respectively, and are part of the family of Newton-Cotes rules which are based on equally-separated points. These rules can be either open[5], in which case they integrate polynomials passing through all including the boundary exactly, or closed, in which case they do not evaluate integrands on the boundary. See Figure 1.2 for an illustration.

Another class of quadrature rule, which can integrate polynomials of order up to $n-1$ with n points are the Gaussian quadrature rules [Golub and Welsch, 1969]. Different examples of Gaussian quadrature rules exist, depending on the measure

[4] See Appendix A.1 for definitions and some additional background on functional analysis.

[5] Note that the concepts of open and closed are different to those used in the QMC literature.

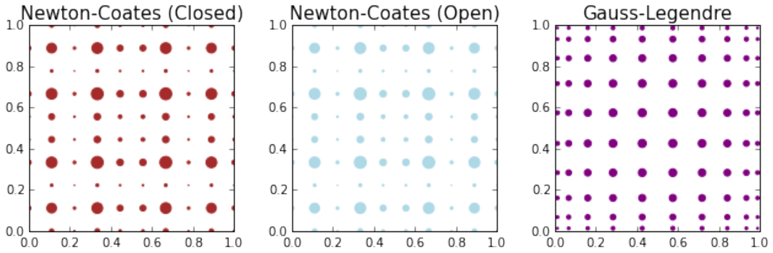

Figure 1.2: *Closed Newton-Coates, open Newton Coates and Gauss-Legendre point sets.* Plot of $n = 100$ points of each method for integration against a uniform distribution on $[0, 1]^2$.

against which the integral is taken. For example, the Gauss-Hermite rule can be used when the measure Π is Gaussian and the Gauss-Legendre rule (see Figure 1.2) when the measure Π is uniform.

Similarly to some QMC sequences, classical deterministic quadrature rules can also be nested: see for example the class of Fejér quadrature rules (also called Clenshaw-Curtis quadrature rules).

The reason for their lack of use in statistics is that they are usually limited to integration over one-dimensional intervals. Several attempts have been proposed to scale these to multiple dimensions (in which case they are called cubature rules), including tensor product structures and sparse quadrature structures such as Smolyak sparse grids. However, these methods have not really been used in statistics due to the fact that most classical deterministic quadrature rules require integrals to be done against very simple measures such as the uniform or Gaussian measure.

Laplace Approximations and Variational Inference

We conclude with a brief discussion of optimisation-based methods such as the Laplace approximation and variational inference. Although these methods do not of themselves replace numerical integration, they are often used in order to approximate distributions, and integrals with respect to the target distribution are then replaced by integrals with respect to these surrogates.

The simplest approach is the Laplace approximation, which consists of fitting a Gaussian with mean at the mode of the posterior and using the local geometry of this mode for the covariance. This will of course be efficient if the posterior is peaked and resembles a Gaussian, but can be extremely poor if the posterior has heavy tails or is multimodal. Efficient modern implementations include the integrated nested Laplace approximation (INLA) [Rue et al., 2009, 2016], which focuses on the class of latent Gaussian models.

An alternative approach is variational inference [Jordan et al., 1999; Blei et al., 2017]. The aim here is to approximate some challenging probability density (usually a Bayesian posterior), by choosing a parametric class of distributions, and approximating the target density by the member of this class which is the closest in some notion of distance (usually a statistical divergence).

This approach has the advantage that it is much less computationally demanding than most advanced Monte Carlo methods, but it also has several disadvantages. Firstly, it can be fairly limited if the variational family is not large enough. Indeed, the main issue with variational inference is that it is not asymptotically exact. That is, even with $n \to \infty$, we have no guarantee that the approximation error will tend to zero if the target measure is not in the variational family. The method is therefore not recommended if precise approximations are required.

Secondly, the divergences used in variational inference are non-convex objectives and therefore cannot be minimised exactly. As a result, variational inference approaches often end up significantly underestimating or overestimating the variance of the target.

1.1.4 Issues Faced by Existing Methods

Following the introduction of common tools in numerical integration, we highlight some of the challenges for these methods.

1) High Computational Cost

The most obvious issue is that of densities or integrands which are expensive to evaluate. The term "expensive" can refer to either computational time or financial cost. For example, complex integrands can take several hours on a computer to be evaluated. Alternatively, in medical applications, evaluating an integrand might mean having to run a set of experiments on some patients, which may incur a large financial cost. These costs mean that a limited set of integrand evaluations are available.

A class of problems for which this occurs is when we have to use a numerical method at each evaluation of the density or integrand (or in fact any of their derivatives). This is the case in the field of uncertainty quantification [Sullivan, 2016] and inverse problems [Stuart, 2010; Dashti and Stuart, 2016] where evaluating the likelihood often requires solving a differential equation numerically. Bui-Thanh and Girolami [2014] give an example of Bayesian inference in a heat conduction problem which requires solving a partial differential equation with finite element methods every time we want to obtain a realisation from the posterior. See also Mohamed et al. [2010] for a similar problem in reservoir simulation, Martin et al. [2012] in seismic models and Petra et al. [2014] for ice sheet models. Other challenging applications include Gaussian process models, which require numerically inverting a potentially large positive definite matrix [Rasmussen and Williams, 2006].

A second class of problems is the so-called "tall data" problem, where the number of samples entering the likelihood is very large. Examples of application fields where this is a problem include astronomy [Sharma, 2017], spatial statistics [Møller and Waagepetersen, 2004], as well as machine learning methods, e.g. topic models [Griffiths and Steyvers, 2004; Blei et al., 2012] or neural networks [Goodfellow et al., 2016].

This is particularly challenging for Bayesian statistics since the posterior distribution can become too computationally expensive to evaluate or simulate from exactly, and has lead researchers to develop a range of new approximate algorithms; see [Angelino et al., 2016] for an overview. Bardenet et al. [2017] also offer a discussion of solutions in the Monte Carlo literature, whilst Hoffman et al. [2013] discuss this issue in the context of variational inference. Another direction of research in the tall data setting has been to consider methods to summarise large datasets with a subset of representative weighted samples. This is called a coreset [Bachem et al., 2017; Huggins et al., 2016; Campbell and Broderick, 2017] and can be used instead of the entire dataset to reduce the computational cost associated with evaluating likelihoods. However, these methodologies are still in their infancy and further developments are required.

2) High Dimensionality

A second challenge is the problem of concentration of measure that is particularly problematic in high dimensions (and hence often called curse of dimensionality). This concentration means that most of the state space has negligible probability mass, and therefore uninteresting from an approximation point of view. Designing samplers which can probe the relevant subset of the sample space $\mathcal{X}$ is therefore challenging yet of critical importance.

To highlight only a few examples, sampling from the posterior over Bayesian neural networks parameters is extremely challenging and requires efficient MCMC proposals [Neal, 1995]. High dimensionality can also be a particular challenge in model selection [Johnson and Rossell, 2012] and its applications to genomics [Li and Zhang, 2010], or in phylogenetics [Larget and Simon, 1999; Mau et al., 1999] when sampling high dimensional structured spaces such as trees. Finally, we point out that it is sometimes desirable to sample from spaces of functions. Applications include fluid dynamics and computational tomography [Cotter et al., 2013].

A common solution for MCMC is to focus on samplers which take into account first and second order gradient information of $\log \pi$, such as the Hamiltonian Monte Carlo samplers discussed above. These can provide more efficient updates as compared to simpler algorithms which do not take into account the density in the proposal. For quadrature rules based on functional approximation, a common solution is to restrict the class of functions we are interested in approximating. See the references in the previous section for an overview in the context of QMC methods.

3) Approximation of Complex Distributions

A particular challenge for sampling methods is when the density π is highly multimodal. The reason is that most of these methods are based on local moves: the next sample is usually obtained by moving away from the location of the current sample. However, in practice it is common for densities to have regions of low probability between the modes, making moves between different modes a rare event. This multimodality problem occurs for example in mixture models [Marin et al., 2005] or in certain models driven by differential equations [Calderhead and Girolami, 2011]. In multimodal cases, it is common to make use of tempering-based algorithms [Swendsen and Wang, 1986; Neal, 1996], but these can be challenging and expensive to implement.

Sampling is also often complicated when the state-space $\mathcal{X}$ is not Euclidean, but instead given by some manifold [Byrne and Girolami, 2013]. Examples of manifolds of interest in statistics include the circle and the sphere [Kent, 1982], which are the central spaces of interest in directional statistics. Alternatively, computing integrals on spaces of structured matrices such as the Stiefel and Grassmann manifold is also useful in signal processing [Srivastava and Klassen, 2004] or computer vision [Turaga et al., 2008]. Finally, another important scenario occurs in model comparison, where sampling is sometimes done jointly across parameters and models and the sample space is hence highly complex [Green, 1995].

4) Quantification of Numerical Error

Of course, it is only ever possible to evaluate an integrand at a finite number of points n, and as such there is usually some numerical error remaining. For this reason, quantifying the error remaining after finite computation is of paramount importance. There is however only very limited work in this area.

In the context of standard MC methods and IS, error estimates are usually based on asymptotic results such as the central limit theorem (recall Equation 1.9). Estimates of the asymptotic variance can be used to approximate the error of the numerical scheme. However, there is in general no guarantee that the finite-sample performance is acceptably close to the asymptotic performance. Similar approaches can also be used for MCMC, with the added difficulty that the Markov structure induces correlation across samples and that convergence of the chain to the target distribution is difficult (if not impossible) to assess. In any case, these estimates tend to be based on very weak assumption and therefore pessimistic in certain cases. Indeed, these estimates are solely based on approximations of the measure with respect to which we are integrating and do not use any properties of the integrand of interest. As such, the same error estimate would be provided regardless of whether we are integrating a constant function or a rough and highly-oscillatory function.

A partial remedy to this problem can be found in the information-based complexity literature [Traub et al., 1983; Novak and Woźniakowski, 2008, 2010; Novak, 2016; Ritter, 2000]. The general approach is to consider some arbitrary function class $\mathcal{H}$, and to study certain types of errors obtained by quadrature rules when integrating functions f in this class. The most popular examples include the worst-case integration error over $\mathcal{H}$, given by:

$$e_{\text{wor}}(\hat{\Pi}; \Pi, \mathcal{H}) \quad := \quad \sup_{\|f\|_{\mathcal{H}}=1} \left| \Pi[f] - \hat{\Pi}[f] \right|. \tag{1.13}$$

Another alternative is the average-case integration error, for which an additional measure $\mu_{\mathcal{H}}$ on the space of functions is required, and which is given by:

$$e_{\text{avg}}(\hat{\Pi}; \Pi, \mu_{\mathcal{H}}) \quad = \quad \int_{\mathcal{H}} \left(\Pi[f] - \hat{\Pi}[f] \right) \mathrm{d}\mu_{\mathcal{H}}(f). \tag{1.14}$$

Unfortunately, these can only be computed for very limited combinations of probability measure Π and function space $\mathcal{H}$ due to the need to compute a supremum over the unit ball of $\mathcal{H}$, or an integral against $\mu_{\mathcal{H}}$. For this reason, these method remain mostly analytical tools which allow theoreticians to guarantee the optimality of certain quadrature rules, rather than a practical tool for the assessment of numerical error.

Finally, it is important to note that certain algorithms have been proposed to approximate the numerical error, and use this approximation to make the methods adaptive to the integrand. However, from an information-based complexity point of view, it can be shown under mild conditions that adaptivity is not helpful in the sense that adaptive algorithms do not lead to faster asymptotic convergence rates for the worst-case or average-case error [Ritter, 2000]. These approaches have however shown to be useful in practice. For example, for classical deterministic quadrature rules, several adaptive schemes have been proposed, usually based on Richardson extrapolation (e.g. Romberg integration, which is a Newton-Coates method with Richardson extrapolation), or epsilon-algorithms [Davis and Rabinowitz, 2007].

1.2 Challenge II: Intractable Models

We have now concluded our initial discussion of numerical integration, the main challenge that will be tackled in this thesis. A second challenge is that of statistical models for which the density is not available. It should be clear from previous sections that both Bayesian inference (see Equation 1.3) and maximum likelihood inference (see Equation 1.7) are likelihood-based inference, meaning that they require us to be able to evaluate the likelihood at different data points and parameter values. However, in the case of complex statistical models this may not be possible, or computationally feasible. We now highlight two such scenarios.

1.2.1 Intractability in Unnormalised Models

A first scenario which is common in applications of statistics is when the likelihood can only be accessed in an unnormalised form:

$$p(\mathbf{x}|\theta) \quad = \quad \frac{\bar{p}(\mathbf{x}|\theta)}{Z(\theta)}, \tag{1.15}$$

where $\tilde{p}(\mathbf{x}|\theta)$ is an unnormalised density which can be evaluated and $Z(\theta) \in \mathbb{R}_+$ is an unknown normalisation constant which depends on the parameter vector θ. Usually this scenario arises due to the high computational cost of evaluating the normalisation constant, or because this constant is itself defined as some intractable integral of the form $Z(\theta) = \int_{\mathcal{D}} \tilde{p}(\mathbf{x}|\theta)\mathrm{d}\mathbf{x}$ (when $\mathcal{D}$ is a continuous domain) or $Z(\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \tilde{p}(\mathbf{x}|\theta)$ (when $\mathcal{D}$ is a discrete, but very large, domain). Examples include Gibbs distributions, which are popular in statistical physics and the study of social networks [Caimo and Mira, 2015], as well as Markov random fields, which are popular in image modelling and spatial statistics [Hyvärinen, 2006, 2007; Moores et al., 2015].

Of course, this can be a particular challenge for maximum likelihood estimation since we need to know the normalisation constant $Z(\theta)$ in order to solve the optimisation problem in Equation 1.7. Such problems have also received a lot of attention in the Bayesian literature, where they are known as “doubly intractable” problems due to the fact that both the normalisation constant of the likelihood and the normalisation constant of the posterior (i.e. the model evidence) are unknown. In these cases, combining Equations 1.3 and 1.15, we get that the posterior distribution takes the form:

$$p(\theta|\mathbf{X}) \quad = \quad \frac{\frac{\bar{p}(\mathbf{X}|\theta)}{Z(\theta)} p_0(\theta)}{p(\mathbf{X})}. \tag{1.16}$$

where both $Z(\theta)$ and $p(\mathbf{X})$ are unknown. To resolve the issue of unknown normalisation constant for the likelihood, several authors have proposed to use plug-in MC and MCMC estimates of the intractable integrals [Geyer, 1991; Lyne et al., 2015] (this clearly this highlights another area where numerical integration is important!).

Other popular approaches have focused on approximations to the likelihood and can be computed at much lower computational cost; see for example the pseudo-likelihood method of Besag [1974] and related composite likelihood methods (see Varin et al. [2011] for an overview). These are however not asymptotically exact and it is not always easy to assess the bias created by the approximations.

In a frequentist setting, issues with these approaches have led to the development of alternative methods to maximum likelihood, most notably score-based inference methods such as score-matching [Hyvärinen, 2006, 2007; Karakida et al., 2016] or proper scoring rules [Gneiting and Raftery, 2007; Dawid, 2007; Parry et al., 2012] (see Chapter 4 for more details). These methods only require access to the gradient of the log-density. Advantages include the fact that we can bypass the computation of expensive normalisation constants whilst still obtaining an asymptotically exact solution since:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) \quad = \quad \nabla_{\mathbf{x}} \log \bar{p}(\mathbf{x}|\theta) + \underbrace{\cancel{\nabla_{\mathbf{x}} \log Z(\theta)}}_{=0} \;=\; \nabla_{\mathbf{x}} \log \bar{p}(\mathbf{x}|\theta). \tag{1.17}$$

where $\nabla_{\mathbf{x}}$ is a vector of partial derivatives with respect to each of the coordinates of $\mathbf{x}$. In a Bayesian setting, pseudo-marginal approaches, including the exchange algorithm [Murray et al., 2006; Møller et al., 2006] have been proposed to sample from posterior distributions efficiently. These usually provide good approximations, but at a high computational cost.

1.2.2 Intractability in Generative Models

A second scenario which recently received renewed interest is that of generative models [Mohamed and Lakshminarayanan, 2016], sometimes also called implicit models or likelihood-free models, for which the likelihood is not available in any form. Instead, we assume that it is possible to obtain IID samples from the model for any value of the parameter vector $\theta \in \Theta$. Let $(\mathcal{U}, \Sigma_{\mathcal{U}}, \mathbb{U})$ be a probability space. Formally we regard generative models as a family of probability measures such that for any value of the parameter $\theta \in \Theta$, we can obtain some IID data $\{\mathbf{x}_i\}_{i=1}^n$ from the corresponding probability measure $\mathbb{P}_\theta$. This data is obtained in two steps: first IID random variables $\{\mathbf{u}_i\}_{i=1}^n$ are obtained from $\mathbb{U}$, then some map $G_\theta : \mathcal{U} \rightarrow \mathcal{X}$ is applied to each of these random variables to obtain $\mathbb{P}_\theta$ distributed random variables: $\mathbf{x}_i = G_\theta(\mathbf{u}_i)$ for $i = 1, \ldots, n$.

Generative models are used throughout the sciences, including in the fields of ecology [Wood, 2010; Beaumont, 2010; Hartig et al., 2011], population genetics [Beaumont et al., 2002] or astronomy [Cameron and Pettitt, 2012]. They also appear in machine learning as black-box models; see for example generative adversarial networks (GANs) [Goodfellow et al., 2014; Dziugaite et al., 2015; Li et al., 2015] and variational autoencoders (VAEs) [Kingma and Welling, 2014].

The problem of inference within generative models is of course very closely related to the classical problem of density estimation [Diggle and Gratton, 1984]. To tackle it, a common approach is the method of simulated moments and its special case of indirect inference [Hall, 2005]. Here, the idea is to simulate data from $\mathbb{P}_\theta$ for a wide range of parameter values $\theta \in \Theta$ and keep the parameter value for which a weighted linear combination of moments (such as the mean or variance) of the samples agree the most with moments of the data simulated from the true data generating process.

Furthermore, another recent approach to this problem relating to optimal transport of measure was discussed in Bassetti et al. [2006]; Bernton et al. [2017]; Genevay et al. [2018], where the authors proposed to minimise the Wasserstein distance, or an approximation thereof, between an empirical probability measure induced by the samples from the true data generating process and the statistical model under consideration.

In a Bayesian context, a common approach to obtain an approximate posterior is approximate Bayesian computation. Here, a parameter value θ is accepted as a sample from the approximate posterior if data generated for this value is close enough (in the sense of some summary statistics) to the data from the true generating process. See Marin et al. [2012]; Lintusaari et al. [2017] for an overview.

1.3 Additional Challenges

We have now introduced the two main challenges studied in this thesis. For completeness, we briefly discuss some of the other contemporary challenges of computational statistics. The list below is of course far from complete.

Parallel Programming First, ongoing research is focusing on how to adapt existing algorithms to new hardware architectures such as GPUs or clusters of computers [Suchard et al., 2010; Lee et al., 2010; Calderhead, 2014]. These new architectures can help scale algorithms significantly, but require reducing communication costs across threads as much as possible. Furthermore, many algorithms in statistics, such as MCMC, are inherently sequential and so completely new algorithms may need to be developed to take advantage of this type of hardware.

Optimisation A second challenge which we will not address in detail in this thesis is that of convex and non-convex optimisation. This is of course useful for solving likelihood-based inference such as the problem of maximum likelihood in Equation 1.7 or profile likelihood approaches [Murphy and van der Vaart, 2000]. Alternatively, it can also be used in regression and functional approximation problems to overcome the high computational costs associated with exact least-squares solutions. Numerical optimisation remains far from a solved problem in similar settings where sampling is challenging: high dimensional, multimodal and expensive applications.

Privacy/Security Finally, the privacy risks associated to the increasing digitalisation of our society have been demonstrated by several authors (see for example de Montjoye et al. [2015]), and recent studies [Kaufman et al., 2009] have demonstrated that the public is getting increasingly sensitive to the risks associated with sharing data about themselves. Another challenge is therefore to develop algorithms for statistical inference and computation which include some notion of privacy. This might mean inference methods with restricted access to data [Graepel et al., 2012], or only access to noisy versions of the data, a common scenario in differential privacy [Dwork, 2008].

1.4 Contributions of the Thesis

We have now concluded our discussion of important challenges in computational statistics. The aim of this thesis is to explore how the theory of kernel methods can be used to address some of the issues discussed in the previous section. Kernel methods can be used for several tasks, most notably functional approximation. They are very flexible since kernel spaces include a wide range of different function spaces with varying regularity and properties such as smoothness and periodicity. They can also be used for tractable computation in high- or infinite-dimensional spaces by making use of a property called the kernel trick.

This thesis makes the following contributions to this area:

- Chapter 1 highlighted some of the main challenges in statistical computation, focusing mainly on issues surrounding numerical integration and statistical inference for models with intractable likelihoods. For numerical integration, we discussed popular approaches in the statistics literature including classical quadrature rules, as well as MC, QMC and MCMC methods. These methods will later be used as a baseline in Chapter 3. This chapter was partly based on Barp et al. [2018].

- Chapter 2 reviews background material, most notably the theory of reproducing kernel Hilbert spaces (RKHS), stochastic processes, and their formal relations. We also discuss how these have been successfully applied in machine learning and statistical modelling, and highlight the strength and weaknesses they provide for computational statistics. Finally, we discuss how stochastic processes can be used in the context of Bayesian nonparametrics, and focus in detail on the particular case of Gaussian processes.

- In Chapter 3, we introduce Bayesian probabilistic numerical methods. We revisit in detail the Bayesian quadrature (BQ) algorithm of O'Hagan [1991], and provide an extensive theoretical analysis of its properties. This includes the first asymptotic convergence results, which will be based on an analysis of quadrature rules in reproducing kernel Hilbert spaces. Later, we discuss details required for the implementation of the method, then study the performance of the method on a wide range of applied problems from computer graphics to inference in dynamical systems. The chapter is partially based on Briol et al. [2015b, 2016]; Oates et al. [2017d]; Briol and Girolami [2018].

- In Chapter 4, we propose several novel extensions to the basic BQ algorithm. The first extension focuses on providing an approach to tackling several numerical integration problems simultaneously by defining the BQ algorithm on a vector-valued function space. This method will be particularly useful when we have an application where multiple integrals of highly correlated functions need to be computed simultaneously or sequentially. The two other novel extensions consist of sampling schemes aimed at taking advantage of the properties of BQ estimator in order to speed up convergence to the solution of the integral. These are based on the Frank-Wolfe algorithm and SMC samplers. This chapter is partially based on Briol et al. [2015a, 2017]; Xi et al. [2018].

- Chapter 5 proposes several applications of kernel methods to solve problems linked to intractable models (including both unnormalised and generative models). In particular, it discusses how Stein's method, a popular tool to assess convergence in probability theory, can be combined with kernel methods to obtain flexible functional approximation tools for unnormalised models. Finally, it discusses inference for unnormalised and generative models in the context of minimal distance estimators. The chapter is partly based on Briol et al. [2017]; Oates et al. [2018]; Chen et al. [2018].
- Chapter 6 concludes with a discussion of the contributions the thesis and discusses potential extensions.

Chapter 2

Kernel Methods, Stochastic Processes and Bayesian Nonparametrics

> "Probability theory is nothing but common sense reduced to calculation."
>
> Pierre-Simon Laplace

Reproducing kernel Hilbert spaces (RKHS) have had a significant impact in the mathematical sciences. This is mainly due to a property called the "reproducing property", by which many quantities of interest are rendered tractable. Working in a RKHS is therefore a convenient and practical choice.

When this is not directly feasible, it is often possible to embed a given space into another, often larger, space using kernels. The reason embeddings are useful is that many operations which are intractable or complex in the original space can be trivial to implement in the embedding space. There are several ways in which embeddings are commonly used. The first is called the "kernel trick", and consists of replacing inner products in the original space by kernel evaluations. This has the advantage that the kernel is implicitly computing inner products in the embedding space; an operation which may be infeasible by direct computation (since the embedding space might be high-dimensional, or even infinite-dimensional). The second type of embedding is the embedding of probability measures into a RKHS, which allows comparison of these measures in a straightforward way. These two types of embeddings will be used throughout this thesis to study quadrature rules.

RKHSs are also useful in a different context: the study of stochastic processes. In particular, it can be shown that every stochastic process with finite second moment has a covariance function which corresponds to a reproducing kernel and vice-versa. This is for example the case for Gaussian processes (GP). The RKHS corresponding to the covariance of a GP is called its native space, and it can be used to understand properties of the process. Finally, GPs and kernel methods have more generally been studied and applied throughout the field of Bayesian nonparametrics, which is concerned with infinite-dimensional Bayesian models.

This thesis makes use of these three intertwined research areas to propose novel algorithms in computational statistics. The following chapter therefore introduces well-known results which will be used throughout later chapters.

2.1 Kernel Methods

2.1.1 Introduction and Characterisations

Although the theory of Banach and Hilbert spaces does already give us a lot of structure to work with, we will focus mainly on a specific subclass of Hilbert spaces called reproducing kernel Hilbert spaces, also sometimes called proper Hilbert spaces[1]. These spaces have the property that functions which are close to one another in the sense of the metric induced by the inner product will have close pointwise values. More precisely, these are classes of functions where the evaluation functional is a continuous mapping. RKHSs were introduced and later studied by many authors, most notably Mercer [1909]; Bergman [1922]; Schoenberg [1937]; Aronszajn [1950] and Schwartz [1964]. For the interested reader, a nice historical survey of these spaces is presented in Stewart [1976]; Fasshauer [2011] and a rigorous modern treatment can be found in Berlinet and Thomas-Agnan [2004]; Schaback and Wendland [2006]; Steinwart and Christmann [2008].

A RKHS on some arbitrary non-empty set $\mathcal{X}$ is characterised by a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ called a kernel[2]. For simplicity, we restrict ourselves to $\mathcal{X} \subseteq \mathbb{R}^d$ for $d \in \mathbb{N}$ in the remainder of this chapter. We say that a kernel is symmetric if it satisfies $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{y}, \mathbf{x})\ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. In the case of a RKHS, we are only interested in a very specific type of kernel called a reproducing kernel:

Definition 1 (Reproducing kernel Hilbert space). *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of the Hilbert space $\mathcal{H}$ if and only if:*

[1] Note that any reader unfamiliar with basic notions in functional analysis and topology is referred to Appendix A.1.

[2] Reproducing kernels are not to be confused with the transition kernels discussed in the previous chapter.

1. $\forall \mathbf{x} \in \mathcal{X}, \quad k(\cdot, \mathbf{x}) \in \mathcal{H}$*,*

2. $\forall \mathbf{x} \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f,\, k(\cdot, \mathbf{x})\rangle_{\mathcal{H}} \;=\; f(\mathbf{x})$*.*

A Hilbert space with such a reproducing kernel is called a reproducing kernel Hilbert space.

The second property above is called the reproducing property and is extremely useful for many applications of these spaces. Due to this property, $k(\cdot, \mathbf{x})$ is often also called the representer of the evaluation functional. $\mathcal{H}$ is known as the native space of k, and a RKHS is often denoted $\mathcal{H}_k$ to emphasise the reproducing kernel k associated to it. Any reproducing kernel leads to a function space with continuous evaluation functionals, as specified by Riesz's representation theorem:

Theorem 1 (**Riesz's representation theorem**. [Berlinet and Thomas-Agnan, 2004], Theorem 1)**.** *A Hilbert space of functions on* $\mathcal{X}$ *has a reproducing kernel if and only if all the evaluation functionals* $e_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}$ *such that* $e_{\mathbf{x}}(f) = f(\mathbf{x}) \; \forall \mathbf{x} \in \mathcal{X}$ *are continuous on* $\mathcal{H}$*.*

An important class of kernels are positive definite kernels. We say that the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite kernel if:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \quad > \quad 0, \tag{2.1}$$

$\forall n \in \mathbb{N}, \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, and for all non-zero $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. Positive definite kernels are important because for every positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exits a unique RKHS $\mathcal{H}_k$ with k as its reproducing kernel and, on the other hand, the reproducing kernel of a RKHS is unique and positive definite. This result, first proved by Aronszajn [1950], is called the Moore-Aronszajn theorem.

Theorem 2 (**Moore-Aronszajn theorem**. Berlinet and Thomas-Agnan [2004], Theorem 3)**.** *Let* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a positive definite kernel. There exists only one Hilbert space* $\mathcal{H}_k$ *of functions on* $\mathcal{X}$ *with* k *as reproducing kernel. The subspace* $\mathcal{H}_0$ *of* $\mathcal{H}_k$ *spanned by the functions* $k(\cdot, \mathbf{x})$ *for* $\mathbf{x} \in \mathcal{X}$ *is dense in* $\mathcal{H}_k$ *and* $\mathcal{H}_k$ *is the set of functions on* $\mathcal{X}$ *which are point-wise limits of Cauchy sequences in* $\mathcal{H}_0$ *with the inner product*

$$\langle f, g\rangle_{\mathcal{H}_0} \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(\mathbf{y}_j, \mathbf{x}_i),$$

where $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$ *and* $g(\mathbf{x}) = \sum_{j=1}^{m} \beta_j k(\mathbf{x}, \mathbf{y}_j)$*.*

Note that good intuition can be gained concerning the functions that actually constitute a RKHS $\mathcal{H}_k$ by looking at the form of f and g in the above theorem. This form makes it clear that many basic properties of the reproducing kernel viewed as a function of one argument in its representer form $k(\cdot, \mathbf{x})$ will be inherited by functions in $\mathcal{H}_k$. This is for example the case for properties such as periodicity or smoothness.

2.1.2 Properties of Reproducing Kernel Hilbert Spaces

The fact that every positive definite kernel leads to a RKHS is useful since we only need to verify Equation 2.1 to check whether the space associated to a kernel is a RKHS. In general, it is hard to check the property directly, but an alternative approach is available through the following result:

Theorem 3 (**Kernel trick**. Berlinet and Thomas-Agnan [2004], Lemma 1)**.** *Let $\mathcal{H}_0$ be some Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ and let $\phi : \mathcal{X} \to \mathcal{H}_0$. Then, any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as $k(\mathbf{x}, \mathbf{y}) := \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}_0}$ is positive definite.*

To verify whether a given kernel leads to a RKHS, we therefore only need to check whether it can be written as an inner product. The kernel trick is also useful since it allows us to write operations involving high, or infinite-dimensional, spaces using only evaluations of the kernel. In the machine learning literature, the map $\phi : \mathcal{X} \to \mathcal{H}_0$ is often called a feature map, whilst $\mathcal{H}_0$ is known as the feature space, and the above lemma is known as the "kernel trick". Note that the relationship between kernel and feature map (or correspondingly feature space) is not one-to-one.

Many algorithms have been "kernelised", meaning that all inner products are replaced by reproducing kernels [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Hofmann et al., 2008]. Popular examples include support vector machines and principal component analysis. In mathematical language, the kernel trick corresponds to an embedding of the space $\mathcal{X}$ into the space $\mathcal{H}_0$, and the feature space corresponds to an embedding space.

We now give several examples of feature maps. First, an obvious choice is $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$ where the feature space is the RKHS itself: $\mathcal{H}_0 = \mathcal{H}_k$. Another example is given by Mercer's theorem, which was originally proposed in [Mercer, 1909]. See Riesz and Nagy [1990] for a detailed discussion and proof, but for convenience in our context we consider the following simplified version as proposed by Muandet et al. [2016].

Theorem 4 (**Mercer Theorem**. [Muandet et al., 2016], Theorem 2.1)**.** *Let $\mathcal{X}$ be a compact Hausdorff space and μ a finite Borel measure with support $\mathcal{X}$. Suppose*

k is a continuous positive definite kernel on $\mathcal{X} \times \mathcal{X}$, and assume that it satisfies $\int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} > 0$ for any non-zero $f \in L^2(\mathcal{X}; \nu)$ where $L^2(\mathcal{X}; \nu)$ is the space of functions with $\int_{\mathcal{X}} f(\mathbf{x})^2 \nu(\mathrm{d}\mathbf{x}) < \infty$. Define the integral operator $\mathcal{K} : L^2(\mathcal{X}; \nu) \to L^2(\mathcal{X}; \nu)$, called the Hilbert-Schmidt operator, as:

$$\mathcal{K}[f](\mathbf{x}) \quad := \quad \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \nu(\mathrm{d}\mathbf{x}').$$

Then there is an orthonormal basis $\{\psi_i\}$ of $L^2(\mathcal{X}; \nu)$ consisting of eigenfunctions of $\mathcal{K}$ such that the corresponding sequence of eigenvalues $\{\lambda_i\}$ are non-negative. The eigenfunctions corresponding to non-zero eigenvalues are continuous functions on $\mathcal{X}$ and the kernel has the representation:

$$k(\mathbf{x}, \mathbf{y}) \quad = \quad \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y}),$$

where the convergence of the series is absolute and uniform.

All reproducing kernels satisfying this theorem are called Mercer kernels. For Mercer kernels, we can easily obtain an explicit expression for a feature map $\phi : \mathcal{X} \to \mathcal{H}_0$ of the form $\phi(\mathbf{x}) = \left(\sqrt{\lambda_1}\psi_1(\mathbf{x}), \sqrt{\lambda_2}\psi_2(\mathbf{x}), \ldots\right)$ where $\mathcal{H}_0$ is the space of square-summable sequences.

We note that Theorem 4 requires continuity of the kernel on $\mathcal{X} \times \mathcal{X}$. From now on we will assume that this is the case for all kernels in this thesis. For $\mathcal{X}$ being a bounded interval in $\mathbb{R}$, this assumption means that all of the functions in the RKHS will be continuous. More general conditions for continuity of the elements of a RKHS can be found in Section 1.5 of Berlinet and Thomas-Agnan [2004] (see for example Theorem 17 therein). Several other properties of RKHSs are also worth mentioning.

The first one is the concept of universality of the RKHS. In general, we say that a RKHS is universal if it is rich enough to approximate any function of interest arbitrarily well in some function class. There exists multiple notions of universality, each depending on the choice of domain $\mathcal{X}$, function space we want to approximate and the type of approximation. See Sriperumbudur et al. [2010a] for an overview.

The second notion is that of a characteristic kernel, which relates to embedding of probability measures in RKHSs [Sriperumbudur et al., 2010b]. Denote by $\mathcal{P}(\mathcal{X})$ the set of all Borel probability measures defined on the topological space $\mathcal{X}$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be characteristic if the function $\Pi[k(\cdot, \mathbf{x})] = \int_{\mathcal{X}} k(\cdot, \mathbf{x}) \Pi(\mathrm{d}\mathbf{x}) \in \mathcal{H}_k$ exists and is injective for all probability measures $\Pi \in \mathcal{P}(\mathcal{X})$ in the set. That is, any element of $\mathcal{P}(\mathcal{X})$ is embedded to a unique element in $\mathcal{H}_k$ called the kernel mean or mean element.

Finally, the third property is that certain kernels induce a metric on the space $\mathcal{X}$, defined as [Schoenberg, 1937]: $d_k(\mathbf{x},\mathbf{y}) := \sqrt{k(\mathbf{x},\mathbf{x}) - 2k(\mathbf{x},\mathbf{y}) + k(\mathbf{y},\mathbf{y})}$; see Berg et al. [1984] Chapter 3 Section 3 for an in-depth discussion. Clearly the metric implies a notion of distance between points which depend on how similar their features are (as given by properties of the feature/embedding map). It is therefore possible to induce geometries of interest by reverse-engineering feature maps and the choice of kernel will thus have considerable impact on applications.

2.1.3 Examples of Kernels and their Associated Spaces

We now highlight some of the popular choices in the literature, focusing mainly on real valued kernels:

1. The family of polynomial kernels is given by:

$$k(\mathbf{x},\mathbf{y}) \quad := \quad \left(\mathbf{x}^\top \mathbf{y} + c\right)^p, \tag{2.2}$$

   where $c > 0$ and $p \in \mathbb{N}$. The RKHS corresponding to this kernel is a finite-dimensional vector space and consists of all real valued p^{th} order polynomials on $\mathcal{X}$.

2. The family of Matérn kernels [Matérn, 1960] is given by:

$$k_\nu(\mathbf{x},\mathbf{y}) \quad := \quad \lambda \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{y}\|_2}{\sigma} \right) J_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{y}\|_2}{\sigma} \right), \tag{2.3}$$

   where J_ν is a Bessel function of the second kind, $\lambda, \sigma > 0$. The RKHS induced from k_ν is norm-equivalent to Sobolev spaces $W_2^\nu(\mathcal{X})$ [Adams and Fournier, 2003][3]. Note that another class of kernels which are norm equivalent to Sobolev spaces are Wendland's polynomial kernels [Wendland, 2005].

   The expression for the Matérn kernels is be tedious to evaluate in general, but simplifies when $\nu = \frac{1}{2} + p$ for some $p \in \mathbb{N}$. A few popular examples are given below:

   - $k_{\frac{1}{2}}(\mathbf{x},\mathbf{y}) = \lambda \exp\left(-\|\mathbf{x}-\mathbf{y}\|_2/\sigma\right)$,
   - $k_{\frac{3}{2}}(\mathbf{x},\mathbf{y}) = \lambda \left(1 + \sqrt{3}\|\mathbf{x}-\mathbf{y}\|_2/\sigma\right) \exp\left(-\sqrt{3}\|\mathbf{x}-\mathbf{y}\|_2/\sigma\right)$,
   - $k_{\frac{5}{2}}(\mathbf{x},\mathbf{y}) = \lambda \left(1 + \sqrt{5}\|\mathbf{x}-\mathbf{y}\|_2/\sigma + 5\|\mathbf{x}-\mathbf{y}\|_2^2/3\sigma^2\right) \exp\left(-\sqrt{5}\|\mathbf{x}-\mathbf{y}\|_2/\sigma\right)$,

[3] See Appendix A.1 for a detailed definition.

where in each case $\lambda, \sigma > 0$.

3. The squared-exponential kernel, also called Gaussian RBF kernel, is given by:

$$k(\mathbf{x}, \mathbf{y}) \quad := \quad \lambda \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}\right), \tag{2.4}$$

   where $\lambda, \sigma > 0$. The functions in this RKHS are smoother and can in fact be shown to be holomorphic (i.e. the functions are infinitely differentiable and equal to their Taylor series). Define $e_{\alpha_i} : \mathcal{X} \to \mathbb{R}$ such that $e_{\alpha_i}(\mathbf{x}) := \sqrt{(\sigma^2)^{-\alpha_i}/(\alpha_i!)}\mathbf{x}^{\alpha_i} \exp\left(-\mathbf{x}^2/2\sigma^2\right)$. Then, Proposition 3.6 in Steinwart et al. [2006] provides the following characterisation: for any function $f : \mathcal{X} \to \mathbb{R}$ (with $\mathcal{X} \subset \mathbb{R}^d$ with non-empty interior) in the RKHS with exponentiated-quadratic kernel with lengthscale σ, $\exists (b_\alpha)$ (where $\alpha := (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ is a multi-index) satisfying $\sum_{\alpha \in \mathbb{N}_0^d} b_\alpha^2 < \infty$ such that $f(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}_0^d} b_\alpha (e_{\alpha_1} \otimes \ldots \otimes e_{\alpha_d})(\mathbf{x})$ where $\otimes$ denotes the tensor product (i.e. $\forall g, h : \mathcal{Y} \to \mathbb{R}$ where $\mathcal{Y} \subseteq \mathbb{R}$, $g \otimes h(y, y') = g(y)h(y') \; \forall y, y' \in \mathcal{Y}$).

The last two kernels above are translation-invariant and radial; i.e. $\exists f, g : \mathbb{R} \to \mathbb{R}$ such that the kernels can be written either as $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x} - \mathbf{y})$ and $k(\mathbf{x}, \mathbf{y}) = g(\|\mathbf{x} - \mathbf{y}\|_2)$ respectively. This is common for most RKHS used in applications, since this property allows us to study the induced RKHS through Fourier analysis [Wendland, 2005]. Together with rotation invariance, these are also convenient modelling assumptions.

To construct more complex reproducing kernels, one strategy consists of combining several base kernels [Rasmussen and Williams, 2006]. For example, given two real-valued reproducing kernels k_1 and k_2, the sum $k(\mathbf{x}, \mathbf{y}) := k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$ and product $k(\mathbf{x}, \mathbf{y}) := k_1(\mathbf{x}, \mathbf{y})k_2(\mathbf{x}, \mathbf{y})$ are also reproducing kernels. Furthermore, given some function $a : \mathcal{X} \to \mathbb{R}$, the rescaling $k(\mathbf{x}, \mathbf{y}) := a(\mathbf{x})k_1(\mathbf{x}, \mathbf{y})a(\mathbf{y})$ and convolution $k(\mathbf{x}, \mathbf{y}) := \int_{\mathcal{X} \times \mathcal{X}} k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{z}, \mathbf{z}')k_1(\mathbf{z}', \mathbf{y})\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{z}'$ are reproducing kernels. Finally, the restriction of a kernel on some domain $\mathcal{X}$ to some domain $\mathcal{Y} \subset \mathcal{X}$ is also a reproducing kernel. An interesting discussion of the resulting kernels can be found in Duvenaud [2014].

2.1.4 Applications and Related Research

The useful properties of RKHSs discussed above have certainly helped spread the use of these spaces to a wide range of applications. Although it is out of the scope of this thesis to give a complete introduction, we now provide a brief overview of these applications.

First, RKHSs had a very large influence in the statistics community, most notably in the theory of kriging [Krige, 1951], a widely used method for interpolation in geostatistics. A detailed historical review of kriging highlighting the importance of kernels (called variograms in this literature) is provided by Cressie [1990]. RKHS theory has also been useful for providing a theoretical study of the closely related spline interpolation [Wahba, 1991]. Kernels have also been used to study Gaussian processes [Stein, 1999; Berlinet and Thomas-Agnan, 2004; Rasmussen and Williams, 2006]. A detailed description of the relationship between kernels and GPs can be found in Kanagawa et al. [2018]. Most recently, it has also been useful in other areas of statistics, such as hypothesis testing [Gretton et al., 2006, 2008, 2012a; Chwialkowski et al., 2016] and sampling methods [Chen et al., 2010; Sejdinovic et al., 2014; Strathmann et al., 2015; Liu and Wang, 2016; Chen et al., 2018].

In the numerical analysis literature, reproducing kernels are used to analyse differential equations [Bergman and Schiffer, 1953], and to design numerical solvers such as meshless methods [Babuska et al., 2003]. They also have a central role in approximation theory; see Buhmann [2003] and Schaback and Wendland [2006].

Finally, reproducing kernels have had a significant impact in machine learning [Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Hofmann et al., 2008], most notably in learning theory [Cucker and Smale, 2002]. They have also helped project support vector machines [Boser et al., 1992] to the forefront of machine learning techniques.

2.2 Stochastic Processes

We have concluded our introduction to RKHSs and now move on to discuss stochastic processes. As we will see, the theory of stochastic processes, and especially GPs, is closely intertwined with that of RKHSs. Understanding this relation will be important in the theoretical developments of further chapters.

2.2.1 Introduction to Stochastic Processes

Stochastic processes are one of the major tools used throughout probability theory and statistics, and providing a complete overview of this topic is out of the scope of this thesis. In this chapter, we will mostly focus on the notions which will be useful in the following chapters, and highlight connections with the theory of RKHSs. Further details can be found in the books of Doob [1953]; Gikhman and Skorokhod [1969]; Karlin and Taylor [1975]; Grimmett and Stirzaker [2001]; Koralov and Sinai [2007]; Pavliotis [2014]. See also the paper by Meyer [2009] for a historical overview.

To avoid delaying this further, we begin with the definition of stochastic process (also called random process).

Definition 2 (**Stochastic process**). *Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space consisting of an index set $\mathcal{X}$ and its corresponding Borel σ-algebra $\mathcal{B}(\mathcal{X})$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $(\mathcal{Y}, \mathcal{G})$ a measurable space. A stochastic process is a collection $\{g(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$ such that for each fixed $\mathbf{x} \in \mathcal{X}$, $g(\mathbf{x}, \cdot) : \Omega \to \mathcal{Y}$ is a random variable.*

Stochastic processes are informally viewed as random functions. For a fixed $\mathbf{x} \in \mathcal{X}$, a stochastic process is a $\mathcal{Y}$-valued random variable, whereas for a fixed $\omega \in \Omega$, it consists of a (deterministic) function $g(\cdot, \omega) : \mathcal{X} \to \mathcal{Y}$.

The set $\mathcal{X}$ is known as the sample space, where $\mathcal{Y}$ is the state space of the stochastic process. In the literature, the sample space is often denoted using the dummy variable T due to the historical context of random functions over time. However, it is now common to have $\mathcal{X}$ be a multidimensional index (e.g. time and space). In particular, when $\mathcal{X} \subseteq \mathbb{R}^2$, the stochastic process is often called a random field. Note that $\mathcal{X}$ can be either a finite or infinite index set.

The stochastic processes that we will look at in later chapters will have $\mathcal{X}$ and $\mathcal{Y}$ being Euclidean spaces. For this reason, we will limit ourselves to this level of generality for the remainder of the chapter.

A first example of stochastic process that we have already encountered in this thesis are the discrete-time Markov chains used in MCMC methods, for example the random-walk Metropolis algorithm with Gaussian proposal (see Chapter 1). In this case $\mathcal{X}$ is clearly discrete and the process is real-valued. A second example is the Langevin diffusion which was used to construct the Metropolis-adjusted Langevin algorithm. In this case the index set is one dimensional and continuous: $\mathcal{X} = \mathbb{R}_+$. Furthermore, the discretisation of the diffusion is itself also a stochastic process, but defined on a discrete space $\mathcal{X}$.

2.2.2 Characterisations of Stochastic Processes

Now that we have introduced stochastic processes, we can ask ourselves how to characterise and classify them further. There are two main ways in which we can characterise stochastic processes, through their finite-dimensional distributions, and through their Karhunen-Loève expansion.

Characterisation via Finite-Dimensional Distributions

The finite-dimensional distributions of a stochastic process is the family of distributions of the $\mathcal{Y}^n$-valued random variables $(g(\mathbf{x}_1, \cdot), \ldots, g(\mathbf{x}_n, \cdot))$ for all $n \in \mathbb{N}$ and $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$.

There are several important properties of stochastic processes which are usually specified using finite-dimensional distributions of the process. First, we say that a stochastic process is stationary if and only if the finite-dimensional distributions are invariant with respect to shifts in the index set. In other words, the process is stationary if the distribution of $(g(\mathbf{x}_1, \cdot), \ldots, g(\mathbf{x}_n, \cdot))$ is the same as that of $(g(\mathbf{x}_1 + \mathbf{x}', \cdot), \ldots, g(\mathbf{x}_n + \mathbf{x}', \cdot))$ for all $\mathbf{x}' \in \mathcal{X}$ such that $\mathbf{x}_i + \mathbf{x}' \in \mathcal{X}$ for all $i = 1, \ldots, n$ and $n \in \mathbb{N}$. Clearly it is important to understand whether the relation between stochastic process and their finite-dimensional distributions is one-to-one. The answer is yes under certain regularity conditions provided by the theorem below. The result below will be given for real-valued stochastic processes, but this can be significantly generalised as in Dudley [2002].

Theorem 5 (**Kolmogorov Consistency Theorem**, Koralov and Sinai [2007], Theorem 12.8)**.** *Let $\{\mathbb{P}_{\{\mathbf{x}_i\}_{i=1}^n} | \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}, n \in \mathbb{N}\}$ be a family of distributions each associated to the product σ-algebra $\mathcal{B}(\mathbb{R}^n)$. Suppose these satisfy:*

- *For every permutation $\{\mathbf{x}'_i\}_{i=1}^n$ of $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and events $A_1, \ldots, A_n \in \mathcal{F}$ with $n \in \mathbb{N}$:*

$$\begin{aligned}&\mathbb{P}_{\{\mathbf{x}_i\}_{i=1}^n} \left[(g(\mathbf{x}_1, \cdot), \ldots, g(\mathbf{x}_n, \cdot)) \in A_1 \times \ldots \times A_n\right] \\ &\quad = \mathbb{P}_{\{\mathbf{x}'_i\}_{i=1}^n} \left[\left(g(\mathbf{x}'_1, \cdot), \ldots, g(\mathbf{x}'_n, \cdot)\right) \in A_1 \times \ldots \times A_n\right].\end{aligned}$$

- *For every points $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and events $A_1, \ldots, A_n \in \mathcal{F}$ with $n \in \mathbb{N}$:*

$$\begin{aligned}&\mathbb{P}_{\{\mathbf{x}_i\}_{i=1}^n} \left[(g(\mathbf{x}_1, \cdot), \ldots, g(\mathbf{x}_n, \cdot)) \in A_1 \times \ldots \times A_n\right] \\ &\quad = \mathbb{P}_{\{\mathbf{x}_i\}_{i=1}^{n+1}} \left[(g(\mathbf{x}_1, \cdot), \ldots, g(\mathbf{x}_n, \cdot), g(\mathbf{x}_{n+1}, \cdot)) \in A_1 \times \ldots \times A_n \times \Omega\right].\end{aligned}$$

Then there is a unique stochastic process whose finite-dimensional distributions coincide with this collection.

The first example goes back to the Markov chains introduced in the previous chapter (the random-walk Metropolis algorithm and Metropolis-adjusted Langevin algorithm). We notice that in both cases, their finite-dimensional distributions are given by

$$\begin{aligned}
&\mathbb{P}_{\{\mathbf{x}_i\}_{i=1}^n}\left[(g(\mathbf{x}_1,\cdot),\ldots,g(\mathbf{x}_n,\cdot)) \in A_1 \times \ldots \times A_n\right] \\
&\quad = \int_{A_1} \ldots \int_{A_n} T(\mathrm{d}\mathbf{x}_1, \mathrm{d}\mathbf{x}_0) \times \ldots \times T(\mathrm{d}\mathbf{x}_n, \mathrm{d}\mathbf{x}_{n-1}).
\end{aligned}$$

for any event $A_1, \ldots, A_n$ in $\mathcal{F}$ where T denotes the transition kernel of the chain.

This theorem also allows us to introduce our first characterisation of GPs. A real-valued GP is a stochastic process $g : \mathcal{X} \times \Omega \to \mathbb{R}$ such that all the finite-dimensional distributions are Gaussian, i.e., $(g(\mathbf{x}_1,\cdot),\ldots,g(\mathbf{x}_n,\cdot))$ is an $\mathcal{N}(m_n, c_n)$ random variable for some vector n-dimensional vector m_n and c_n an $n \times n$ symmetric non-negative definite matrix $\forall n \in \mathbb{N}$ and $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$. GPs will be the basis of most of the work in later chapters. Extended introductions can be found in Adler [1990]; Stein [1999]; Rasmussen and Williams [2006]. An important property is that two GPs defined on the same measurable space are either equivalent or mutually singular [Feldman, 1958].

Another example of stochastic process are Dirichlet processes [Ferguson, 1973]. We say a stochastic process is a Dirichlet process with base measure $\mathbb{G}$ and concentration parameter α if and only if its finite-dimensional distributions are Dirichlet distributions; i.e. given any finite measurable partition $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ of $\mathcal{X}$, we have that $(g(\mathcal{X}_1,\cdot),\ldots,g(\mathcal{X}_n,\cdot))$ are $\mathrm{Dir}(\alpha\mathbb{G}(\mathcal{X}_1),\ldots,\alpha\mathbb{G}(\mathcal{X}_n))$ distributed for some concentration parameter $\alpha > 0$. Here, the notation Dir is used to denote a Dirichlet distribution. Note that this case would require a more general version of the Kolmogorov extension theorem than that presented in this thesis (see for example Dudley [2002]).

Characterisation via the Karhunen-Loève Expansion

A second characterisation of stochastic processes is as an infinite series of basis functions with random coefficients called a Karhunen-Loève expansion [Loève, 1978]. This expansion will depend on the first two moments of the stochastic process, which are the mean function $m : \mathcal{X} \to \mathcal{Y}$ and covariance function $c : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$. Denote by $\mathbb{E}_{\mathbb{P}}[X]$ the expectation of some random variable X under $\mathbb{P}$. The mean and covariance function are defined as:

$$\begin{aligned}
m(\mathbf{x}) &:= \mathbb{E}_{\mathbb{P}}\left[g(\mathbf{x},\omega)\right], \\
c(\mathbf{x},\mathbf{y}) &:= \mathbb{E}_{\mathbb{P}}\left[\left(g(\mathbf{x},\omega) - m(\mathbf{x})\right)\left(g(\mathbf{y},\omega) - m(\mathbf{y})\right)\right].
\end{aligned}$$

Theorem 6 (**Karhunen–Loève Theorem**. Sullivan [2016], Theorem 11.4)**.** *Sup-*

pose that $g : \mathcal{X} \times \Omega \to \mathbb{R}$ is a stochastic process such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$: (i) $g(\mathbf{x}, \cdot) \in L^2(\Omega; \mathbb{P})$, (ii) $m(\mathbf{x}) = 0$ and (iii) the covariance function $c(\mathbf{x}, \mathbf{y})$ is a continuous function of both $\mathbf{x}$ and $\mathbf{y}$. Then:

$$g(\mathbf{x}, \omega) \quad = \quad \sum_{j=1}^{\infty} Z_j(\omega)\psi_j(\mathbf{x}),$$

where $\{\psi_j(\mathbf{x})\}_{j=1}^{\infty}$ are orthonormal eigenfunctions of the Hilbert-Schmidt operator $\mathcal{C} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ defined as $\mathcal{C}[f] := \int_{\mathcal{X}} c(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \mathrm{d}\mathbf{y}$ and the eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ are non-negative (assumed without loss of generality to be ordered $\lambda_1 > \lambda_2 > \ldots$). The convergence of the series is in $L^2(\Omega; \mathbb{P})$ and uniform among compact families of $\mathbf{x} \in \mathcal{X}$, with:

$$Z_j(\omega) \quad = \quad \int_{\mathcal{X}} g(\mathbf{x}, \omega)\psi_j(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Furthermore, the random variables Z_j are centred, uncorrelated, and have variance λ_j: $\mathbb{E}_{\mathbb{P}}[Z_j] = 0$ and $\mathbb{E}_{\mathbb{P}}[Z_j Z_k] = \lambda_j \delta_{jk}$.

This characterisation can be particularly useful for approximating the stochastic process. First, it orthogonalises the stochastic and deterministic parts of the stochastic process. Furthermore, since we have assumed that the eigenvalues are in decreasing order, a truncation $\sum_{j=1}^{L} Z_j(\omega)\psi_j(\mathbf{x})$ for $L > 0$ of this series is the best L-dimensional approximation of the stochastic process in an $L^2(\Omega; \mathbb{P})$ sense. Such a truncation is therefore the analogue of principal component analysis for stochastic processes. The truncation can also be useful for approximate sampling of a stochastic process. Indeed, all that is required is to sample IID random variables $\{Z_j\}_{j=1}^{L}$. See Huang et al. [2001] for a detailed study.

The Karhunen-Loeve characterisation therefore provides us with a second definition of a GP as the series $g(\mathbf{x}, \omega) := \sum_{j=1}^{\infty} \sqrt{\lambda_j}\epsilon_j \psi_j(\mathbf{x})$, where $\{\epsilon_j\}_{j=1}^{\infty}$ are IID $\mathcal{N}(0, 1)$ random variables and $\{\lambda_j, \psi_j\}_{j=1}^{\infty}$ are the eigenvalues and eigenfunctions of the Hilbert-Schmidt operator.

2.2.3 Connection Between Kernels and Covariance Functions

As hinted at previously, there is a close relationship between reproducing kernels and covariance functions. Consider without loss of generality a stochastic process with $m = 0$ and covariance function c. We say that a stochastic process is a second-order stochastic process if $\mathbb{E}_{\mathbb{P}}[|g(\mathbf{x}, \omega)|^2] < \infty$ for all $\mathbf{x} \in \mathcal{X}$ (i.e. the process has finite second moment). It turns out that reproducing kernels correspond to covariance functions of second order stochastic processes:

Theorem 7 (**Loève's Theorem.** Loève [1978], p132)**.** *A function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance function of a second-order stochastic process if and only if it is positive definite.*

Focusing on the special case of Gaussian processes, we have that for any pair of mean function m and reproducing kernel k, there exists a GP with mean m and covariance k and vice versa; see Theorem 12.1.3 in Dudley [2002].

An important point however is that any realisation of a Gaussian process (or in fact any second-order stochastic process) will usually not lie in the RKHS associated with its kernel/covariance function. Several conditions for these functions to lie in the RKHS are provided in [Driscoll, 1973; Lukić and Beder, 2001; Pillai et al., 2007]. See also the extended discussion in Kanagawa et al. [2018].

2.3 Bayesian Nonparametric Models

We have now concluded our introduction to reproducing kernel Hilbert spaces and stochastic processes, and highlighted their connections. These two research areas are commonly used in Bayesian inference in cases where the parameter belongs to a function space. This subfield of Bayesian inference is commonly called Bayesian nonparametrics[4] [Dey et al., 1998; Müller et al., 2015; Gine and Nickl, 2016; Ghosal and van der Vaart, 2017].

2.3.1 Bayesian Models in Infinite Dimensions

It is often said that priors and posteriors on functions are infinite-dimensional. An intuitive justification for this name is that assigning a distribution on a function is often equivalent to assigning a distribution on an infinite sequence of scalars. Consider the problem of constructing a prior model which is a stochastic process. Assigning such a prior to a function is equivalent to selecting a prior distribution for the stochastic part of the Karhunen-Loeve expansion of the stochastic process; i.e. selecting a prior for the sequence $\{Z_j\}_{j=1}^{\infty}$ in Theorem 6.

Two canonical examples of priors in Bayesian nonparametrics are the Gaussian process and the Dirichlet process (both introduced in the previous section). Recently, infinite-dimensional models have become popular in the literature due to the fact that they place probability mass on a wider range of models. This is not the

[4] This name is particularly misleading since the likelihoods are indeed parametric, but the parameter in this case is a function.

case for parametric models, which place very restrictive assumptions on the data-generating mechanism being modelled, and it that sense correspond to very strong prior knowledge.

When working with Bayesian statistics for infinite-dimensional models, it is necessary to to tread carefully as finite-dimensional intuitions and results do not always carry through. Recall the statement of Bayes' Theorem in Chapter 1, Equation 1.3: $\pi(\theta|\mathbf{X}) = \pi_0(\theta)\pi(\mathbf{X}|\theta)/\pi(\mathbf{X})$. Here, $\theta \in \Theta$ was some parameter of interest in some Euclidean space and this identity assumed the existence of densities with respect to the Lebesgue measure on Θ. However, when Θ is not a subset of some Euclidean space but some function space, this identity cannot hold since there is no infinite-dimensional equivalent of the Lebesgue measure and so these densities do not exist.

Instead, a generalisation of Bayes' theorem in infinite dimensions can be given in terms of Radon-Nikodym derivative of the posterior measure with respect to the prior measure. The reason is that under regularity conditions on the likelihood, the posterior will be absolutely continuous with respect to the prior (see Stuart [2010]; Dashti and Stuart [2016]). In these cases, denote by Π_0 the prior measure and by Π the posterior measure. Bayes' Theorem can be expressed as:

$$\frac{\mathrm{d}\Pi}{\mathrm{d}\Pi_0}(\theta) \quad = \quad \frac{1}{Z(\mathbf{X})}\exp(-\Phi(\theta;\mathbf{X})), \tag{2.5}$$

where Φ is called a potential and encapsulates information from the likelihood. The normalisation constant ensures that the posterior Π is a probability measure:

$$Z(\mathbf{X}) \quad = \quad \int_{\Theta} \exp(-\Phi(\theta;\mathbf{X}))\Pi_0(\mathrm{d}\theta). \tag{2.6}$$

There are several additional challenges when working with priors on infinite-dimensional spaces. First, specifying a prior on a large parameter space is hard. In infinite dimensions, any choice of prior will be mutually singular with respect to infinitely many measures (i.e. these priors put zero mass on a very large class of functions), and so infinite-dimensional priors can be thought of as being "infinitely informative". Eliciting a representative subjective prior is therefore challenging, and establishing objective priors even more so.

An important property for Bayesian inference is posterior consistency. Posterior consistency is the property that the posterior eventually concentrates in a small neighborhood of the true parameter. In the finite-dimensional case, this will hold under very mild conditions as long as the true parameter value is in the support of the prior; however, this will not be the case in infinite-dimensions. Early results for consistency include the work of Doob [1949] and Schwartz [1965]. Several more recent results, both positive and negative, have also been established [Diaconis and Freedman, 1986; Freedman, 1999; Choi and Schervish, 2007; Owhadi et al., 2015]. One should therefore always verify that this property holds on a case-by-case basis.

Due to the fact that infinite-dimensional priors have support in an infinite-dimensional space, they also tend to require more data than their parametric counterparts. General asymptotic consistency rates, also sometimes called contraction rates, have been established in the IID setting by Ghosal et al. [2000]; Shen and Wasserman [2001] and the non-IID setting (including Gaussian time series and Markov processes) by Ghosal and van Der Vaart [2007]. See also Chapter 8 of Ghosal and van der Vaart [2017] for an overview of more recent results.

Finally, computation is challenging when the posterior is not in a well-known family of models and needs to be approximated. For example, choosing good proposals for MCMC in high dimensions is complicated, and the use of standard proposals such as random walks can lead to acceptance rates tending to zero as the dimension of the problem increases. Extensive research is dedicated to the design of algorithms with an acceptance rate which does not degrade as the dimension increases; see for example Beskos et al. [2011, 2017]; Cotter et al. [2013].

Keeping all of the drawbacks above in mind, it is important to point out the main philosophical appeal of the Bayesian methodology. Given a carefully chosen prior, the posterior provides a full characterisation of the uncertainty about the unknown parameter of interest (rather than a simple point estimate as would be available with alternative methodologies).

2.3.2 Gaussian Processes as Bayesian Models

In the next section, we discuss in more detail one of the most popular models in the Bayesian nonparametrics literature: Gaussian processes. These models will be used extensively throughout Chapters 3 and 4. Over the years, GPs have been used in a Bayesian setting on a range of applications. Examples include the field of computer experiments [Kennedy and Hagan, 2001], Machine learning (including as a regression and classification model) [Rasmussen and Williams, 2006], Bayesian inverse problems [Stuart, 2010; Dashti and Stuart, 2016] and Bayesian numerical methods [Larkin, 1972; Diaconis, 1988; O'Hagan, 1992].

GPs have been particularly popular models due to their conjugacy property: under the assumption of exact function evaluations or function evaluations with Gaussian noise, the posterior resulting from a GP is also a GP. More precisely,

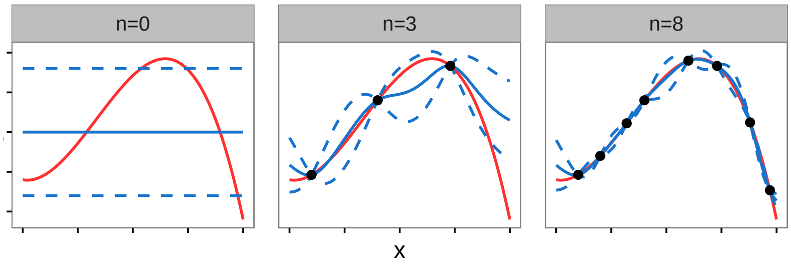

Figure 2.1: *Sketch of a Gaussian process prior and posterior.* The one-dimensional function f (red) is increasingly well approximated by the posterior mean m_n (blue) as the number n of function evaluations is increased. The dashed lines represent pointwise 95% posterior credible intervals.

denote by $g : \mathcal{X} \times \Omega \to \mathbb{R}$ a GP prior on some function $f : \mathcal{X} \to \mathbb{R}$, with mean $m : \mathcal{X} \to \mathbb{R}$ and covariance $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Denote by $\mathbf{f} \in \mathbb{R}^n$ the vector of values $f_i = f(\mathbf{x}_i)$, $\mathbf{m} \in \mathbb{R}^n$ the vector of values $m_i = m(\mathbf{x}_i)$ for some data points in the set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$. Furthermore, let $\mathbf{c}(\mathbf{x}, \mathbf{X}) = \mathbf{c}(\mathbf{X}, \mathbf{x})^\top$ denote the $1 \times n$ vector whose ith entry is $c(\mathbf{x}, \mathbf{x}_i)$ and $\mathbf{C}$ for the matrix with entries $(\mathbf{C})_{i,j} = c(\mathbf{x}_i, \mathbf{x}_j)$. After conditioning the GP prior g on some data $\mathbf{X}$ observed with IID noise which is $\mathcal{N}(0, \sigma^2)$ distributed, the posterior $g_n : \mathcal{X} \times \Omega \to \mathbb{R}$ is a GP with mean $m_n : \mathcal{X} \to \mathbb{R}$ and covariance $c_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by ([Rasmussen and Williams, 2006, Chapter 2]):

$$\begin{aligned} m_n(\mathbf{x}) &= m(\mathbf{x}) + \mathbf{c}(\mathbf{x}, \mathbf{X})(\mathbf{C} + \sigma^2 I_{n\times n})^{-1}(\mathbf{f} - \mathbf{m}), & (2.7) \\ c_n(\mathbf{x}, \mathbf{x}') &= c(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x}, \mathbf{X})(\mathbf{C} + \sigma^2 I_{n\times n})^{-1}\mathbf{c}(\mathbf{X}, \mathbf{x}'), & (2.8) \end{aligned}$$

where $I_{n\times n}$ is the identity matrix of dimension $n \times n$. The formal definition of g_n as a posterior model is tricky in the noiseless case since defining a likelihood is not straightforward. However, we can formally see g_n as a conditioned stochastic process to avoid technical obfuscation. A sketch of the conditioning procedure is provided in Figure 2.1.

The expression for the posterior given above only considers function evaluations, but our observation model could be more complex. In fact, the conjugacy property of GPs holds when the data consists of any bounded linear functional of f. Useful example of observations include integrals over parts of the domain $\int_{\mathcal{Y}} f(\mathbf{x}) \mathrm{d}\mathbf{x}$ for $\mathcal{Y} \subseteq \mathcal{X}$, or derivative observations $\nabla_{\mathbf{x}} f(\mathbf{x})$.

Unfortunately, the conjugacy property of GPs can break down in several cases. First, if the hyperparameters γ of the covariance function $c(\mathbf{x}, \mathbf{y}; \gamma)$ are unknown and a hierarchical Bayesian approach is taken (i.e. a prior is specified over hyperparameters), then the problem is usually not conjugate anymore. These problems then require advanced MCMC methods to sample from the posterior [Filippone and Girolami, 2014]. Another example where the conjugacy property is lost is in the case of deep GPs [Damianou and Lawrence, 2013; Dunlop et al., 2017; Monterrubio-Gómez et al., 2018], where the covariance function itself is modelled as a realisation from a GP (up to several levels).

2.3.3 Practical Issues with Gaussian Processes

GPs will be used extensively throughout the remainder of this thesis, and we therefore pause to discuss issues relating to their practical implementation. This includes how to select a particular type of GP prior for Bayesian inference, stability of the numerical systems underlying conditional distributions, and issues relating to their scalability in high dimensional or large data settings.

Prior Specification

Prior specification (also called model selection) is an important consideration for working with GPs [Stein, 1999; Xu and Stein, 2017]. It consists of selecting the mean function $m : \mathcal{X} \to \mathbb{R}$ and the covariance function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of the GP prior; see Oakley [2002] for elicitation of priors in the area of computer experiments. Since prior models are "infinitely informative" in the nonparametric case, this choice will be of prime importance as it will significantly influence the result of the Bayesian analysis. Care is therefore required.

The choice of mean function for GPs has received relatively little attention. This is mainly due to the fact that an appropriate choice of prior mean should be guided by problem-specific knowledge. A common practice is to set the mean function to $m = 0$, then let the data influence the posterior. In cases where n is large and the dimension of the domain $\mathcal{X}$ is low, this may be an acceptable approach. However when this is not the case, the data will not be informative about the function on the entire domain and, as a result, the posterior will revert to the prior in areas which are unexplored. An arbitrary choice of prior such as $m = 0$ can therefore have severe consequences in these cases. To avoid this problem, it is also possible to use a parametric model as prior mean, for example using a linear combination of basis functions [Kennedy and Hagan, 2001], or use meta-learning; see for example Fortuin and Ratsch [2019].

A problem which has received significantly more attention is the choice of

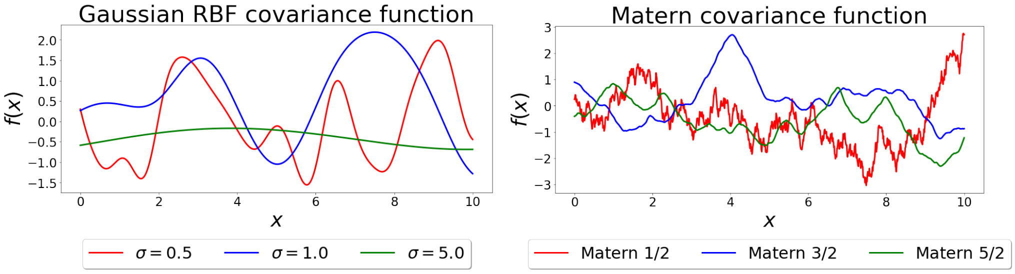

Figure 2.2: *Importance of model selection for Gaussian processes. Left:* Draws from a Gaussian Process prior with mean zero and covariance a Gaussian RBF kernel with lengthscale $\sigma = 0.1$ (red), $\sigma = 1$ (blue) and $\sigma = 5$ (green). *Right:*

covariance function. This is because the covariance function will determine essential properties of the realisations and mean of the posterior. Some popular covariance functions have already been introduced in Section 2.1.3 and other examples can also be found in [Duvenaud, 2014]. It is common to base this choice on smoothness, periodicity and tail properties.

As was seen in these examples, covariance functions also tend to have several hyperparameters (jointly denoted by the vector γ) which need to be selected, and will have a significant influence on the prior obtained. This is for example illustrated in Figure 2.2 (left), where realisations from a GP with Gaussian RBF covariance function (see Equation 2.4) are plotted for various values of the lengthscale σ but fixed amplitude $\lambda = 1$. Similarly, Figure 2.2 (right) contains realisations from a GP with Matérn covariance (see Equation 2.3) with lengthscale $\sigma = 1$ and amplitude $\lambda = 1$ but varying smoothness hyperparameter ν. In both case, the hyperparameters have a significant impact on the realisations obtained.

Consider a parametric covariance function $c(\mathbf{x}, \mathbf{x}'; \gamma_l, \gamma_s)$, with a distinction drawn here between scale hyperparameters γ_l and smoothness hyperparameters γ_s. The former are defined as parameterising the norm of the associated RKHS, whereas the latter affect the corresponding RKHS itself. Selection of γ_l, γ_s based on data can only be successful in the absence of acute sensitivity to these hyperparameters. For scale hyperparameters, a wide body of evidence demonstrates that this is usually not a concern [Stein, 1999]. We now outline several approaches, which are described in more details by Rasmussen and Williams [2006]:

- **Marginalisation**: A natural approach, from a Bayesian perspective, is to set a prior on the hyperparameters γ and then to marginalise over the posterior distribution on these parameters. Recent results for certain infinitely differentiable covariance functions establish minimax optimal rates for this approach, including in the practically relevant setting where π is supported on a low-dimensional sub-manifold of the ambient space $\mathcal{X}$ [Yang and Dunson, 2016]. However, the act of marginalisation itself involves an intractable integral which will usually break the conjugacy property of GPs. It is therefore important to keep in mind the additional computational resources required when assessing the advantages provided by marginalisation.

- **Cross-Validation**: Another approach to the choice of covariance function is cross-validation. It consists of separating the data into $M \in \mathbb{N}$ subsets then, for a given hyperparameter value, conditioning the GP on $M-1$ subsets and assessing its predictive performance using the data points in the last subset. The procedure is then repeated over all choices of $M-1$ subsets, to obtain an indication of how good the hyperparameter value is for prediction. This procedure can then be repeated for several hyperparameter values, and the best performing hyperparameter is retained.

  Clearly, this method will be a robust approach to selecting hyperparameters since it is less prone to suffer from outliers. However, it can be considered to be less principled than marginalisation from a Bayesian point of view since it selects a prior using the data. Another issue is that it can perform poorly when the number n of data points is small, since the data needs to be further reduced into M subsets. The performance estimates are known to have large variance in those cases.

- **Empirical Bayes**: An alternative to the above approaches is empirical Bayes. This consists in selecting hyperparameters γ to maximise the log-marginal likelihood of the data $\{f(\mathbf{x}_i)\}_{i=1}^n$:

  $$l(\gamma) \quad = \quad -\frac{1}{2}\mathbf{f}^\top \mathbf{C}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{C}| - \frac{n}{2}\log 2\pi,$$

  where $|\mathbf{C}|$ denotes the determinant of the matrix $\mathbf{C}$. In practice, this objective can be maximised using any numerical optimisation routine. Empirical Bayes has the advantage of providing an objective function that is easier to optimise relative to cross-validation but it is not fully Bayesian since it also makes use of the data to select the hyperparameters. Empirical Bayes can lead to over-confidence when n is very small, since the full irregularity of the function has yet to be uncovered [Szabó et al., 2015]. In addition, it can be shown that empirical Bayes estimates need not converge as $n \to \infty$. This is for example the case when the GP is supported on infinitely differentiable functions [Xu and Stein, 2017].

Selection of smoothness hyperparameters is a much harder problem and an active area of theoretical research; see Szabó et al. [2015]. In some cases it is possible to elicit a smoothness hyperparameter from physical or mathematical considerations, such as a known number of derivatives of the function. Alternatively, the three methods highlighted above can also be used for smoothness hyperparameters but are much less well understood in this case.

Stability of the Numerical System

The main computational challenge associated with the use of GPs is inverting the $n \times n$ Gram matrix $\mathbf{C}$. This is required in order to obtain the posterior mean and variance in Equation 2.7 and 2.8. When n is large, or in unfavourable hyperparameter regimes, the inverse of the covariance matrix can become numerically unstable. Understanding when this may happen is of great practical importance, and we refer the reader to Chapter 12 of Wendland [2005] for a detailled discussion.

Consider Figure 2.3 where we highlight this problem for the simple case of GP regression with Gaussian RBF covariance function where the function is evaluated at 100 equidistant points on $[0, 10]$. When the covariance function has a large lengthscale σ, the matrix is ill-conditioned since neighboring rows or columns are very similar to one another. This may not be an issue from a theoretical viewpoint, but it is likely that the matrix will become numerically singular. Schaback and Wendland [2006] point out that this behaviour occurs for a large class of radial kernels. Another observation in this paper is that the conditioning of the Gram matrix will worsen with the smoothness of the covariance function.

Often it is the case that we need to compute the product $\mathbf{C}^{-1}\mathbf{b}$ where $\mathbf{b}$ is a vector of length n. In this case, first solving the linear system $\mathbf{b} = \mathbf{C}\mathbf{a}$ for $\mathbf{a}$, then computing the matrix-vector product tends to be more numerically stable than computing the matrix inverse directly.

Several approaches to further improve stability include multipole expansions [Greegard and Strain, 1991], domain decomposition methods [Beatson et al., 2001], partition of unity methods [Babuska and Melenk, 1997], compactly supported kernels [Floater and Iske, 1996; Wendland, 2005] and preconditioning of the covariance matrix [Mouat, 2001].

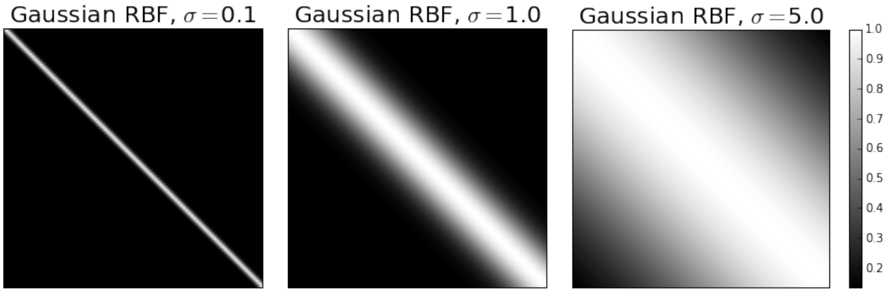

Figure 2.3: *Ill-conditioning of the Gram matrix in Gaussian process regression.* We continue the example in Figure 2.2 and plot the Gram matrices corresponding to 100 equidistant points in $[0, 10]$ for a GP with Gaussian RBF kernel with amplitude hyperparameter $\lambda = 1$ and lengthscale hyperparameter $\sigma = 0.1$ (left) $\sigma = 1$ (middle) and $\sigma = 5$ (right).

Scalability

In situations where obtaining data is cheap, the naive $O(n^3)$ computational cost associated with inverting the covariance matrix renders GP regression slow. It is then natural to ask whether the uncertainty quantification provided by GPs is worth the increased off-line computational overhead. Below, several approaches to reducing the computational overhead of GPs are highlighted.

Exact inversion can be achieved at low cost through exploiting structure in the kernel matrix. Examples include: tensor product kernels [O'Hagan, 1991], circulant embeddings [Davies and Bryant, 2013] and low-rank kernels such as polynomial kernels. In addition there are many approximate inversion techniques. We highlight a few below: reduced rank approximations [Quinonero-Candela and Rasmussen, 2005; Bach, 2013; El Alaoui and Mahoney, 2015], explicit feature maps designed for additive kernels [Vedaldi and Zisserman, 2012], local approximations [Gramacy and Apley, 2015], multi-scale approximations [Iske, 2004; Katzfuss, 2017], random approximations of the kernel itself, such as random Fourier features [Rahimi and Recht, 2007], spectral methods [Lazaro-Gredilla et al., 2010; Bach, 2017], hash kernels [Shi et al., 2009], parallel programming [Dai et al., 2014] and efficient use of data structures [Wendland, 2005][Section 14].

Furthermore, several approach to improve conditioning of the linear system discussed in the previous also reduce the computational cost as a by-product. These include the fast multipole methods and compactly supported covariance functions.

This, of course, does not represent an exhaustive list of the (growing) literature on kernel matrix methods. Note that the majority of approximate kernel methods do not come with probability models for the additional source of numerical error introduced by the approximation.

Chapter 3

Bayesian Numerical Integration: Foundations

"We believe that they demonstrate very strongly that, in a fundamental sense, Monte Carlo is statistically unsound."

[O'Hagan, 1984]

Our objective in this thesis will be to make use of the theory of reproducing kernel Hilbert spaces to tackle challenges in computational statistics.

The first challenge that was previously highlighted is the numerical integration of expensive functions, and this will be the focus of the current chapter. In particular, we will review an existing algorithm called Bayesian quadrature (BQ). We will begin with a brief overview of the field of Bayesian probabilistic numerical methods, then move on to the analysis of BQ. This will include consistency rates and numerical experiments on a wide range of statistical applications. In Chapter 4, we will then propose novel extensions of the algorithm.

3.1 Bayesian Probabilistic Numerical Methods

3.1.1 Numerical Analysis in Statistics and Beyond

Numerical analysis is a subfield of mathematics extensively used throughout applications across the sciences (and beyond). There are several competing definitions of the field, but it is often described by researchers as "the study of algorithms for the problems of continuous mathematics"; see Trefethen [1992]. More precisely, numerical analysis is concerned with how to best project continuous problems into discrete scales. Canonical examples include approximating the solution of integral equations [Davis and Rabinowitz, 2007], differential equations [Hairer et al., 1993; Hairer and Wanner, 1996], or even the solution of problems in interpolation [Wahba, 1991; Wendland, 2005], optimisation [Boyd and Vandenberghe, 2004; Nocedal and Wright, 2006] and linear algebra [Trefethen and Bau, 1997].

In each of these cases, a continuous mathematical quantity, such as a function or an operator, is discretised in a way such that the solution to the discretised problem can be computed in closed form by a computer. Discretising the quantity of interest in different manners leads to different algorithms, and the approximation properties of each discretisation scheme is, of course, of great importance. For this reason, a second less glamorous, yet popular, definition of numerical analysis is the "study of numerical errors".

A standard approach to developing a new algorithm in numerical analysis goes as follows. First, it is important to check that the problem at hand is well-posed and that the proposed algorithm is numerically stable. That is, we want to verify that the method does not magnify approximation errors. A second step consists of studying the convergence of the algorithm and the associated order of this convergence in the size of the discretisation grid or mesh. This is done by defining some notion of error and studying how this error decreases as the number of iterations increases. These types of errors are usually chosen to be worst-case errors over some set of assumptions on the problem, and therefore provide rather conservative bounds on the error incurred by the use of a given algorithm; see Chapter 1 for a brief discussion.

Numerical methods are essential to mathematical modelling in many applied settings, as well as specifically within statistics and machine learning. Take for example the field of Bayesian statistics: the biggest challenge here (as described in the previous chapters) is the approximation of expensive or high dimensional integrals which are required to obtain a posterior distribution on quantities of interest. At this stage, numerical methods are usually considered by practitioners as computational black-boxes that return a point estimate for the integral, and whose numerical error is then neglected. This means that the posterior distribution on the quantity of interest will not account for the numerical error. Numerical integration is thus one part of Bayesian inference for which uncertainty is not routinely accounted for in a fully Bayesian way.

This lack of Bayesian uncertainty quantification should of course be alarming to Bayesian statisticians. Can one really trust a posterior distribution which was obtained by approximating an integral or differential equation? How would the posterior differ if this additional source of uncertainty was incorporated? These questions are often ignored, but in many cases should probably not be.

3.1.2 Numerical Methods as Bayesian Inference Problems

The description of numerical analysis as the discretisation of a continuous quantity should sound familiar to statisticians. It is in fact very much related to statistical inference, where we are interested in inferring some unknown quantity by observing a finite number of values, and in studying asymptotic properties of associated estimators. In this case, the unknown quantity would be the solution of the continuous mathematical problem (e.g. some intractable integral), and the data consists of functionals of some underlying function (e.g. integrand evaluations).

As a Bayesian, a natural procedure would therefore be to think of the unknown quantity as a random variable and specify a prior over it, then update one's beliefs using the observations available. This is the approach proposed by Bayesian probabilistic numerical methods, which originate in the work of Poincaré [1896] and were independently proposed by a number of eminent mathematicians and statisticians in the 1970s, 1980s and 1990s [Larkin, 1972; Diaconis, 1988; O'Hagan, 1992; Kadane and Wasilkowski, 1985; Skilling, 1991], and most recently reviewed in Hennig et al. [2015]; Briol et al. [2015b] and Cockayne et al. [2017]. See also Oates and Sullivan [2019] for a review of the early history of the field.

The Bayesian approach to numerical analysis did not see many significant developments between the 1990s and 2010s. As Diaconis [1988] puts it in the late 1980s: "most people, even Bayesians, think this sounds crazy when they first hear about it". Indeed, back then the idea of using Bayesian statistics to solve numerical analysis problems was unusual and the advantages against classical methods (which had already been developed for decades) were unclear. To some extent, although Bayesian methods are now more widely accepted, the criticism remains valid. This question will be studied throughout this chapter.

Is it really useful to formulate numerical problem from the Bayesian viewpoint, or are we just reformulating known algorithms in the Bayesian language? This thesis argues that there is much more to Bayesian numerical methods than a change of vocabulary. A first advantage of specifying a prior distribution is that we are making all of our assumptions on the quantity of interest, or any subject of computation, explicit. The prior also allows the user to add additional information which does not fit into any of the existing methods. For example, BQ is very flexible and can incorporate a wide range of prior knowledge on the integrand through selection of the mean and covariance function of a GP. We might know that the integrand is periodic or monotonic, perhaps through inspection of the functional form of f or via domain-specific knowledge, and this can directly be encoded in the algorithm.

Second of all, Bayesian statistics is a principled way of performing uncertainty quantification [Robert, 1994]. The recent work by Cockayne et al. [2017] outlines how many Bayesian probabilistic numerical methods can be framed as Bayesian inverse problems [Stuart, 2010; Dashti and Stuart, 2016]. Instead of point estimates for the quantities of interest, these methods can in fact provide entire Bayesian posteriors on these quantities. These posteriors should provide much more faithful representations of our uncertainty than the classic worst-case error bounds. Note that the notions of uncertainty and error discussed here are very different to those used in numerical analysis. We are talking about epistemic uncertainty (representing our personal lack of knowledge about a problem) rather than aleatoric uncertainty (which concerns inherent randomness in a system).

Thirdly, the Bayesian approach is also useful in cases where numerical methods are used in a sequential manner. Of course, in many situations numerical error will be negligible and no further action is required, but if numerical errors are propagated through a computational pipeline and allowed to accumulate, then failure to properly account for such errors could potentially have drastic consequences on subsequent statistical inferences. Such consequences could be akin to the Lorenz's butterfly effect in chaos theory, where small changes to the initial state of a system could have large consequences on later states. See Mosbach and Turner [2009] for an example of numerical error accumulating when solving differential equations, and Oates et al. [2017b] for a large-scale application of Bayesian probabilistic numerical methods to a problem in electrical impedance tomography.

Finally, the Bayesian framework allows us to frame numerical problems in the setting of transfer learning, where we re-use the computations performed for a first numerical problem to improve the performance when solving a second numerical problem. This will be illustrated in the case where we have multiple integrals of interest $\Pi[f_1], \dots, \Pi[f_P]$ ($P \in \mathbb{N}$) in Chapter 4, and we will demonstrate how knowledge of the correlation structure between $f_1, \dots, f_P$ can be used to improve the estimate of each of these integrals. This setting is not usually considered in numerical analysis, but arises naturally from the statistical formulation.

3.1.3 Recent Developments in Bayesian Numerical Methods

Since the early 1980s, a range of Bayesian probabilistic numerical methods have been invented and developed to address most canonical problems in numerical analysis.

Of course, there are many statistical methods for functional approximation; see for example GPs and Dirichlet processes as introduced in the previous chapter. We now provide a brief overview for other canonical problems:

- **Ordinary Differential Equations:** The first method for ordinary differential equations was proposed by Skilling [1991] and an approximate Bayesian framework using GPs was introduced in Chkrebtii et al. [2016]. Raissi et al. [2017] later proposed a version using multifidelity GPs.

  Hennig and Hauberg [2014]; Schober et al. [2014] provided a probabilistic version of Runge-Kutta methods, which takes the form of a filtering model which was further studied in Schober et al. [2018]. Various priors were later explored in [Magnani et al., 2017], and Kersting and Hennig [2016] explored the close link between the solutions of the ordinary differential equations and quadrature. Similar algorithms with a filtering flavour were also proposed by Teymur et al. [2016, 2018] in the case of multi-step methods.

  In a separate line of research, Conrad et al. [2017] proposed an uncertainty quantification framework which proceeds by introducing random noise at each step of any existing numerical solver. This was further studied from a theoretical point of view in Lie et al. [2017] and extended to a random time-step formulation in Abdulle and Garegnani [2017].

- **Partial Differential Equations**: Extensions of Chkrebtii et al. [2016]; Conrad et al. [2017] were also proposed for partial differential equations. The first independent work for partial differential equations is due to Owhadi [2015], who framed the problem of numerical homogenisation as a Bayesian inference problem. Later, [Cockayne et al., 2016; Oates et al., 2017b] developed probabilistic meshless methods which allow for uncertainty quantification within the popular stochastic collocation methods. Finally, Owhadi [2017]; Owhadi and Zhang [2017]; Owhadi and Scovel [2017] developed gamblets, a computationally efficient approach in the case of hierarchical information. More generally, [Owhadi and Scovel, 2017] also provided an in-depth discussion of the link between methods which are optimal in Bayesian and game-theoretic settings.

- **Optimisation:** By far the most popular Bayesian numerical method is called Bayesian optimisation [Mockus, 1989]. It is widely applicable and used extensively throughout machine learning [Snoek et al., 2012], both in academic research and throughout industry. Further work includes Hennig and Kiefel [2013], who provided a probabilistic perspective on Quasi-Newton methods, Mahsereci and Hennig [2015], who introduced a probabilistic line search algorithm, and Wills and Schön [2017], who combined both of the methods above.

- **Linear Algebra:** So far, less work has been done at the intersection of linear algebra and Bayesian statistics. Hennig [2015] has proposed a probabilistic interpretation for certain solvers of unconstrained linear systems, whilst Bartels and Hennig [2016] proposed an extension for the specific case of least-squares. Fitzsimons et al. [2017] proposed a Bayesian approach to inferring log-determinants. Finally, Cockayne et al. [2018] proposed a Bayesian version of the conjugate gradient method.

Of course, one of the canonical problems which has not been discussed so far is numerical integration. This will be the focus of the remainder of this chapter. We conclude this section with a remark on the definition of Bayesian numerical methods. Although most of the methods above claim to be Bayesian, they do not all satisfy some of the main Bayesian principles, such as propagation of uncertainty by conditioning and marginalisation. A formal definition of a Bayesian probabilistic numerical method was proposed by Cockayne et al. [2017], and a discussion of which methods satisfy this definition can be found in Table 1 in that paper.

Although some of these methods are not fully Bayesian, they may be considered as being approximately Bayesian. The goal of such methods is then to strike a good balance between the useful properties of Bayesian methods and the computational challenges and other practicalities surrounding implementation.

3.2 Bayesian Quadrature

For the remainder of this chapter, we study an algorithm called Bayesian Quadrature (BQ) [O'Hagan, 1991] which proposes a Bayesian approach to numerical integration. In this section, we introduce BQ and relate it to the study of quadrature rules in RKHSs. We then provide theoretical results for several variants of BQ in Section 3.3, before discussing details of importance for its efficient implementation in Section 3.4. Finally, we study its performance on several problems in statistics and engineering in Sections 3.5 and 3.6

3.2.1 Introduction to Bayesian Quadrature

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space with $\mathcal{X} \subset \mathbb{R}^d$ for $d \in \mathbb{N}$ and let $\mathcal{B}(\mathcal{X})$ be a Borel σ-algebra. Let $f : \mathcal{X} \to \mathbb{R}$ be some function for which we would like to compute the integral $\Pi[f]$.

Recall that a quadrature rule describes any functional of the form of a linear combination of function values: $\hat{\Pi}[f] = \sum_{i=1}^n w_i f(\mathbf{x}_i)$ for some states (or samples) $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n \subset \mathbb{R}$. The notation $\hat{\Pi}[f]$ is motivated by the fact that this expression can be re-written as the integral of f with respect to an empirical measure $\hat{\Pi} = \sum_{i=1}^n w_i \delta(\mathbf{x}_i)$, where $\delta(\mathbf{x}_i)$ is a Dirac measure (i.e. for all $A \in \mathcal{B}(\mathcal{X})$, $\delta_{\mathbf{x}_i}(A) = 1$ if $\mathbf{x}_i \in A$, $\delta_{\mathbf{x}_i}(A) = 0$ if $\mathbf{x}_i \notin A$). The weights w_i can be negative and need not satisfy $\sum_{i=1}^n w_i = 1$.

BQ begins by defining a stochastic process $g : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$ formally seen as a prior model for the integrand f. The most popular choice, originally made by Larkin [1972], is to consider a GP, but others could also be used. Recall from Chapter 2 that a GP can be characterised by its mean function and its covariance function: $m(\mathbf{x}) = \mathbb{E}_{\mathbb{P}}[g(\mathbf{x}, \omega)]$ and $c(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\mathbb{P}}[(g(\mathbf{x}, \omega) - m(\mathbf{x}))(g(\mathbf{x}', \omega) - m(\mathbf{x}'))]$. From now on, we assume without loss of generality that $m \equiv 0$. Conditioning the GP at quadrature points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ gives a new GP denoted $g_n : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$. This GP has mean $m_n(\mathbf{x}) = m(\mathbf{x}) + \mathbf{c}(\mathbf{x}, \mathbf{X})\mathbf{C}^{-1}(\mathbf{f} - \mathbf{m})$ and covariance function $c_n(\mathbf{x}, \mathbf{x}') = c(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x}, \mathbf{X})\mathbf{C}^{-1}\mathbf{c}(\mathbf{X}, \mathbf{x}')$. For simplicity we will assume there is no measurement error. After obtaining the conditioned GP g_n, the final step is to produce a distribution on the value of the integral $\Pi[g_n]$ by considering the pushforward of the process g_n through the integration operator. A sketch of the procedure is presented in Figure 3.1 and the relevant formulae are now provided.

Proposition 1 (BQ posterior distribution on the solution of the integral). *The distribution of* $\Pi[g_n]$ *is Gaussian with mean and variance*[1]

$$\begin{aligned} \mathbb{E}[\Pi[g_n]] &= \Pi[\mathbf{c}(\cdot, \mathbf{X})]\mathbf{C}^{-1}\mathbf{f}, & (3.1) \\ \mathbb{V}[\Pi[g_n]] &= \Pi\Pi[c(\cdot, \cdot)] - \Pi[\mathbf{c}(\cdot, \mathbf{X})]\mathbf{C}^{-1}\Pi[\mathbf{c}(\mathbf{X}, \cdot)]. & (3.2) \end{aligned}$$

All of the proofs in this thesis can be found in Appendix B, ordered by Chapter and in the order in which they appear in the main text. In particular, see B.1 for all the proofs in this chapter.

Here, $\Pi\Pi[c(\cdot, \cdot)]$ denotes the integral of c with respect to each argument. It can be seen that the computational cost of obtaining this full posterior (in the worst-case $O(n^3)$) is much higher than that of obtaining a point estimate for the integral using MC methods. However, many methods for scaling GPs (discussed in the previous chapter) can be used to speed this up. Karvonen and Särkkä [2018] also proposed a novel scalable method specifically targeted to scaling BQ.

[1] The mean and variance are taken with respect to $\mathbb{P}$, but we do not repeatedly specify this to avoid overloading the notation.

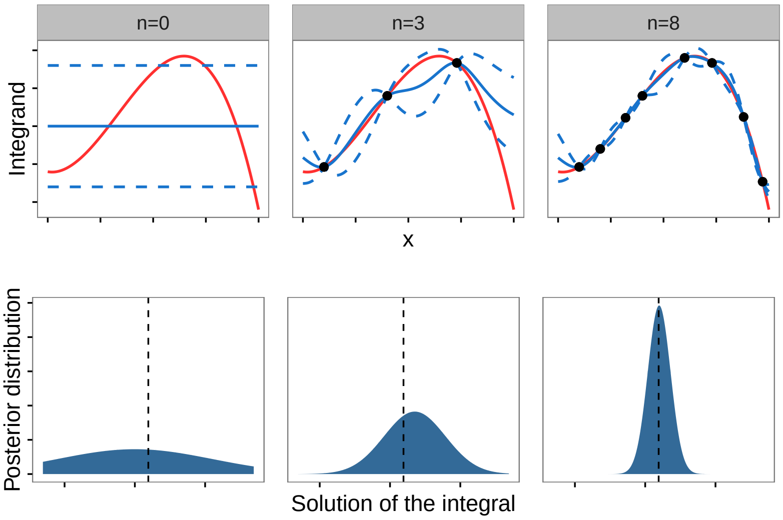

Figure 3.1: *Sketch of Bayesian quadrature.* The top row shows the approximation of the integrand f (red) by the posterior mean m_n (blue) as the number n of function evaluations is increased. The dashed lines represent point-wise 95% posterior credible intervals. The bottom row shows the Gaussian distribution with mean $\mathbb{E}[\Pi[g_n]]$ and variance $\mathbb{V}[\Pi[g_n]]$ and the dashed black line gives the true value of the integral $\Pi[f]$.

Since BQ formally associates g with a prior on f, $\Pi[g_n]$ in turn provides a posterior distribution over the value of the integral $\Pi[f]$ representing our epistemic uncertainty. An interesting remark is that Equation 3.1 takes the form of a quadrature rule:

$$\mathbb{E}[\Pi[g_n]] \;=\; \hat{\Pi}_{\mathrm{BQ}}[f] \;:=\; \sum_{i=1}^{n} w_i^{\mathrm{BQ}} f(\mathbf{x}_i), \tag{3.3}$$

with weight vector given by $\mathbf{w}^{\mathrm{BQ}} := (\Pi[\mathbf{c}(\mathbf{X},\cdot)]\mathbf{C}^{-1})^{\top}$. Furthermore, the posterior variance in Equation 3.2 does not depend on function values $\{f(\mathbf{x}_i)\}_{i=1}^{n}$, but only on the location of the states $\{\mathbf{x}_i\}_{i=1}^{n}$ and the choice of covariance function c. This is useful as it allows state locations and weights to be precomputed and reused. However, it also means that the variance is completely driven by the choice of prior. A valid quantification of uncertainty thus relies on a well-specified prior; we consider this issue further in Section 3.4[2].

The BQ mean (Equation 3.1) coincides with classical quadrature rules for specific choices of covariance function c. For example, in one dimension a Brownian covariance function $c(x,x') = \min(x,x')$ leads to a posterior mean m_n that is a

[2] Note that other choices of priors for f will give posteriors which do not necessarily have this property.

piecewise linear interpolant of f between the states $\{x_i\}_{i=1}^n$, i.e. the trapezium rule [Suldin, 1959]. Similarly, Särkkä et al. [2016] constructed a covariance function c for which Gauss-Hermite quadrature is recovered, and Karvonen and Särkkä [2017] showed how other polynomial-based quadrature rules can be recovered. In another research direction, Karvonen et al. [2018] showed how it is possible to design a BQ rule whose mean corresponds to the point estimate of any cubature rule.

Clearly the point estimator in Equation 3.3 is a natural object; it has also received attention in both the kernel quadrature literature [Sommariva and Vianello, 2006] and empirical interpolation literature [Kristoffersen, 2013]. In those contexts, the point estimator is derived from different assumptions on the integrand: namely, that it is an element of a RKHS with kernel c, rather than a draw from a GP with covariance c.

Although other stochastic processes could of course be used as priors [Cockayne et al., 2017], GPs are popular due to their conjugacy properties, and the terminology Bayesian quadrature usually refers to this case. Note that other names for BQ with GP priors include Gaussian-process quadrature or kernel quadrature. Alternative prior which are conjugate include Student-t process, and these could afford heavier tails for values assumed by the integrand.

There has been a wide range of applications of BQ, including to other numerical methods in optimisation, linear algebra and functional approximation [Kersting and Hennig, 2016; Fitzsimons et al., 2017], inference in complex computer models [Oates et al., 2017d], and problems in econometrics [Oettershagen, 2017] and computer graphics [Brouillat et al., 2009; Marques et al., 2013; Briol et al., 2015b; Xi et al., 2018].

3.2.2 Quadrature Rules in Reproducing Kernel Hilbert Spaces

Next we review how analysis of the approximation properties of the quadrature rule $\hat{\Pi}_{\mathrm{BQ}}[f]$ can be carried out in terms of functional approximation in some RKHS. Denote by $\mathcal{H}_k$ a RKHS with some kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Furthermore, denote its inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and associated norm $\| \cdot \|_{\mathcal{H}_k}$. In the remainder of this thesis all kernels k are assumed to satisfy $\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}) \Pi(\mathrm{d}\mathbf{x}) < \infty$. In particular this guarantees $\int_{\mathcal{X}} f(\mathbf{x})^2 \Pi(\mathrm{d}\mathbf{x}) < \infty$ for all $f \in \mathcal{H}_k$. An important object in the study of quadrature rules is the kernel mean $\mu(\Pi) : \mathcal{X} \to \mathbb{R}$, defined as

$$\mu(\Pi)(\mathbf{x}) \quad := \quad \Pi[k(\cdot, \mathbf{x})]. \tag{3.4}$$

The kernel mean is an element of the RKHS $\mathcal{H}_k$ as a consequence of $\int_{\mathcal{X}} k(\mathbf{x},\mathbf{x})\Pi(\mathrm{d}\mathbf{x}) < \infty$ [Smola et al., 2007]. The kernel mean is often also called the representer of integration, which is justified by the fact that $\forall f \in \mathcal{H}_k$:

$$\begin{aligned}\Pi[f] \;=\; \int_{\mathcal{X}} f(\mathbf{x})\Pi(\mathrm{d}\mathbf{x}) \;&=\; \int_{\mathcal{X}} \langle f, k(\cdot,\mathbf{x})\rangle_{\mathcal{H}_k}\Pi(\mathrm{d}\mathbf{x}) \\ &=\; \left\langle f, \int_{\mathcal{X}} k(\cdot,\mathbf{x})\Pi(\mathrm{d}\mathbf{x})\right\rangle_{\mathcal{H}_k} \;=\; \langle f, \mu(\Pi)\rangle_{\mathcal{H}_k}.\end{aligned}$$

since the integral and inner product commute due to the existence of $\mu(\Pi)$ as a Bochner integral [Steinwart and Christmann, 2008, p510].

The main reason that RKHSs are popular in the study of quadrature rule is the reproducing property, which permits an elegant theoretical analysis with many quantities of interest, such as worst-case and average-case errors, becoming tractable. In the language of kernel means, quadrature rules of the form $\hat{\Pi}[f] = \sum_{i=1}^n w_i f(\mathbf{x}_i)$ can be written as $\hat{\Pi}[f] = \langle f, \mu(\hat{\Pi})\rangle_{\mathcal{H}_k}$ where $\mu(\hat{\Pi})$ is the approximation to the kernel mean given by $\mu(\hat{\Pi})(\mathbf{x}) = \hat{\Pi}[k(\cdot,\mathbf{x})]$ (or equivalently, it is the kernel mean with respect to the empirical measure $\hat{\Pi}$). For fixed $f \in \mathcal{H}_k$, the integration error associated with $\hat{\Pi}[f]$ can be expressed as

$$\hat{\Pi}[f] - \Pi[f] \;=\; \langle f, \mu(\hat{\Pi})\rangle_{\mathcal{H}_k} - \langle f, \mu(\Pi)\rangle_{\mathcal{H}_k} \;=\; \langle f, \mu(\hat{\Pi}) - \mu(\Pi)\rangle_{\mathcal{H}_k}.$$

A tight upper bound for the error is obtained by the Cauchy-Schwarz inequality:

$$\left|\hat{\Pi}[f] - \Pi[f]\right| \;\leq\; \|f\|_{\mathcal{H}_k} \left\|\mu(\hat{\Pi}) - \mu(\Pi)\right\|_{\mathcal{H}_k}. \tag{3.5}$$

The expression above, sometimes called the Koksma-Hlawka inequality [Hickernell, 1998], decouples the magnitude in $\mathcal{H}_k$ of the integrand f from the kernel mean approximation error. The first term in this bound is a constant on which we have no control since it depends on the integrand f. However, since the second term does not depend on f, it is common to design quadrature rules to minimise it, as this will lead to an integration error which is small for all functions in $\mathcal{H}_k$. The following sections discuss how quadrature rules can be tailored to target this term.

3.2.3 Optimality of Bayesian Quadrature Weights

An interesting well-known fact is that the worst-case error (WCE) in the RKHS $\mathcal{H}_k$ is characterised as the error in estimating the kernel mean (also called maximum mean discrepancy (MMD) [Gretton et al., 2006]):

Proposition 2 (The WCE in a RKHS corresponds to the MMD).

$$e(\hat{\Pi}; \Pi, \mathcal{H}_k) \quad := \quad \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \Pi[f] - \hat{\Pi}[f] \right| \; = \; \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}_k}.$$

Minimisation of the WCE in $\mathcal{H}_k$ is natural and corresponds to solving a least-squares problem in the feature space induced by the kernel: Let $\mathbf{X}$ denote quadrature points $\{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ denote the vector of quadrature weights, $\mathbf{z} \in \mathbb{R}^n$ be a vector such that $z_i = \mu(\Pi)(\mathbf{x}_i)$, and $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the matrix with entries $(\mathbf{K})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. Combining Proposition 2 with direct calculation gives a tractable formula for the WCE in $\mathcal{H}_k$:

$$\begin{aligned} e(\hat{\Pi}; \Pi, \mathcal{H}_k)^2 \quad &= \quad \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}_k}^2 \\ &= \quad \sum_{i,j=1}^n w_i w_j k(\mathbf{x}_i, \mathbf{x}_j) - 2 \sum_{i=1}^n w_i \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}_i) \Pi(\mathrm{d}\mathbf{x}) \qquad (3.6) \\ &\qquad + \int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') \, \Pi(\mathrm{d}\mathbf{x}) \Pi(\mathrm{d}\mathbf{x}') \\ &= \quad \mathbf{w}^\top \mathbf{K} \mathbf{w} - 2\mathbf{w}^\top \Pi[k(\mathbf{X}, \cdot)] + \Pi\Pi[k(\cdot, \cdot)]. \qquad (3.7) \end{aligned}$$

Several optimality properties for integration in RKHSs were provided in Section 4.2 of Novak and Woźniakowski [2008]. Relevant to this work is that given n evaluations of the function, an optimal estimate in the sense of the WCE in $\mathcal{H}_k$ can, without loss of generality, take the form of a quadrature rule Bakhvalov [1971]. To be more precise, any non-linear and/or adaptive estimator (where the location of function evaluations are chosen adaptively) can be matched in terms of asymptotic WCE in $\mathcal{H}_k$ by a quadrature rule as we have defined. Note that of course, adaptive quadrature may provide superior performance for a single fixed function f, and the minimax result may not be true in general outside the RKHS framework [Novak, 1996].

To relate these ideas to BQ, consider the challenge of deriving an optimal quadrature rule, conditional on fixed states $\{\mathbf{x}_i\}_{i=1}^n$, that minimises the WCE in the RKHS $\mathcal{H}_k$ over weights $\mathbf{w}$. The solution to this convex problem is $\mathbf{w} = \mathbf{K}^{-1}\mathbf{z}$ and is called kernel quadrature in the literature.

Clearly, if the reproducing kernel k is equal to the covariance function c of the GP prior, then the posterior mean from BQ is identical to the optimal quadrature rule in the RKHS [Kadane and Wasilkowski, 1985]. Furthermore, with $k = c$, the BQ posterior variance can be obtained in terms of WCE. In fact the following inequality can be obtained: $\mathbb{V}[\Pi[g_n]] = e(\hat{\Pi}_{\mathrm{BQ}}; \Pi, \mathcal{H}_k)^2$. Regarding optimality, the problem is thus reduced to selection of states $\{\mathbf{x}_i\}_{i=1}^n$.

3.2.4 Selection of States

Optimal Point Sets An optimal Bayesian Quadrature rule would select states to globally minimise the variance $\mathbb{V}[\Pi[g_n]]$, or equivalently the WCE in $\mathcal{H}_k$:

$$\left\{\mathbf{x}_i^{\text{OBQ}}\right\}_{i=1}^n \quad := \quad \underset{\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}}{\arg\min}\ e(\hat{\Pi}_{\text{BQ}}; \Pi, \mathcal{H}_k).$$

Optimal BQ corresponds to classical quadrature rules (e.g. Gauss-Hermite) for specific choices of kernels [Karvonen and Särkkä, 2017]. However it cannot in general be implemented because optimising the states is in general NP-hard [Schölkopf and Smola, 2002, Section 10.2.3].

In earlier work, several approaches have been made for the choice of quadrature points. For example, O'Hagan [1991] considered states $\{\mathbf{x}_i\}_{i=1}^n$ that are employed in Gaussian quadrature methods. Rasmussen and Ghahramani [2002] used MC realisations. Recent work by Gunter et al. [2014]; Briol et al. [2015a] selected states using experimental design to target the variance $\mathbb{V}[\Pi[g_n]]$. These different approaches are now briefly recalled.

Monte Carlo Methods MC, IS, MCMC and QMC (all introduced in Chapter 1) are widely used in statistical computation. Here we pursue the idea of using these algorithms to generate states for BQ, with the aim to exploit BQ to account for the possible impact of numerical integration error on inferences made in statistical applications. In MCMC it is possible that two states $\mathbf{x}_i = \mathbf{x}_j$ are identical. To prevent the covariance matrix $\mathbf{C}$ from becoming singular, duplicate states should be discarded. This is justified since the information contained in function evaluations $f_i = f_j$ is not lost. This does not introduce additional bias into BQ methods, in contrast to MC methods.

We define the following Bayesian estimators, which correspond to BQ algorithms where the integrand is conditioned at MC, IS, MCMC and QMC states:

$$\begin{aligned} \hat{\Pi}_{\text{BMC}}[f] &:= \textstyle\sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{MC}}), & \hat{\Pi}_{\text{BIS}}[f] &:= \textstyle\sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{IS}}), \\ \hat{\Pi}_{\text{BMCMC}}[f] &:= \textstyle\sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{MCMC}}), & \hat{\Pi}_{\text{BQMC}}[f] &:= \textstyle\sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{QMC}}), \end{aligned}$$

where $\{\mathbf{x}_i^{\text{MC}}\}_{i=1}^n$ are IID realisations from Π, $\{\mathbf{x}_i^{\text{IS}}\}_{i=1}^n$ are IID realisations from some importance distribution Π', $\{\mathbf{x}_i^{\text{MCMC}}\}_{i=1}^n$ are samples from a Markov chain with invariant distribution Π and $\{\mathbf{x}_i^{\text{QMC}}\}_{i=1}^n$ is a QMC point set.

This two-step procedure requires no modification to existing sampling methods, and has the advantage that each estimator is associated with a full posterior distribution. We refer to quadrature rules of the form $\hat{\Pi}_{\text{BMC}}[f]$ as Bayesian Monte Carlo (BMC), $\hat{\Pi}_{\text{BIS}}[f]$ as Bayesian importance sampling (BIS), $\hat{\Pi}_{\text{BMCMC}}[f]$ as Bayesian Markov chain Monte Carlo (BMCMC) and $\hat{\Pi}_{\text{BQMC}}[f]$ as Bayesian quasi-Monte Carlo (BQMC). As previously discussed, BMC and BIS were first proposed by Rasmussen and Ghahramani [2002], but to date we are not aware of any previous use of BMCMC, presumably due to analytic intractability of the kernel mean when π is unnormalised. BQMC has been described by Hickernell et al. [2005]; Marques et al. [2013]; Särkkä et al. [2016]. Note that other Monte Carlo sampling methods could also be used. For example, Briol et al. [2017] proposed to combine SMC methods with BQ weights. This will be further discussed in Chapter 4).

Experimental Design Methods An alternative approach to the choice of states comes from the experimental design literature. The simplest example is a greedy algorithm that sequentially minimises $\mathbb{V}[\Pi[g_n]]$. This method, commonly referred to as sequential Bayesian Quadrature [Osborne et al., 2012; Gunter et al., 2014] consists of repeating the following step: $\mathbf{x}_n^{\text{SBQ}} := \arg\max_{\mathbf{x}\in\mathcal{X}} \Pi[c(\cdot,\mathbf{X})]\mathbf{C}^{-1}\Pi[c(\mathbf{X},\cdot)]$ where $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x})^\top$. More sophisticated optimisation algorithms have also been used. For example, Eftang and Stamm [2012] proposed adaptive procedures to iteratively divide the domain of integration into subdomains and Briol et al. [2015a] used conditional gradient algorithms (this will be discussed in detail in Chapter 4). Several gradient-based global optimisation algorithms were also considered in Oettershagen [2017].

3.3 Theoretical Results for Bayesian Quadrature

The role of the following section is to derive convergence rates for BQ algorithms. We begin by discussing a general set of tools which can be used to derive such rates, then focus specifically on the case of $\hat{\Pi}_{\text{MC}}[f]$, $\hat{\Pi}_{\text{MCMC}}[f]$ and $\hat{\Pi}_{\text{QMC}}[f]$. Further results for an experimental design-based BQ rule will later be given in Chapter 4.

The main setting we consider assumes that the true integrand f belongs to a RKHS $\mathcal{H}_k$ and that the GP prior is based on a covariance function c which is identical to the kernel k of $\mathcal{H}_k$. This assumption is of course more natural from a kernel approximation point of view (in which case the algorithm is called kernel quadrature) than for the Bayesian viewpoint. Indeed, it would be more natural for the latter case to assume that f was a realisation from a GP with covariance k; an event which has probability zero of happening in most cases of interest [Driscoll, 1973; Lukić and Beder, 2001] . Further results which build up on our results and consider prior misspecification can be found in [Kanagawa et al., 2016, 2017]. These could be used to provide theoretical guarantees under the more natural Bayesian assumptions.

3.3.1 Convergence and Contraction Rates

The convergence results in this thesis are based on two simple lemmas. The first lemma, shows that probabilistic integrators provide a point estimate that is at least as good as their non-probabilistic counterparts:

Lemma 1 (**Bayesian reweighting bound**). *Consider the quadrature rule* $\hat{\Pi}[f] = \sum_{i=1}^n w_i f(\mathbf{x}_i)$ *and the corresponding BQ rule* $\hat{\Pi}_{BQ}[f] = \sum_{i=1}^n w_i^{BQ} f(\mathbf{x}_i)$*. Then* $e(\hat{\Pi}_{BQ}; \Pi, \mathcal{H}_k) \leq e(\hat{\Pi}; \Pi, \mathcal{H}_k)$*.*

Proof. Since the BQ rule corresponds to the optimally weighted quadrature rule in $\mathcal{H}_k$, we must have that:

$$\begin{aligned} e(\hat{\Pi}_{\mathrm{BQ}}; \Pi, \mathcal{H}_k)^2 &= \left(\inf_{\mathbf{w} \in \mathbb{R}^n} \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \Pi[f] - \sum_{i=1}^n w_i f(\mathbf{x}_i) \right| \right)^2 \\ &\leq \left(\sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \Pi[f] - \hat{\Pi}[f] \right| \right)^2 \; = \; e(\hat{\Pi}; \Pi, \mathcal{H}_k)^2 \end{aligned}$$

□

Clearly, whenever we have a BQ rule based on reweighting an existing quadrature rule (e.g. BMC, BIS, BMCMC or BQMC), it is straightforward to obtain an upper bound on the WCE convergence rate if a convergence rate is known for the original rule. Results based on Lemma 1 are useful in that they provide us with some guarantees on the performance of the method, but tend to be unsatisfying for several reasons. First, they can lead to loose upper bounds since they do not take into account any gains in reweighing. Furthermore, since the BQ weights tend to be more expensive to compute, it is questionable whether reweighting can actually be beneficial from a point estimate point of view (there is of course still the advantage, from an uncertainty quantification point of view, of having a Bayesian estimator).

The second lemma which we use to derive convergence rates often leads to more satisfying results although it is not sharp in general [Ritter, 2000, Proposition II.4]. The lemma shows that the convergence of $\hat{\Pi}_{\mathrm{BQ}}[f]$ is controlled by quality of the GP mean approximation m_n:

Lemma 2 (**Regression bound**). *Let $f \in \mathcal{H}_k$ and fix states $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$. Then we have $\left|\Pi[f] - \hat{\Pi}_{\mathrm{BQ}}[f]\right| \leq \|f - m_n\|_{L^2(\mathcal{X};\Pi)}$.*

Proof. Applying Jensen's inequality we get:

$$\begin{aligned}\left|\Pi[f] - \hat{\Pi}_{\mathrm{BC}}[f]\right|^2 &= \left(\int_{\mathcal{X}} f(\mathbf{x}) - m_n(\mathbf{x})\Pi(\mathrm{d}\mathbf{x})\right)^2 \\ &\leq \int_{\mathcal{X}} (f(\mathbf{x}) - m_n(\mathbf{x}))^2 \Pi(\mathrm{d}\mathbf{x}) \;=\; \|f - m_n\|^2_{L^2(\mathcal{X};\Pi)},\end{aligned}$$

Taking square roots gives the required result. □

Using the lemma above, we can transfer known results from the literature on approximation with kernel interpolants, (or equivalently GP means) to results for BQ rules. These results usually depend on the kernel and on space-filling properties of the point set selection method for the domain $\mathcal{X}$. An important quantity to formalise this statement is the fill distance of a point set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$:

$$h_{\mathbf{X}} \;=\; \sup_{\mathbf{x} \in \mathcal{X}} \min_{i=1,\ldots,n} \|\mathbf{x} - \mathbf{x}_i\|_2. \tag{3.8}$$

Other quantities of interest include $q_{\mathbf{X}} := \frac{1}{2}\min_{j \neq k} \|\mathbf{x}_j - \mathbf{x}_k\|_2$, the separation radius, and $\rho_{\mathbf{X}} := h_{\mathbf{X}}/q_{\mathbf{X}}$, the mesh ratio. For most sensible choices of point sets, we have $h_{\mathbf{X}} \to 0$ as $n \to \infty$. For kernel interpolants, it is common to have an upper bound on the error: $|f(\mathbf{x}) - m_n(\mathbf{x})| \leq Cv(h_{\mathbf{X}})\|f\|_{\mathcal{H}_k}$, where the role of v can be compared with that of the power function in the scattered data approximation literature (see Wendland [2005][Section 11.1] for more details) and will depend on the kernel k. Such results can clearly be combined with Lemma 2 to get rates for BQ.

We now have two results, Lemma 1 and 2, which refer to the point estimators provided by BQ and can be used to provide convergence rates. However, we also aim to quantify the change in probability mass as the number of samples increases and a contraction rate is therefore also of interest. Fortunately, it is also possible to obtain such rates from convergence rates of the point estimators:

Lemma 3 (**BQ contraction bound**). *Assume $f \in \mathcal{H}_k$ and a GP prior with covariance k was specified. Suppose that $e(\hat{\Pi}_{BQ}; \Pi, \mathcal{H}_k) \leq \gamma_n$ where $\gamma_n \to 0$ as $n \to \infty$. Let $I_\delta = [\Pi[f] - \delta, \Pi[f] + \delta]$ be an interval of radius $\delta > 0$ centred on the true value of the integral. Then*

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \;=\; 1 - O(\exp(-(\delta^2/2)\gamma_n^{-2})).$$

This result demonstrates that the posterior distribution is well-behaved;

probability mass concentrates in a neighbourhood of $\Pi[f]$. Hence, if our prior is well calibrated (see Chapter 2), the posterior provides uncertainty quantification over the solution of the integral as a result of performing a finite number n of integrand evaluations.

3.3.2 Monte Carlo, Important Sampling and MCMC Point Sets

The three lemmas from the previous section provide us with a set of tools which can be used to analyse BQ rules based on specific point sets. In this section, we provide results for BQ rules based on points obtained through several Monte Carlo methods.

All of the results in this section will be on $\mathcal{X} = [0,1]^d$ with $d \in \mathbb{N}$ for simplicity, although this assumption can be relaxed in all cases. As a baseline, we begin by noting a general result for BMC estimation under weak conditions on the RKHS which is based on Lemma 1:

Theorem 8 (Consistency and contraction of BMC for functions in RKHSs with bounded kernel). *Let $\mathcal{X} = [0,1]^d$, $d \in \mathbb{N}$ and $\mathcal{H}_k$ be a RKHS satisfying $\sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x},\mathbf{x}) < \infty$. Then: $e(\hat{\Pi}_{BMC};\Pi,\mathcal{H}_k) = O_P(n^{-\frac{1}{2}})$. Furthermore, if $f \in \mathcal{H}_k$ and $\delta > 0$:*

$$\mathbb{P}\{\Pi[f]-\delta < \Pi[g_n] < \Pi[f]+\delta\} \quad = \quad 1 - O_P(\exp(-C_\delta n^{\frac{1}{d}})),$$

where $C_\delta > 0$ depends on δ.

In fact, similar results can also be obtained for IS and MCMC. As previously discussed, results based on Lemma 1 can be far from tight. This is clearly highlighted by the result in the theorem below, obtained using Lemma 2. These will assume that our RKHS is norm-equivalent to $\mathcal{H}_\alpha$, a classical Sobolev space of order α. We will need one of the following conditions on the point sets:

(A1) The states are generated IID from the measure Π, which is assumed to have a density bounded away from zero on $\mathcal{X}$.

(A2) The states are generated IID from some importance measure Π', which is assumed to have a density bounded away from zero on $\mathcal{X}$.

(A3) The states are generated by a reversible, uniformly ergodic Markov chain that targets the measure Π, which is assumed to have a density bounded away from zero on $\mathcal{X}$.

Furthermore, a minor technical assumption that enables us to simplify the presentation of results below is that the set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ may be augmented with a finite, predetermined set $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^m$ where m does not increase with n. Clearly this has no bearing on asymptotics.

Theorem 9 (Consistency and contraction of BMC, BIS and BMCMC in $\mathcal{H}_\alpha$). *Let $\mathcal{X} = [0,1]^d$ and let $\mathcal{H}_k$ be norm-equivalent to $\mathcal{H}_\alpha$ where $\alpha > d/2$, $\alpha \in \mathbb{N}$. Suppose $\hat{\Pi}_{BQ}[f]$ is a BQ rule with point set satisfying (A1), (A2) or (A3). Then:*

$$e(\hat{\Pi}_{BQ}; \Pi, \mathcal{H}_k) = O_P\left(n^{-\alpha/d+\epsilon}\right)$$

for all $\epsilon > 0$ arbitrarily small. Furthermore, if $f \in \mathcal{H}_k$ and $\delta > 0$,

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \quad = \quad 1 - O_P\left(\exp\left(-C_\delta n^{2(\frac{\alpha}{d}-\epsilon)}\right)\right),$$

where $C_\delta > 0$ depends on δ and ϵ is arbitrarily small.

The use of ϵ is common in the literature on quadrature rules in order to hide $O(\log n)$ terms. The convergence rate improves with smoothness, but it suffers from a curse of dimensionality. When the assumption that $\alpha > \frac{d}{2}$ does not hold, a root-n rate can still be recovered from Theorem 8. The Matérn kernel leads to function spaces which are norm-equivalent to the Sobolev spaces $\mathcal{H}_\alpha$, and so the result above provide a bound for some of the most common BQ rules. A lower bound for the WCE of randomised algorithms in $\mathcal{H}_\alpha$ in this setting is $O_P(n^{-\alpha/d-1/2})$ [Novak and Woźniakowski, 2010]. Thus our result shows that the point estimate is at most one MC rate away from being optimal.

The control variate trick of Bakhvalov [2015] can be used to achieve the optimal randomised WCE, but this steps outside of the Bayesian framework. Bach [2017] obtained a similar result for fixed n and a specific importance sampling distribution. However, this specific importance sampling distribution is difficult to sample from in general, and his analysis does not directly imply our asymptotic results and vice versa. After completion of this work, similar results appeared in Oettershagen [2017]; Bauer et al. [2017]; Kanagawa et al. [2017].

A slight extension of Theorem 9 shows that certain infinitely differentiable kernels lead to exponential rates. These include the Gaussian RBF kernel introduced in Chapter 2, as well as the multiquadric kernel: $k(\mathbf{x}, \mathbf{y}) = (c^2 + \|\mathbf{x} - \mathbf{y}\|_2^2)^{\frac{1}{2}}$ and the inverse-multiquadric kernel $k(\mathbf{x}, \mathbf{y}) = (c^2 + \|\mathbf{x} - \mathbf{y}\|_2^2)^{-\frac{1}{2}}$ for $c > 0$.

Theorem 10 (Consistency and contraction of BMC, BIS and BMCMC in RKHSs with infinitely differentiable kernels). *Let $\mathcal{X} = [0,1]^d$, and assume* *that $\mathcal{H}_k$ is a RKHS which is norm-equivalent to the RKHS with Gaussian RBF kernel, multiquadric kernel or inverse-multiquadric kernel. Suppose $\hat{\Pi}_{BQ}[f]$ is a BQ rule with states satisfying either (A1), (A2) or (A3). Then:*

$$e(\hat{\Pi}_{BQ}; \Pi, \mathcal{H}_k) = O_P\big(\exp(-Cn^{1/d-\epsilon})\big)$$

for $C > 0$ and $\forall \epsilon > 0$ arbitrarily small. Furthermore, if $f \in \mathcal{H}_k$ and $\delta > 0$,

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \;\; = \;\; 1 - O_P\left(\exp\left(-C_\delta \exp(2n^{\frac{1}{d}+\epsilon})\right)\right),$$

where $C_\delta > 0$ depends on δ and $\epsilon > 0$ can be arbitrarily small.

Once again, there is a curse of dimensionality kicking in, but it is difficult to assess how strong it will be since C_δ is usually not known in practice.

The theorems above can be generalised in several directions:

1. We can consider more general domains $\mathcal{X}$. Specifically, the scattered data approximation bounds that are used in our proof apply to any compact domain $\mathcal{X} \subset \mathbb{R}^d$ that satisfies an interior cone condition [Wendland, 2005, p.28]. Following Wendland [2005][Section 3.3], a domain $\mathcal{X} \subset \mathbb{R}^d$ is said to satisfy an interior cone condition if there exists an angle $\theta \in (0, \pi/2)$ and a radius $r > 0$ such that $\forall \mathbf{x} \in \mathcal{X}$, a unit vector $\xi(\mathbf{x})$ exists such that the cone $\{\mathbf{x} + \lambda \mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_2 = 1, \mathbf{y}^\top \xi(\mathbf{x}) \geq \cos\theta, \lambda \in [0, r]\}$ is contained in $\mathcal{X}$. This condition essentially excludes domains with pinch-points on the boundaries, i.e. regions with a $\prec$ shape. Technical results in this direction were established in Oates et al. [2018]; Kanagawa et al. [2017].

2. We can consider other spaces $\mathcal{H}_k$. Similar results can be obtained for power kernels, thin-plate splines and compact support kernels using the bounds provided in [Wendland, 2005, Section 11].

3. As discussed in [Kanagawa et al., 2017; Xi et al., 2018], the results above will also hold for certain quasi-uniform point sets. We say $\mathbf{X}$ is a quasi-uniform grid on $\mathcal{X} \subset \mathbb{R}^d$ if it satisfies $h_{\mathbf{X}} \leq C_1 n^{-\frac{1}{d}}$ for some $C_1 > 0$. If such a quasi-uniform point set also satisfies $h_{\mathbf{X}} \leq C_2 q_{\mathbf{X},\mathcal{X}}$ for some $C_2 > 0$, then the same rates as for MC/IS/MCMC will be attained.

4. All of the proofs in this section are based on showing that MC, IS and MCMC all lead to realisations which are close to a grid with high probability, and therefore reduce the fill distance at the same rate as this grid (once again with high probability). A sensible question is therefore “Should we should be

using a grid in the first place?". The answer will be no in general. Indeed, for integration, we are only interested in approximating the integrand well in regions of high probability for the distribution we are integrating against. Evaluating it in regions of low probability would be wasteful since these do not contribute much to the value of the integral. In general, these considerations will enter the rate constants, but our proof techniques do not allow us to track the dependence explicitly.

3.3.3 Quasi-Monte Carlo Point Sets

The previous section provided some theoretical results for BMC, BIS and BMCMC under various assumptions on the RKHS. The cases considered were (i) RKHSs with bounded kernel (Theorem 8), (ii) RKHSs norm-equivalent to a Sobolev space (Theorem 9) and (iii) RKHSs with infinitely differentiable kernel norm-equivalent to either the Gaussian RBF, multiquadric or inverse-multiquadric kernel (Theorem 10). Clearly, the stronger were the assumptions on the function class, the faster the convergence rates were. In this section, we provide a similar set of theorems for QMC point sets, under slightly different assumptions on the kernel (dictated by the QMC point sets studied).

The most commonly used QMC sequences are called low-discrepancy sequences, and include (amongst others) the Halton and Sobol sequences. In this case, the notion of discrepancy is given by the star discrepancy [Dick and Pillichshammer, 2010]: $D^*(\{\mathbf{x}_i\}_{i=1}^n) = \sup_{a\in[0,1]^d} |\frac{1}{n}\sum_{i=1}^n 1_{\{\mathbf{x}_i \in I_a\}} - \int_{\mathcal{X}} \mathbf{x}\mathrm{d}\mathbf{x}|$ where $I_a = [0,a_1) \times \ldots \times [0,a_d)$. A low-discrepancy sequence is a point sequence such that $D^*(\{\mathbf{x}_i\}_{i=1}^n) = O(\log(n)^d n^{-1})$. Our first result, based on Lemma 1, provides convergence and contraction rates for the WCE under the assumption that the RKHS is norm-equivalent to a Sobolev space of smoothness α:

Theorem 11 (Consistency and contraction of BQMC in Sobolev spaces). *Consider $\mathcal{X} = [0,1]^d$ with Π uniform on $\mathcal{X}$. Let $\mathcal{H}_k$ be a RKHS norm equivalent to $\mathcal{H}_\alpha$, a Sobolev space of smoothness α ($\alpha \in \mathbb{N}$ and $\alpha \geq \frac{d}{2}$). Suppose that states $\{\mathbf{x}_i\}_{i=1}^n$ are obtained from a low-discrepancy sequence. Then:*

$$e(\hat{\Pi}_{BQMC}; \Pi, \mathcal{H}_k) \;=\; O(n^{-1+\epsilon})$$

for all $\epsilon > 0$ arbitrarily small. Furthermore if $f \in \mathcal{H}_k$ and $\delta > 0$,

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \;=\; 1 - O(\exp(-C_\delta n^{2-\epsilon})),$$

where $C_\delta > 0$ depends on δ and $\epsilon > 0$ can be arbitrarily small.

Note that this result improves on the result for BMC for $\frac{d}{2} \leq \alpha < d$, is the same as BMC for $\alpha = d$, and is suboptimal for $\alpha > d$. The sub-optimal in the latter case is due to the use of a crude upper bound in the proof and it should be possible to improve on this in future work.

We now consider more interesting spaces of functions whose mixed partial derivatives exist and for which even faster convergence rates can be obtained using BQMC. Denote the Sobolev space of dominating mixed smoothness by $\mathcal{S}^\alpha$, where α is the order of the space. To build intuition, note that $\mathcal{S}^\alpha$ is norm-equivalent to the RKHS generated by a tensor product of Matérn kernels [Sickel and Ullrich, 2009], or indeed a tensor product of any other univariate Sobolev space-generating kernel. For $\mathcal{S}^\alpha$, an appropriate QMC point set would be a higher-order digital net; for details see Dick and Pillichshammer [2010].

Theorem 12 (Consistency and contraction of BQMC in Sobolev spaces of mixed dominating smoothness). *Consider $\mathcal{X} = [0,1]^d$ with Π uniform on $\mathcal{X}$. Let $\mathcal{H}_k$ be norm-equivalent to $\mathcal{S}^\alpha$, where $\alpha \geq 2$, $\alpha \in \mathbb{N}$. Suppose states are chosen according to a higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ net over $\mathbb{Z}_b$ for some prime b where $n = b^m$. Then:*

$$e(\hat{\Pi}_{BQMC}; \Pi, \mathcal{H}_k) \;=\; O(n^{-\alpha+\epsilon})$$

for all $\epsilon > 0$ arbitrarily small. If $f \in \mathcal{H}_k$ and $\delta > 0$,

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \quad = \quad 1 - O(\exp(-C_\delta n^{2\alpha-\epsilon})),$$

where $C_\delta > 0$ depends on δ and $\epsilon > 0$ can be arbitrarily small.

This result shows that the posterior is again well-behaved. Indeed, the rates of convergence and contraction are much faster in $\mathcal{S}^\alpha$ compared to $\mathcal{H}_\alpha$. In terms of point estimation, this is the optimal rate for any deterministic algorithm for integration of functions in $\mathcal{S}^\alpha$ [Novak and Woźniakowski, 2010]. These results should be understood to hold on the subsequence $n = b^m$, as QMC methods do not in general give guarantees for all $n \in \mathbb{N}$. It is not clear how far this result can be generalised, in terms of π and $\mathcal{X}$ since this would require the use of different QMC point sets. The case of QMC for infinitely differentiable kernels was recently studied in Fasshauer et al. [2012]; the results therein for Smolyak point sets imply (exponential) convergence and contraction rates for BQMC via the same arguments that we have made explicit for the space $\mathcal{S}^\alpha$.

3.4 Considerations for Practical Implementation

This concludes our theoretical study of the use of Monte Carlo methods for BQ. We have so far discussed BQ algorithms, and proved that they can provide point estimators with optimal, or near-optimal, consistency rates in most cases of interest. Chapter 4 will highlight extensions based on experimental design strategies. In the next section, we discuss details which are relevant to the practical implementation of BQ. In particular, we discuss our strategy for prior selection and highlight problems relating to the (lack of) tractability of kernel means.

3.4.1 Prior Specification for Integrands

An important point to make is that the theoretical results in the previous section do not address the important issue of whether the scale of the posterior uncertainty provides an accurate reflection of the actual numerical error. This is closely related to the well-studied problem of prior specification, which was discussed in Chapter 2. In the context of BQ, cross-validation and marginalisation should be reserved for cases where the integrand is very expensive to evaluate, in which case these approaches will be worthwhile from a computational point of view. When this is not the case, it will be preferable to use empirical Bayes. This is the main approach we will use in the remainder of this paper.

Note that it is sometimes possible to analytically marginalise certain types of scale parameters without impacting the conjugacy of the stochastic process. For example, the result below highlights how to marginalise an amplitude parameter using an objective prior:

Proposition 3 (**BQ with marginalised amplitude parameter**)**.** *Suppose our covariance function takes the form* $c(\mathbf{x},\mathbf{y};\lambda) = \lambda c_0(\mathbf{x},\mathbf{y})$ *where* $c_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is itself a covariance function and* $\lambda > 0$ *is an amplitude parameter. Consider the improper prior* $p(\lambda) \propto \frac{1}{\lambda}$*. Then the induced distribution on* $\Pi[g_n]$ *is a Student-t distribution with mean and variance*

$$\begin{aligned}\mathbb{E}\left[\Pi[g_n]\right] &= \Pi\left[c_0(\cdot,\mathbf{X})\right]\mathbf{C}_0^{-1}\mathbf{f},\\ \mathbb{V}\left[\Pi[g_n]\right] &= \frac{\mathbf{f}^\top\mathbf{C}_0^{-1}\mathbf{f}}{n}\left(\Pi\Pi[c_0(\cdot,\cdot)] - \Pi[\mathbf{c}_0(\cdot,\mathbf{X})]\mathbf{C}_0^{-1}\Pi[\mathbf{c}_0(\mathbf{X},\cdot)]\right),\end{aligned}$$

and n *degrees of freedom. Here* $(\mathbf{C}_0)_{i,j} = c_0(\mathbf{x}_i,\mathbf{x}_j)$*,* $(\mathbf{c}_0(\cdot,\mathbf{X}))_i = c_0(\cdot,\mathbf{x}_i)$*,* $\mathbf{c}_0(\cdot,\mathbf{X}) = \mathbf{c}_0(\mathbf{X},\cdot)^\top$*.*

Empirical results in the remainder of this section support the use of this approach, though we do not claim that this strategy is optimal.

$\mathcal{X}$	Π	c	Reference
$[0,1]^d$	Unif($\mathcal{X}$)	Wendland Tensor Product	Oates et al. [2017c]
$[0,1]^d$	Unif($\mathcal{X}$)	Matérn Weighted Tensor Product	Section 3.6.3
$[0,1]^d$	Unif($\mathcal{X}$)	Exponentiated Quadratic	Use of error function
$\mathbb{R}^d$	Mixt. of Gaussian	Exponentiated Quadratic	Kennedy [1998]
$\mathbb{S}^d$	Unif($\mathcal{X}$)	Gegenbauer	Section 3.6.4
Arbitrary	Unif($\mathcal{X}$) / Mixt. of Gauss.	Trigonometric	Integration by parts
Arbitrary	Unif($\mathcal{X}$)	Splines	Wahba [1991]
Arbitrary	Known moments	Polynomial Tensor Product	Briol et al. [2015a]
$\mathbb{R}^d$	$\alpha-$stable	$\alpha-$stable	Nishiyama and Fukumizu [2016], Section 6.
$\mathbb{R}^d$	Generalized hyperbolic	Generalized hyperbolic	Nishiyama and Fukumizu [2016], Section 6.

Table 3.1: A non-exhaustive list of distribution Π and covariance function c pairs that provide a closed-form expression for both the mean $\mu(\Pi)(\mathbf{x}) = \Pi[c(\cdot,\mathbf{x})]$ and the initial error $\Pi[\mu(\Pi)]$.

3.4.2 Tractable and Intractable Kernel Means

Recall that the BQ posterior mean is of the form $\hat{\Pi}_{\text{BQ}}[f] = \Pi[c(\cdot,\mathbf{X})]c(\mathbf{X},\mathbf{X})^{-1}f(\mathbf{X})$, and it should therefore be clear that the method can only ever be applied when the kernel mean $\Pi[c(\cdot,\mathbf{x})]$ can be evaluated in closed form. This section highlights the limited range of scenarios when this can be achieved, and highlights alternative strategies when this is not possible.

Tractable Kernel-Distribution Pairs

A few cases of covariance-measure pairs (c,Π) where the kernel mean is available in closed form are recorded in Table 3.1. In the event that the covariance function-distribution pair (c,Π) of interest does not lead to a closed-form covariance function mean, it is sometimes possible to determine another covariance function-density pair (c',Π') for which $\Pi'[c'(\cdot,\mathbf{x})]$ is available and such that $f(\mathbf{x})\pi(\mathbf{x})/\pi'(\mathbf{x}) \in \mathcal{H}_{c'}$. Then one can construct an importance sampling estimator

$$\Pi[f] \;=\; \int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x} \;=\; \int_{\mathcal{X}} f(\mathbf{x})\frac{\pi(\mathbf{x})}{\pi'(\mathbf{x})}\pi'(\mathbf{x})\mathrm{d}\mathbf{x} \;=\; \Pi'\left[f\frac{\pi}{\pi'}\right],$$

and proceed as above.

Bayesian Quadrature with Approximate Kernel Means

When obtaining a tractable kernel mean is not feasible, an alternative is to work with an approximate Bayesian quadrature rule as described in this section. Our approach is to consider a BQ rule based on a quadrature approximation of the kernel mean denoted ${}_a\hat{\Pi}_{\text{BQ}}[f]$. The weights of this quadrature rule are ${}_a\mathbf{w}_{\text{BQ}} = \mathbf{C}^{-1}{}_a\Pi[\mathbf{c}(\mathbf{X},\cdot)]$ and these approximate the optimal BQ weights based on a quadrature approximation ${}_a\Pi[\mathbf{c}(\mathbf{X},\cdot)]$ of the kernel mean [see also Proposition 1 in Sommariva and Vianello, 2006]. The following lemma demonstrates that we can bound the contribution of this error and inflate our posterior to reflect the additional uncertainty due to the approximation, so that uncertainty quantification is still provided.

Proposition 4 (WCE for BQ with approximate kernel mean). *Consider an empirical measure* ${}_a\Pi = \sum_{j=1}^m {}_a w_j \delta(\mathbf{x}_j)$ *which approximates the measure* Π. *Then BQ can be performed analytically with respect to* ${}_a\Pi$*; denote this estimator by* ${}_a\hat{\Pi}_{BQ}[f]$. *Moreover,*

$$e({}_a\hat{\Pi}_{BQ};\Pi,\mathcal{H}_c)^2 \;\leq\; e(\hat{\Pi}_{BQ};\Pi,\mathcal{H}_c)^2 + \sqrt{n}e({}_a\Pi;\Pi,\mathcal{H}_c)^2.$$

Under approximate BQ, the posterior variance cannot be computed in closed-form, but computable upper-bounds can be obtained and these can then be used to propagate numerical uncertainty through the remainder of our statistical task. The idea here is to make use of the triangle inequality:

$$e({}_a\hat{\Pi}_{\mathrm{BQ}};\Pi,\mathcal{H}_c) \;\;\leq\;\; e({}_a\hat{\Pi}_{\mathrm{BQ}};{}_a\Pi,\mathcal{H}_c) \;+\; e({}_a\Pi;\Pi,\mathcal{H}_c). \tag{3.9}$$

The first term on the RHS is now available analytically and its square is given by: $e({}_a\hat{\Pi}_{\mathrm{BQ}};{}_a\Pi,\mathcal{H}_c)^2 = {}_a\Pi_a\Pi[c(\cdot,\cdot)] - {}_a\Pi[\mathbf{c}(\cdot,\mathbf{X})]\mathbf{C}^{-1}{}_a\Pi[\mathbf{c}(\mathbf{X},\cdot)]$. For the second term, explicit upper bounds exist in the case where states ${}_a\mathbf{x}_i$ are independent random samples from Π. For instance, from [Song, 2008, Theorem 27] we have, for a radial covariance function c, uniform ${}_a w_j = m^{-1}$ and independent ${}_a\mathbf{x}_i \sim \Pi$,

$$e({}_a\Pi;\Pi,\mathcal{H}_c) \;\;\leq\;\; \frac{2}{\sqrt{m}}\sup_{\mathbf{x}\in\mathcal{X}}\sqrt{c(\mathbf{x},\mathbf{x})} + \sqrt{\frac{\log(2/\delta)}{2m}} \tag{3.10}$$

with probability at least $1-\delta$. See also Altun and Smola [2006]; Szabó et al. [2016]. (For dependent ${}_a\mathbf{x}_j$, the m in Equation 3.10 can be replaced with an estimate for the effective sample size). More efficient quadrature rules, such as QMC methods could of course also be used.

Write $C_{n,\gamma,\delta}$ for a $100(1-\gamma)\%$ credible interval for $\Pi[f]$ defined by the conservative upper bound described in Equations 3.9 and 3.10. Then we conclude that $C_{n,\gamma,\delta}$ is a $100(1-\gamma)\%$ credible interval with probability at least $1-\delta$. Note that, even though the credible region has been inflated, it still contracts to the truth, since the first term on the right-hand side in Proposition 4 can be bounded by the sum of $e({}_a\hat{\Pi}_{\mathrm{BQ}};\Pi,\mathcal{H}_c)$ and $e({}_a\hat{\Pi};\Pi,\mathcal{H}_c)$, both of which vanish as $n,m\to\infty$.

We pause to briefly discuss the utility and significance of such an approach. Obviously, the new approximation problem (that of approximating Π with ${}_a\Pi$) could also be computed with a BQ method, and we may hence end up in an "infinite regress" scenario [O'Hagan, 1991], where the new kernel mean is itself unknown and so on. However, one level of approximation may be enough in many scenarios. Indeed, by using MC to select $\{{}_a\mathbf{x}_j\}_{j=1}^m$ and increasing m sufficiently faster than n, the error term $\sqrt{n}e({}_a\Pi;\Pi,\mathcal{H}_c)^2$ can be made to vanish faster than $e(\hat{\Pi}_{\text{BQ}};\Pi,\mathcal{H}_c)^2$ and hence the WCE for ${}_a\hat{\Pi}_{\text{BQ}}$ will be asymptotically identical to the WCE for the (intractable) exact BQ estimator $\hat{\Pi}_{\text{BQ}}$. Therefore, it will be reasonable to expend computational effort on raising m in settings where evaluation of the integrand constitutes the principal computational. This is because approximating the kernel mean only requires sampling m times, but does not require us to evaluate the integrand. A formal analysis of the trade-off between m and n in terms of statistical efficiency and computational cost will be important future work.

There are also several possible alternative approaches. First, Oates et al. [2017c] proposed a fully Bayesian approach to the problem. The idea is to provide two prior models: one on the integrand f and one on the measure Π. One potential choice of model for Π is a Dirichlet process mixture model, in which case the posterior distribution on the kernel mean remains tractable for certain classes of covariance functions such as the Gaussian RBF covariance. This approach hence allows us to work without direct access to the kernel mean, but is also useful more generally when using BQ with intractable measures (such as an unnormalised or generative model). Another approach using Bayesian estimators of the kernel mean (following the methodology proposed in Flaxman et al. [2016]) was also discussed in the supplementary material of Oates et al. [2017c].

Another alternative is to construct a covariance function for which the kernel mean will always be available in closed form. Such an approach is possible using tools from Stein's method [Oates et al., 2017c, 2018; Oates and Girolami, 2016]. This will be discussed in detail in Chapter 5.

3.5 Simulation Study

We have now completed our introduction and theoretical study of BQ. The aims of the remainder of this chapter are two-fold. Firstly, in this section, we validate the preceding theoretical analysis and in particular:

1. Assess the uncertainty quantification properties of BQ estimators when using marginalisation and empirical Bayes to select parameters of the GP covariance

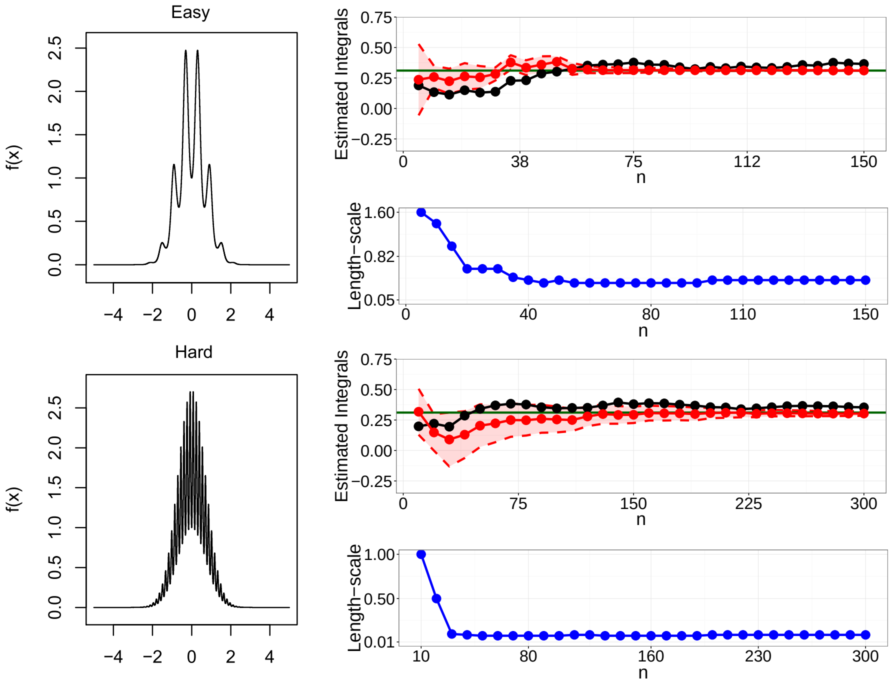

Figure 3.2: *Non-isotropic test functions for evaluation of the uncertainty quantification provided by Bayesian Monte Carlo and Bayesian quasi-Monte Carlo.* Empirical Bayes was used for σ whilst λ was marginalised. *Left:* The test functions f_1 (top), f_2 (bottom) in $d = 1$ dimension. *Right:* Solutions provided by Monte Carlo (MC; black) and Bayesian MC (BMC; red), for one typical realisation. 95% credible regions are shown for BMC and the green horizontal line gives the exact value of the integral. The blue curve gives the corresponding lengthscale parameter selected by empirical Bayes.

function. This is done by studying the frequentist coverage of our posterior distributions for several test functions of varying regularity.

2. Verify that the convergence rates from Section 3.3 hold in practice and verify how tight these are (all of the results are upper bounds on the error).

Secondly, in the next section, we will explore the use of BQ in a range of problems arising in contemporary statistical applications each demonstrating some advantages and disadvantages of the approach.

3.5.1 Assessment of Uncertainty Quantification

Our baseline problems for studying the uncertainty quantification provided by BQ include a non-isotropic test function with an "easy" setting $f_1(\mathbf{x}) = \exp\left(\sin(5x_1)^2 - \|\mathbf{x}\|_2^2\right)$ and a "hard" setting: $f_2(\mathbf{x}) = \exp\left(\sin(20x_1)^2 - \|\mathbf{x}\|_2^2\right)$. The easy test function does not vary much (see Figure 3.2), and as such will be easy to integrate using an interpolation-based method such as BQ. On the other hand, the hard test function is more variable and will hence be more difficult to approximate for the GP underlying BQ, but will not be significantly more difficult for MC since it is not based on interpolation. One realisation of states $\{\mathbf{x}_i\}_{i=1}^n$, generated independently and uniformly over $\mathcal{X} = [-5, 5]^d$ (initially $d = 1$), was used to estimate $\Pi[f_1]$ and $\Pi[f_2]$. We work in a RKHS characterised by tensor products of Matérn kernels

$$k_\alpha(\mathbf{x}, \mathbf{x}') = \lambda \prod_{i=1}^{d} \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}|x_i - x_i'|}{\sigma_i{}^2} \right)^\alpha K_\alpha \left(\frac{\sqrt{2\alpha}|x_i - x_i'|}{\sigma_i{}^2} \right),$$

where K_α is the modified Bessel function of the second kind. Closed-form kernel means exist in this case for $\alpha = p + 1/2$ whenever $p \in \mathbb{N}$.

In this setup, empirical Bayes was used to select the lengthscale parameters $\sigma = (\sigma_1, \ldots, \sigma_d) \in (0, \infty)^d$ of the kernel, while the amplitude parameter λ was marginalised as in Proposition 3. The smoothness parameter was fixed at $\alpha = 7/2$. Note that all test functions will be in the space $\mathcal{H}_\alpha$ for any $\alpha > 0$ and there is a degree of arbitrariness in this choice of prior.

Results are shown in Figure 3.2. Error-bars are used to denote the 95% posterior credible regions for the value of the integral and we also display the values $\hat{\sigma}_i$ of the length scale σ_i selected by empirical Bayes. The term "credible" is used loosely since the $\hat{\sigma}_i$ are estimated rather than marginalised. The $\hat{\sigma}_i$ appear to converge rapidly as $n \to \infty$; this is encouraging but we emphasise that Section 3.3 does not provide theoretical guarantees for empirical Bayes (all the results assume a fixed covariance function). On the negative side, over-confidence is possible at small values of n. Indeed, the BQ posterior is liable to be over-confident under empirical Bayes, since in the absence of evidence to the contrary, empirical Bayes selects large values for σ that correspond to more regular functions; this is most evident in the "hard" case.

Next we computed coverage frequencies for $100(1 - \gamma)\%$ credible regions. For each sample size n, the process was repeated over many realisations of the states $\{\mathbf{x}_i\}_{i=1}^n$, shown in Figure 3.3. It may be seen that (for n large enough) the uncertainty quantification provided by empirical Bayes is over-cautious for the easier function f_1, whilst being well-calibrated for the more complicated functions such as f_2. As expected, we observed that the coverage was over-confident for small values of n.

We can also study the performance in the case where both lengthscale σ and amplitude λ are optimised using empirical Bayes. In general this performed worse

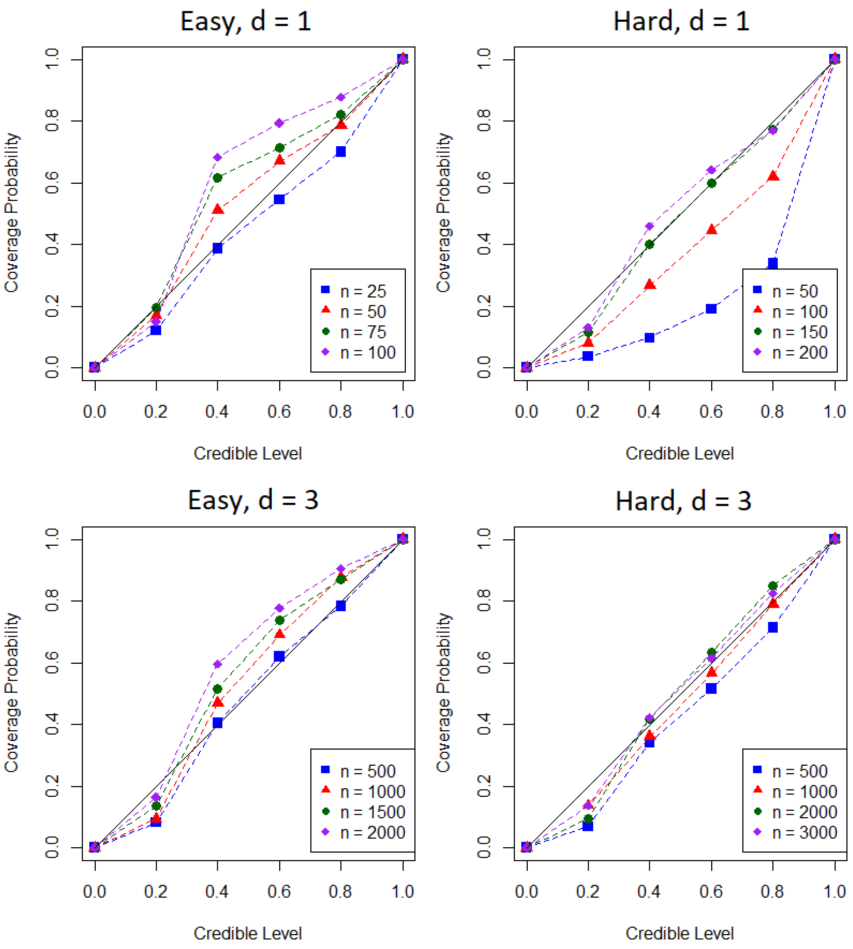

Figure 3.3: *Coverage of Bayesian Monte Carlo (with marginalisation) on the test functions* . Here we used empirical Bayes for σ with λ marginalised in dimensions $d = 1$ (top) and $d = 3$ (bottom). Coverage frequencies (computed from 500 (top) or 150 (bottom) realisations) were compared against notional $100(1-\gamma)\%$ Bayesian credible regions for varying level γ and number of observations n. The upper-left quadrant represents conservative credible intervals whilst the lower-right quadrant represents over-confident intervals. *Left:* “Easy” test function f_1. *Right:* “Hard” test function f_2.

than when λ was marginalised. In Figure 3.4 (top row) we study this case for the “easy” and “hard” test functions for $d = 1$. We notice that empirical Bayes led to over-confident inferences in the “low n” regime, but attains approximately correct frequentist coverage for larger n. Results are also shown in Figure 3.4 (bottom row) for when $d = 5$ but we have a single lengthscale parameter $\sigma = \sigma_1 = \ldots = \sigma_5$. Clearly more integrand evaluations are required for empirical Bayes to attain a

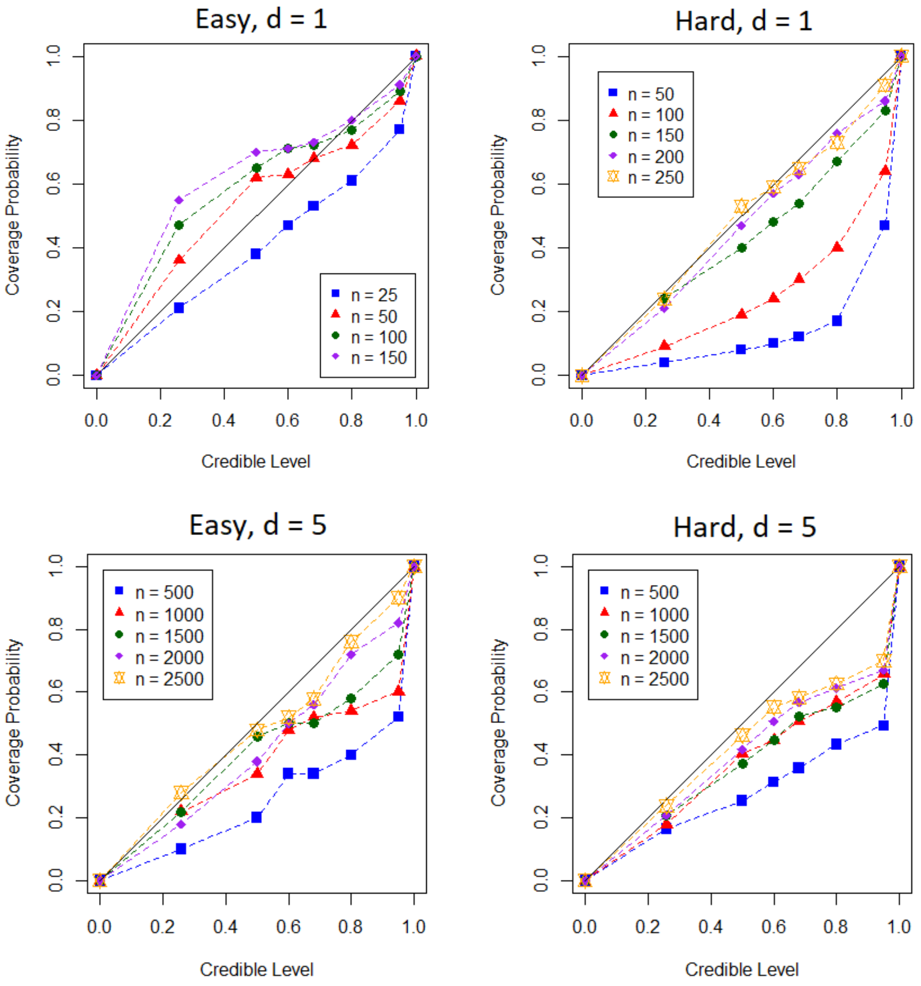

Figure 3.4: *Coverage of Bayesian Monte Carlo (without marginalisation) on the test functions.* Here, both σ and λ were picked using empirical Bayes. Results are shown for $d = 1$ (top) and $d = 5$ (bottom). Coverage frequencies $C_{n,\gamma}$ (computed from 100 (top) or 50 (bottom) realisations) were compared against notional $100(1-\gamma)\%$ Bayesian credible regions for varying level γ. *Left:* "Easy" test function f_1. *Right:* "Hard" test function f_2.

good frequentist coverage of the credible intervals, due to the curse of dimension. However, the frequentist coverage was once again reasonable for large n.

In summary, the results above illustrate the extent to which uncertainty quantification in possible using BQ. In particular, for our examples, we observed reasonable frequentist coverage if the number n of samples was not too small.

3.5.2 Validation of Convergence Rates

Our second set of experiments attempts to study whether the asymptotic convergence rates are realised in practice. We note that for kernels with a fixed lengthscale and amplitude parameter, the variance $\mathbb{V}[\Pi[g_n]]$, or equivalently the worst-case error in $\mathcal{H}_c$, is independent of the integrand and may be plotted as a function of n. The results below demonstrate that theoretical rates are observed in practice for $d = 1$ for BMC and BQMC; however, at large values of d, more data are required to achieve accurate estimation and increased numerical instability was observed.

BMC In Section 3.3.2, it was proven that the square-root of the BMC posterior variance converges at the rate $O_P(n^{-\alpha/d+\epsilon})$ when $\mathcal{H}_c$ is a Sobolev space of order $\alpha > d/2$. Figure 3.5 (top row) depicts empirical convergence results obtained for $d = 1$ (left) and $d = 5$ (right), for one typical realisation. In the one dimensional case, the $O_P(n^{-\alpha/d+\epsilon})$ theoretical convergence rates are broadly attained and indeed exceeded by at most one Monte Carlo rate. At larger values of n, numerical regularisation takes effect and damages the rate of convergence. In the higher dimensional case, the only rate proven in this work is $O_P(n^{-1/2})$ since $\alpha < d/2$ in all cases $p = \alpha + 1/2 \in \{3/2, 5/2, 7/2\}$ considered. These results show that a faster rate is attainable in practice, illustrating a gap in our theory.

BQMC In Section 3.3.3 it was proven that the square-root of the BQMC standard variance converges at the rate $O(n^{-\alpha+\epsilon})$ when $\mathcal{H}_c$ is a Sobolev space of dominating mixed smoothness and order $\alpha > 1/2$. Figure 3.5 (bottom row) depicts empirical convergence results obtained for $d = 1$ (left) and $d = 5$ (right), for one typical realisation. In the one dimensional case, the $O(n^{-\alpha+\epsilon})$ theoretical convergence rate is broadly attained in all cases $p = \alpha + 1/2 \in \{3/2, 5/2, 7/2\}$ considered. However, in the $d = 5$ case, the rates are not observed for the number n of evaluations considered. This helps us demonstrate the important point that the rates we provide are asymptotic, and may require large values of n before being observed in practice.

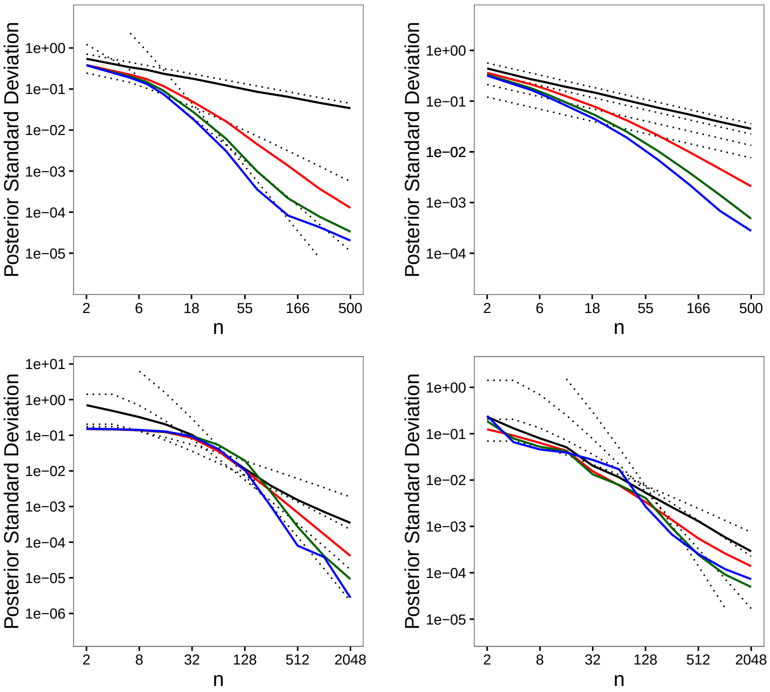

Figure 3.5: *Convergence rates for Bayesian Monte Carlo and Bayesian quasi-Monte Carlo.* WCE (or posterior standard deviation) for one realisation of BMC and BQMC on $[0,1]^d$ for $d = 1$ (left) and $d = 5$ (right). Here we considered BMC in Sobolev spaces $\mathcal{H}_\alpha$ (top row), and BQMC in Sobolev spaces of mixed dominating smoothness $\mathcal{S}^\alpha$ (bottom row). The results are obtained using tensor product Matérn kernels of smoothness $\alpha = 3/2$ (red), $\alpha = 5/2$ (green) and $\alpha = 7/2$ (blue). Dotted lines represent the theoretical convergence rates established for each kernel. The black line represents the corresponding standard Monte Carlo or quasi-Monte Carlo rate. Kernel parameters were fixed to $(\sigma, \lambda) = (0.02, 1)$ (top left), $(\sigma, \lambda) = (1.2, 1)$ (top right), $(\sigma, \lambda) = (0.005, 1)$ (bottom left) and $(\sigma, \lambda) = (1, 0.5)$ (bottom right).

3.6 Some Applications to Statistics and Engineering

Now that we have studied the suitability of BQ as a tool for uncertainty quantification in numerical analysis, we explore possible roles for BMC, BMCMC and BQMC in statistical applications. Four case studies, carefully chosen to highlight both the strengths and the weaknesses of BQ are presented:

1. A problem of Bayesian model selection, which is usually solved using thermodynamic integration, and for which we would like to model numerical error in the computation of model evidences.

2. A problem of computing posterior expectations over the parameters of some large-scale partial differential equation-based computer model of subsurface

flow for which MCMC sampling is computationally expensive.

3. A high-dimensional numerical integration problem occurring in semi-parametric random effect models when trying to access the observed data likelihood which was previously solved using QMC.

4. A problem of numerical integration in computer graphics for the rendering of virtual environments, for which BQ was previously used without theoretical guarantees.

3.6.1 Case Study 1: Large-Scale Model Selection

Consider the problem of selecting a single best model among a set $\{\mathcal{M}_i\}_{i=1}^M$, based on data $\mathbf{y}$ assumed to arise from a true model in this set. The Bayesian solution is to select the maximum a-posteriori model. We focus on the case with uniform prior on models $p(\mathcal{M}_i) = 1/M$, and this problem hence reduces to finding the largest marginal likelihood $p_i = p(\mathbf{y}|\mathcal{M}_i)$ since the maximum-a-posteriori model satisfies $p(\mathcal{M}_i|\mathbf{y}) = p(\mathbf{y}|\mathcal{M}_i)/M\sum_{j=1}^M p(\mathbf{y}|\mathcal{M}_j) \propto p(\mathbf{y}|\mathcal{M}_i)$. The p_i are usually intractable integrals over the parameters θ_i associated with model $\mathcal{M}_i$. One widely-used approach to model selection is to estimate each p_i in turn, say by $\hat{p}_i$, then to take the maximum of the $\hat{p}_i$ over $i = 1, \ldots, M$. In particular, thermodynamic integration is one approach to approximation of marginal likelihoods p_i for individual models [Gelman and Meng, 1998; Friel and Pettitt, 2008].

In many contemporary applications the maximum a-posteriori model is not well-identified, for example in variable selection where there are very many candidate models. Then, the computation becomes sensitive to numerical error in the $\hat{p}_i$, since an incorrect model $\mathcal{M}_i$, $i \neq k$ can be assigned an overly large value of $\hat{p}_i$ due to numerical error, in which case it could be selected in place of the correct maximum a-posteriori model. Below we explore the potential to exploit BQ to surmount this problem.

Thermodynamic Integration with Bayesian Quadrature

To simplify notation below we consider computation of a single p_i and suppress dependence on the index i corresponding to model $\mathcal{M}_i$. Denote the parameter space by Θ. For $t \in [0, 1]$ (an inverse temperature) define the power posterior Π_t, a measure over Θ with density $\pi_t(\theta) \propto p(\mathbf{y}|\theta)^t p(\theta)$. The thermodynamic identity is formulated as a double integral [Gelman and Meng, 1998]:

$$\log p(\mathbf{y}) = \int_0^1 \int_\Theta \log p(\mathbf{y}|\theta)\pi_t(\theta)\mathrm{d}\theta\mathrm{d}t.$$

The thermodynamic integral can be re-expressed as $\log p(\mathbf{y}) = \int_0^1 g(t)\mathrm{d}t$, $g(t) = \int_\Theta f(\theta)\pi_t(\theta)\mathrm{d}\theta$, where $f(\theta) = \log p(\mathbf{y}|\theta)$. Standard practice is to discretise the outer integral using a quadrature rule and estimate the inner integral using MCMC. Intuitively, this may be a convenient way of computing the model evidence since it requires sampling over the power posteriors, which will be tempered versions of the posterior. This will tend to be easier since tempering reduces the difficulties which come with multimodality of a target distribution.

Letting $0 = t_1 < \cdots < t_m = 1$ denote a fixed temperature schedule, we thus use the trapezium rule to obtain:

$$\log p(\mathbf{y}) \approx \sum_{i=2}^{m} (t_i - t_{i-1}) \frac{\hat{g}_i + \hat{g}_{i-1}}{2}, \qquad \hat{g}_i = \frac{1}{n}\sum_{j=1}^{n} \log p(\mathbf{y}|\theta_{i,j}),$$

where $\{\theta_{i,j}\}_{j=1}^n$ are MCMC samples from π_{t_i}. Several improvements have been proposed, including the use of higher-order numerical quadrature for the outer integral [Friel et al., 2014; Hug et al., 2016] and the use of control variates for the inner integral [Oates et al., 2016, 2017c].

Our proposal is to apply BQ to both the inner and outer integrals. This is instructive, since nested integrals are prone to propagation and accumulation of numerical error. In the Bayesian approach, the two integrands f and g are each assigned prior probability models. For the inner integral we assign a prior $f \sim \mathcal{N}(0, c_f)$. Our data here are the $nm \times 1$ vector $\mathbf{f}$ where $f_{(i-1)n+j} = f(\theta_{i,j})$. For estimating g_i with BQ we have m times as much data as for the MC estimator $\hat{g}_i$, which makes use of only n function evaluations. Here, information transfer across temperatures is made possible by the explicit model for f underpinning BQ.

In the posterior, $\mathbf{g} = (g(t_1), \ldots, g(t_T))$ is a Gaussian random vector with $\mathbf{g}|\mathbf{f} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ where the mean and covariance are given by $\mu_a = \Pi_{t_a}[\mathbf{c}_f(\cdot, \mathbf{X})]\mathbf{C}_f^{-1}\mathbf{f}$ and $\Sigma_{a,b} = \Pi_{t_a}\Pi_{t_b}[c_f(\cdot,\cdot)]] - \Pi_{t_a}[\mathbf{c}_f(\cdot,\mathbf{X})]\mathbf{C}_f^{-1}\Pi_{t_b}[\mathbf{C}_f(\mathbf{X},\cdot)]$, where $\mathbf{X} = \{\theta_{i,j}\}_{j=1}^n$ and $\mathbf{C}_f$ is a $nm \times nm$ covariance matrix defined by c_f.

For the outer integral, it is known that discretisation error can be substantial; Friel et al. [2014] proposed a second-order correction to the trapezium rule to mitigate this bias, while Hug et al. [2016] pursued the use of Simpson's rule. Attacking this problem from the probabilistic perspective, we do not want to place a stationary prior on $g(t)$, since it is known from extensive empirical work that $g(t)$ will vary more at smaller values of t. Indeed the rule-of-thumb $t_i = (i/m)^5$ is commonly used [Calderhead and Girolami, 2009].

We would like to encode this information into our prior. To do this, we proceed with an importance sampling step $\log p(\mathbf{y}) = \int_0^1 g(t)\mathrm{d}t = \int_0^1 h(t)\pi(t)\mathrm{d}t$. The rule-of-thumb implies an importance distribution $\pi(t) \propto 1/(\epsilon + 5t^{4/5})$ for some small $\epsilon > 0$, which renders the function $h = g/\pi$ approximately stationary (made precise in the following subsection). A stationary GP prior $h \sim \mathcal{N}(0, c_h)$ on the transformed integrand h provides the encoding of this prior knowledge that was used. Under this construction, the posterior $\log p(\mathbf{y})$ is Gaussian with mean and covariance defined as $\mathbb{E}[\log p(\mathbf{y})] = \Pi[\mathbf{c}_h(\cdot, \mathbf{T})]\mathbf{C}_h^{-1}\mu$ and

$$\mathbb{V}[\log p(\mathbf{y})] \quad = \quad \underbrace{\Pi\Pi[c_h(\cdot,\cdot)] - \Pi[\mathbf{c}_h(\cdot,\mathbf{T})]\mathbf{C}_h^{-1}\Pi[\mathbf{c}_h(\mathbf{T},\cdot)]}_{(*)} + \underbrace{\Pi[\mathbf{c}_h(\cdot,\mathbf{T})]\mathbf{C}_h^{-1}\mathbf{\Sigma}\mathbf{C}_h^{-1}\Pi[\mathbf{c}_h(\mathbf{T},\cdot)]}_{(**)},$$

where $\mathbf{T} = \{t_i\}_{i=1}^m$ and $\mathbf{C}_h$ is an $m \times m$ covariance matrix defined by c_h. The term $(*)$ arises from BQ on the outer integral, while the term $(**)$ arises from propagating numerical uncertainty from the inner integral through to the outer integral.

Experimental Setup

As a test-bed that captures the salient properties of model selection discussed above, we considered variable selection for logistic regression:

$$\begin{aligned} p(\mathbf{y}|\beta) &= \prod_{i=1}^N p_i(\beta)^{y_i}[1 - p_i(\beta)]^{1-y_i}, \\ \operatorname{logit}(p_i(\beta)) &= \gamma_1\beta_1 x_{i,1} + \ldots \gamma_d\beta_d x_{i,d}, \quad \gamma_1, \ldots, \gamma_d \in \{0, 1\} \end{aligned}$$

where the model $\mathcal{M}_k$ specifies the active variables via the binary vector $\gamma = (\gamma_1, \ldots, \gamma_d)$. A model prior $p(\gamma) \propto d^{-\|\gamma\|_1}$ was employed. Given a model $\mathcal{M}_k$, the active parameters β_j were endowed with independent priors $\beta_j \sim \mathcal{N}(0, \tau^{-1})$, where here $\tau = 0.01$.

A single dataset of size $N = 200$ were generated from model $\mathcal{M}_1$ with parameter $\beta = (1, 0, \ldots, 0)$; as such the problem is under-determined (there are in principle $2^{10} = 1024$ different models) and the true model is not well-identified. The selected model is thus sensitive to numerical error in the computation of marginal likelihood. In practice we limited the model space to consider only models with $\sum \gamma_i \leq 2$; this speeds up the computation and, in this particular case, only rules out models that have much lower posterior probability than the actual maximum a-posteriori model. There were thus 56 models being compared.

In this work we used the manifold Metropolis-adjusted Langevin algorithm [Girolami and Calderhead, 2011] in combination with population MCMC. Population MCMC shares information across temperatures during sampling, yet previous work has not leveraged evaluation of the log-likelihood f from one sub-chain t_i to inform estimates derived from other sub-chains $t_{i'}$, $i' \neq i$. In contrast, this occurs naturally in the BQ framework.

Here MCMC was used to generate a small number, $n = 200$, of samples on a per-model basis, in order to simulate a scenario where numerical error in computation of marginal likelihood will be non-negligible. A temperature ladder with $m = 10$ runs was employed, for the same reason, according to the recommendation of Calderhead and Girolami [2009]. No convergence issues were experienced; the same MCMC setup has previously been successfully used in Oates et al. [2016].

We motivate a prior for the unknown function g based on the work of Calderhead and Girolami [2009], who advocated the use of a power-law schedule $t_i = (\frac{i-1}{m-1})^5$, $i = 1, \ldots, m$, based on an extensive empirical comparison of possible schedules. A "good" temperature schedule approximately satisfies the criterion $|g(t_i)(t_{i+1} - t_i)| \approx m^{-1}$, on the basis that this allocates equal area to the portions of the curve g that lie between t_i and t_{i+1}, controlling bias for the trapezium rule. Substituting $t_i = (\frac{i-1}{m-1})^5$ into this optimality criterion produces $|g(t_i)|((i+1)^5 - i^5) \approx m^4$. Now, letting $i = \theta m$, we obtain $|g(\theta^5)|(5\theta^4 m^4 + o(m^4)) \approx m^4$. Formally treating θ as continuous and taking the $m \to \infty$ limit produces $|g(\theta^5)| \approx 0.2\theta^{-4}$ and so $|g(t)| \approx 0.2t^{-4/5}$. From this we conclude that the transformed function $h(t) = 5t^{4/5}g(t)$ is approximately stationary and can reasonably be assigned a stationary GP prior. However, in an importance sampling transformation we require that $\pi(t)$ has support over $[0, 1]$. For this reason we took $\pi(t) = 1.306/(0.01 + 5t^{4/5})$ in our experiment.

The covariance matrix $\mathbf{\Sigma}$ cannot be obtained in closed-form due to intractability of the kernel mean $\Pi_{t_i}[c_f(\cdot, \theta)]$. We therefore explored an approximation ${}_a\mathbf{\Sigma}$ such that plugging in ${}_a\mathbf{\Sigma}$ in place of $\mathbf{\Sigma}$ provides an approximation to the posterior variance $\mathbb{V}[\log p(\mathbf{y})]$ for the log-marginal likelihood. This took the form

$$ {}_a\Sigma_{i,j} \quad := \quad {}_a\Pi_{t_i} {}_a\Pi_{t_j}[c_f(\cdot,\cdot)] - {}_a\Pi_{t_i}[\mathbf{c}_f(\cdot, \mathbf{X})]\mathbf{C}_f^{-1} {}_a\Pi_{t_j}[\mathbf{c}_f(\mathbf{X}, \cdot)], $$

where an empirical distribution ${}_a\pi = \frac{1}{100}\sum_{i=1}^{100} \delta(\mathbf{x}_i)$ was employed based on the first $m = 100$ samples, while the remaining samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=101}^{200}$ were reserved for the covariance computation. This heuristic approach becomes exact as $m \to \infty$, in the sense that ${}_a\Sigma_{i,j} \to \Sigma_{i,j}$, but under-estimates covariance at finite m.

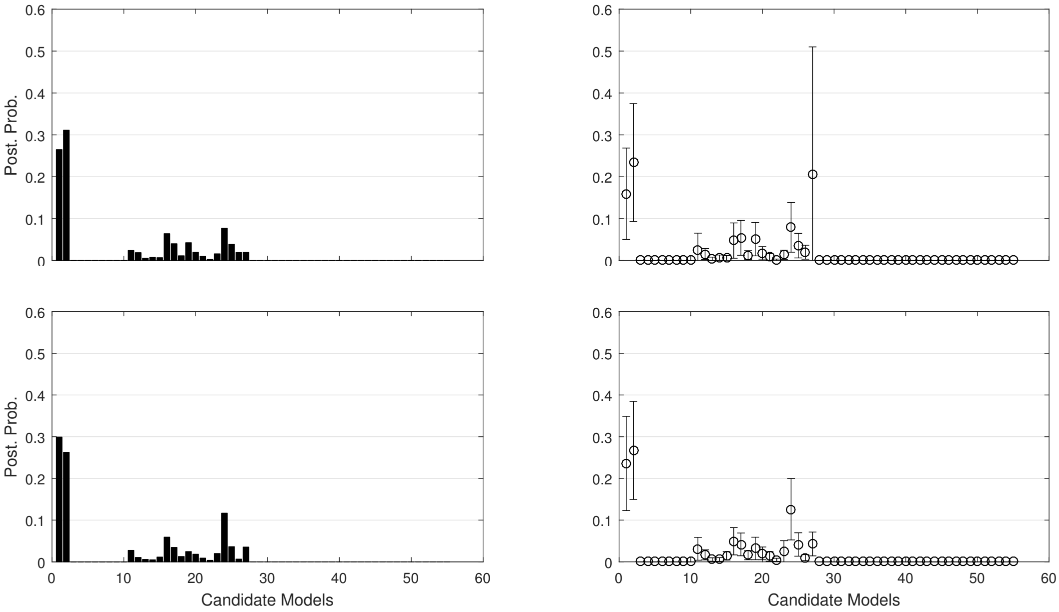

Figure 3.6: *Bayesian quadrature for thermodynamic integration.* Illustration on variable selection for logistic regression (with true model $\mathcal{M}_1$). Standard and probabilistic thermodynamic integration were used to approximate marginal likelihoods and, hence, the posterior over models. Each row represents an independent realisation of MCMC, while the data $\mathbf{y}$ were fixed. *Left:* Standard Monte Carlo, where point estimates for marginal likelihood were assumed to have no associated numerical error. *Right:* BQ, where a model for numerical error on each integral was propagated through into the posterior over models. The probabilistic approach produces a "probability distribution over a probability distribution", where the numerical uncertainty is modelled on top of the usual uncertainty associated with model selection.

In experiments below, both c_f and c_h were taken to be Gaussian covariance functions; for example: $c_f(\mathbf{x}, \mathbf{x}') = \lambda_f \exp\big(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\sigma_f^2\big)$ parameterised by λ_f and σ_f. This choice was made to capture smoothness of both integrands f and h involved. For this application we found that, while the σ parameters were possible to learn from data using empirical Bayes, the λ parameters required a large number of data to pin down. Therefore, for these experiments we fixed $\lambda_f = 0.1 \times \text{mean}(f_{i,j})$ and $\lambda_h = 0.01 \times \text{mean}(h_i)$. In both cases the remaining covariance function parameters σ were selected using empirical Bayes.

Results

Results are shown in Figure 3.6. Here we compared approximations to the model posterior obtained using the standard method versus the probabilistic method, over

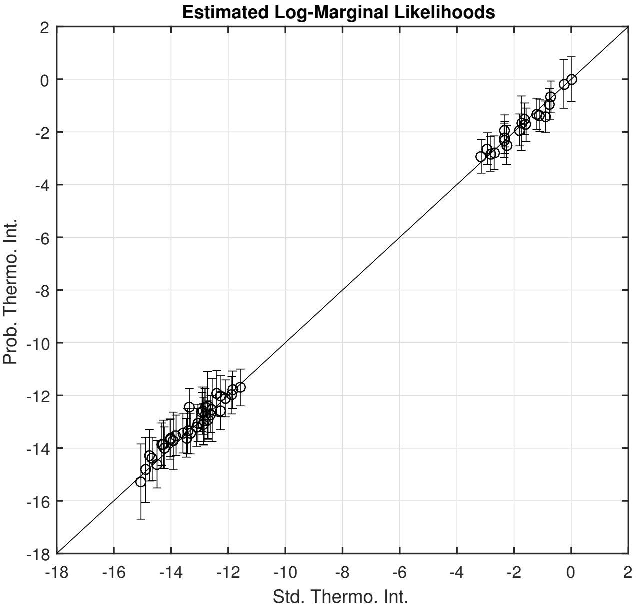

Figure 3.7: *Calibration of Bayesian quadrature for thermodynamic integration in model selection.* Estimates of marginal likelihoods $p_i = p(\mathbf{y}|\mathcal{M}_i)$. On the x-axis we show point estimates obtained by ignoring numerical error (the standard approach). On the y-axis we present the posterior mean estimates and ± one posterior standard deviation that aims to capture the extent of numerical error.

two realisations of the MCMC (the data $\mathbf{y}$ were fixed). The computation associated with BQ required less time, in total, than the time taken by MCMC.

An advantage of the Bayesian probabilistic numerical method approach is that it models numerical uncertainty on top of the usual statistical uncertainty. The same model was not always selected when numerical error was ignored and depended on the MCMC random seed. In contrast, under the probabilistic approach, either $\mathcal{M}_1$ or $\mathcal{M}_2$ could feasibly be the maximum a-posteriori under any of the MCMC realisations, up to numerical uncertainty. The top row of Figure 3.6 shows a large posterior uncertainty over the marginal likelihood for $\mathcal{M}_{27}$. This could be used as an indicator that more computational effort should be expended on this particular integral. The posterior variance was dominated by uncertainty due to discretisation error in the outer integral, rather than the inner integral. This suggests that numerical uncertainty could be reduced by allocating more computational resources to the outer integral rather than the inner integral.

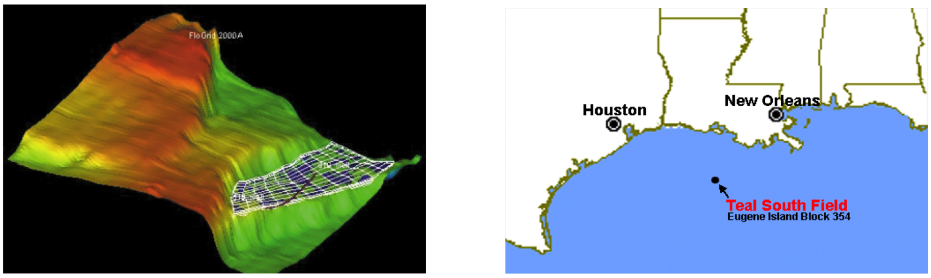

Figure 3.8: *The Teal South oil field. Left:* Computer model for the Teal South oil field. Simulation of this model requires significant computational resources. This renders any statistical analysis challenging due to the small number of data points (i.e. simulations) which can be obtained. *Right:* Location of the oil field.

3.6.2 Case Study 2: Computer Experiments

For our second case study, we consider an industrial scale computer model for the Teal South oil field, New Orleans [Hajizadeh et al., 2011] (see Figure 3.8). Conditional on field data, posterior inference was facilitated using state-of-the-art MCMC [Lan et al., 2016]. Oil reservoir models are generally challenging for MCMC. First, simulating from those models can be time-consuming, making the cost of individual MCMC samples a few minutes to several hours. Second, the posterior distribution will often exhibit strongly non-linear concentration of measure. Here we computed statistics of interest using BMCMC, where the uncertainty quantification afforded by BQ aims to enable valid inferences in the presence of relatively few MCMC samples.

Quantification of the uncertainty associated with predictions is a major topic of ongoing research in this field [Mohamed et al., 2010; Hajizadeh et al., 2011; Park et al., 2013] due to the economic consequences associated with inaccurate predictions of quantities such as future oil production rate. A probabilistic model for numerical error in integrals associated with prediction could provide a more complete uncertainty assessment.

The Teal South model is a partial differential equation computer model for an oil reservoir. The model studied is on an 11×11 grid with 5 layers. It has 9 parameters representing physical quantities of interest. These include horizontal permeabilities for each of the 5 layers, the vertical to horizontal permeability ratio, aquifer strength, rock compressibility and porosity. For our experiments, we used an emulator of the likelihood model documented in Lan et al. [2016] in order to speed up MCMC; however this might be undesirable in general due to the additional uncertainty associated with the approximation in the results obtained.

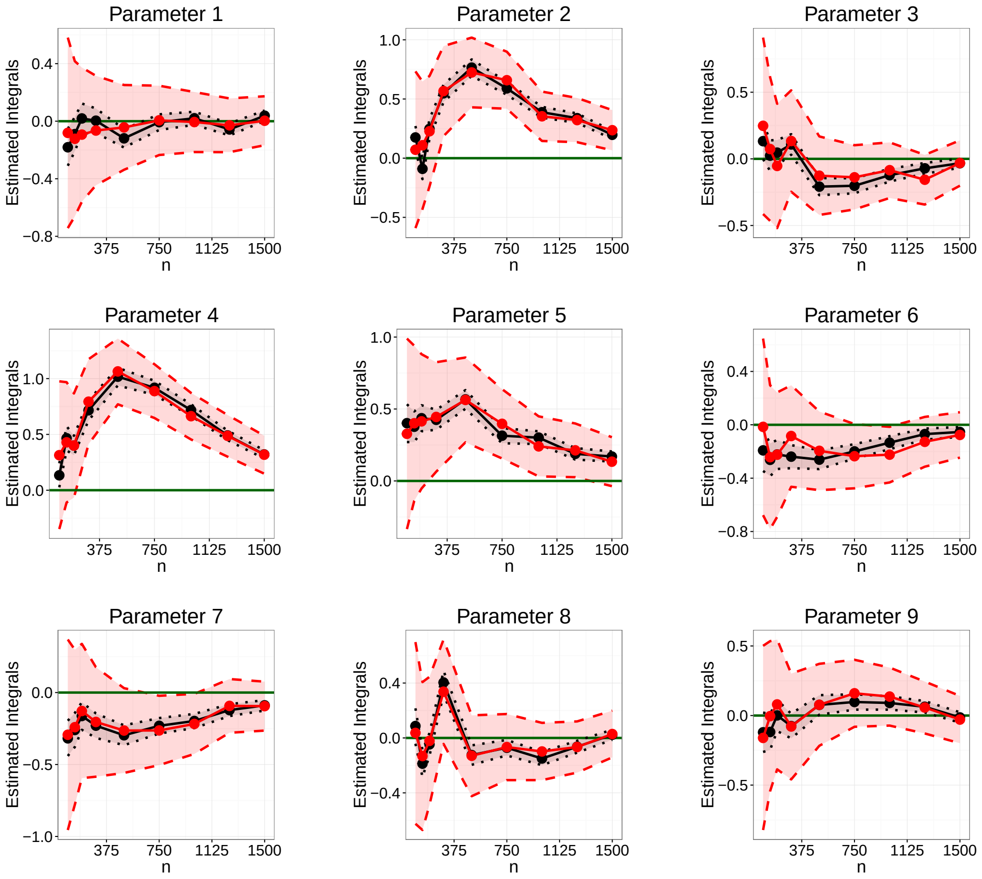

Figure 3.9: *Bayesian Markov chain Monte Carlo estimates of posterior means on the parameter of the Teal South oil field model (centered around the exact values).* The green line gives the exact value of the integral. The MCMC (black line) and BMCMC point estimates (red line) provided similar performance. The MCMC 95% confidence intervals, based on estimated asymptotic variance (black dotted lines), are poorly calibrated whereas with the BMCMC 95% credible intervals (red dotted lines) provide a more honest uncertainty assessment.

The particular integrals that we considered are posterior means for each model parameter, and we compared against an empirical benchmark obtained with brute force MCMC. BMCMC was employed with a Matérn $\alpha = 3/2$ covariance function whose lengthscale parameter was selected using empirical Bayes and the amplitude parameter was fixed to $\lambda = 1$.

Due to intractability of the posterior distribution, the kernel mean is unavailable in closed form. To overcome this, the methodology in Section 3.4.2 was employed to obtain an empirical estimate of the kernel mean (half of the MCMC samples were used with BQ weights to approximate the integral and the other half with MC weights to approximate the kernel mean). Equation 3.9 was used to upper bound the intractable BQ posterior variance. For the upper bound to hold, states ${}_a\mathbf{x}_j$ must be independent samples from Π, whereas here they were obtained using MCMC and were therefore not independent. In order to ensure that MCMC samples were "as independent as possible" we employed sophisticated MCMC methodology developed by Lan et al. [2016]. Nevertheless, we emphasise that there is a gap between theory and practice here that we hope to fill in future research.

Estimates for posterior means were obtained using both standard MCMC and BMCMC, shown in Figure 3.9. For this example the posterior distribution provides sensible uncertainty quantification for integrals 1, 3, 6-9, but was over-confident for integrals 2, 4, 5. The point accuracy of the BMCMC estimator matched that of the standard MCMC estimator. The lack of faster convergence for BMCMC appears to be due to inaccurate estimation of the kernel mean and we conjecture that alternative exact approaches, such as Oates et al. [2017c], may provide improved performance in this context. However, standard confidence intervals obtained from the central limit theorem for MCMC with a plug-in estimate for the asymptotic variance were over-confident for parameters 2-9.

3.6.3 Case Study 3: High-Dimensional Random Effects

Our aim here was to explore whether more flexible representations afforded by weighted combinations of Hilbert spaces could help scale BQ when $\mathcal{X}$ is high-dimensional. The focus was BQMC, but the methodology could be applied to BQ rules with other point sets.

Weighted Spaces

The formulation of high (and infinite)-dimensional QMC can be achieved with a construction known as a weighted Hilbert space. These spaces, defined below, are motivated by the observation that many integrands encountered in applications seem to vary more in lower dimensional projections compared to higher dimensional projections. Our presentation below follows Section 2.5.4 and 12.2 of Dick and Pillichshammer [2010], but the idea goes back at least to Wahba [1991, Chapter 10].

As usual with QMC, we work in $\mathcal{X} = [0,1]^d$ and Π uniform over $\mathcal{X}$. Let $\mathcal{I} = \{1, 2, \dots, d\}$. For each subset $u \subseteq \mathcal{I}$, define a weight $\gamma_u \in (0, \infty)$ and denote the collection of all weights by $\gamma = \{\gamma_u\}_{u \subseteq \mathcal{I}}$. Consider the space $\mathcal{H}_\gamma$ of functions of the form $f(\mathbf{x}) = \sum_{u \subseteq \mathcal{I}} f_u(\mathbf{x}_u)$, where f_u belongs to a RKHS $\mathcal{H}_{c_u}$ with kernel c_u and $\mathbf{x}_u$ denotes the components of $\mathbf{x}$ that are indexed by $u \subseteq \mathcal{I}$. This is not restrictive, since any function can be written in this form by considering only $u = \mathcal{I}$. We turn $\mathcal{H}_\gamma$ into a Hilbert space by defining an inner product $\langle f, g \rangle_\gamma := \sum_{u \subseteq \mathcal{I}} \gamma_u^{-1} \langle f_u, g_u \rangle_u$ where $\gamma = \{\gamma_u : u \subseteq \mathcal{I}\}$. Constructed in this way, $\mathcal{H}_\gamma$ is a RKHS with kernel $c_\gamma(\mathbf{x}, \mathbf{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u c_u(\mathbf{x}, \mathbf{x}')$. Intuitively, the weights γ_u can be taken to be small whenever the function f does not depend heavily on the $|u|$-way interaction of the states $\mathbf{x}_u$. Thus, most of the γ_u will be small for a function f that is effectively low-dimensional. A measure of the effective dimension of the function is given by $\sum_{u \subseteq \mathcal{I}} \gamma_u$; in an extreme case d could even be infinite provided that this sum remains bounded [Dick et al., 2013].

The (canonical) weighted Sobolev space of dominating mixed smoothness $\mathcal{S}_\gamma^\alpha$ is defined by taking each of the component spaces to be $\mathcal{S}^\alpha$. Constructed in this way, $\mathcal{S}_\gamma^\alpha$ is a RKHS with kernel

$$c_{\alpha,\gamma}(\mathbf{x}, \mathbf{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u \prod_{i \in u} \left(\sum_{k=1}^{\alpha} \frac{B_k(x_i) B_k(x_i')}{(k!)^2} - (-1)^\alpha \frac{B_{2\alpha}(|x_i - x_i'|)}{(2\alpha)!} \right),$$

where the B_k are Bernoulli polynomials. In finite dimensions, BQMC rules based on a higher-order digital nets attain optimal WCE rates $O(n^{-\alpha+\epsilon})$ for this RKHS:

Proposition 5 (**Consistency of BQMC in weighted Sobolev spaces of mixed dominating smoothness**). *Let $\mathcal{H}_c$ be a RKHS that is norm-equivalent to $\mathcal{S}_\gamma^\alpha$. Then BQMC based on a digital $(t, \alpha, 1, \alpha m \times m, d)$-net over $\mathbb{Z}_b$ attains the optimal rate $e(\hat{\Pi}_{\mathrm{BQMC}}; \Pi, \mathcal{H}_c) = O(n^{-\alpha+\epsilon})$ for any $\epsilon > 0$, where $n = b^m$.*

The QMC rules in Proposition 5 do not explicitly take into account the values of the weights γ. The net used in the proposition above is a popular QMC point set for $\mathcal{S}_\gamma^\alpha$, and we refer the reader to Dick et al. [2013] for more details.

An algorithm that tailors QMC states to specific weights γ is known as the component by component (CBC) algorithm; further details can be found in [Kuo, 2003]. In principle the CBC algorithm can lead to improved rate constants in high dimensions, because effort is not wasted in directions where f varies little, but the computational overheads are also greater. We did not consider CBC algorithms for BQMC in this work.

Note that the weighted Hilbert space framework allows us to bound the WCE independently of dimension providing that $\sum_{u \in \mathcal{I}} \gamma_u < \infty$ [Sloan and Woźniakowski, 1998]. This justifies the use of "high-dimensional" in this context. Analogous results for functional approximation were provided by Fasshauer et al. [2012] for the Gaussian kernel. Further details are provided in Section 4.1 of Dick et al. [2013].

Semi-Parametric Random Effects Regression

For illustration we considered generalised linear models, and focus on a Poisson semi-parametric random effects regression model studied by Kuo et al. [2008, Example 2]. The context is inference for the parameters β of the following model

$$\begin{aligned} Y_j|\lambda_j &\sim \mathrm{Po}(\lambda_j), \\ \log(\lambda_j) &= \beta_0 + \beta_1 z_{1,j} + \beta_2 z_{2,j} + u_1 \phi_1(z_{2,j}) + \cdots + u_d \phi_d(z_{2,j}), \\ u_j &\sim N(0, \tau^{-1}) \text{ independent.} \end{aligned}$$

Here $z_{1,j} \in \{0,1\}$, $z_{2,j} \in (0,1)$ and $\phi_j(z) = [z - \kappa_j]_+$ where $\kappa_j \in (0,1)$ are predetermined knots. We took $d = 50$ equally spaced knots in $[\min \mathbf{z}_2, \max \mathbf{z}_2]$. Inference for β requires multiple evaluations of the observed data likelihood $p(\mathbf{y}|\beta) = \int_{\mathbb{R}^d} p(\mathbf{y}|\beta, \mathbf{u}) p(\mathbf{u}) \mathrm{d}\mathbf{u}$ and therefore is a candidate for BQ methods, in order to model the cumulative uncertainty of estimating multiple numerical integrals.

In order to transform this integration problem to the unit cube we perform the change of variables $x_j = \Phi^{-1}(u_j)$ so that we wish to evaluate $p(\mathbf{y}|\beta) = \int_{[0,1]^d} p(\mathbf{y}|\beta, \Phi(\mathbf{x})) \mathrm{d}\mathbf{x}$. Here $\Phi^{-1}(\mathbf{x})$ denotes the standard Gaussian inverse cumulative distribution function applied to each component of $\mathbf{x}$. BQ proceeds under the hypothesis that the integrand $f(\mathbf{x}) = p(\mathbf{y}|\beta, \Phi(\mathbf{x}))$ belongs to (or at least can be well approximated by functions in) $\mathcal{S}_\gamma^\alpha$ for some smoothness parameter α and some weights γ. Intuitively, the integrand $f(\mathbf{x})$ is such that an increase in the value of x_j at the knot κ_j can be compensated for by a decrease in the value of x_{j+1} at a neighbouring knot κ_{j+1}, but not by changing values of $\mathbf{x}$ at more remote knots. Therefore we expect $f(\mathbf{x})$ to exhibit strong individual and pairwise dependence on the x_j, but expect higher-order dependency to be weaker. This motivates the weighted space assumption. Sinescu et al. [2012] provides theoretical analysis for the choice of weights γ. Here, weights γ of order two were used; $\gamma_u = 1$ for $|u| \leq d_{\max}$, $d_{\max} = 2$, $\gamma_u = 0$ otherwise, which corresponds to an assumption of low-order interaction terms (though f can still depend on all d of its arguments).

Results

Results in Figure 3.10 showed that the 95% posterior credible regions more-or-less cover the truth for this problem, suggesting that the uncertainty estimates are appropriate. On the negative side, the BQMC method does not encode non-negativity of the integrand and, consequently, some posterior mass is placed on negative values for the integral, which is not meaningful. To understand the effect

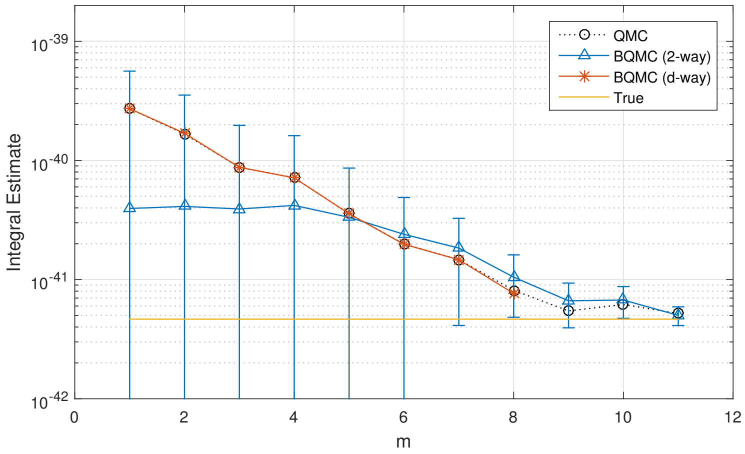

Figure 3.10: *Bayesian quasi-Monte Carlo for semi-parametric random effects regression in $d = 50$ dimensions, based on $n = 2^m$ samples from a higher-order digital net.* [Error bars show 95% credible regions. To improve visibility results are shown on the log-scale; error bars are symmetric on the linear scale. A brute-force QMC estimate was used to approximate the true value of the integral $p(\mathbf{y}|\beta)$ where $\beta = (0, 1, 1)$ was the data-generating value of the parameter.]

of the weighted space construction here, we compared against the BQMC point estimate with d-way interactions ($u \in \{\emptyset, \mathcal{I}\}$). An interesting observation was that these point estimates closely followed those produced by QMC.

3.6.4 Case Study 4: Computer Graphics

BQ can be defined on arbitrary manifolds, with formulations on non-Euclidean spaces suggested as far back as Diaconis [1988] and recently exploited in the context of computer graphics [Brouillat et al., 2009; Marques, 2013].

Global Illumination Integrals

Below we analyse BQMC on the d-sphere $\mathbb{S}^d = \{\mathbf{x} = (x_1, \ldots, x_{d+1}) \in \mathbb{R}^{d+1} : \|\mathbf{x}\|_2 = 1\}$ in order to estimate integrals of the form $\Pi[f] = \int_{\mathbb{S}^d} f(\mathbf{x})\Pi(\mathrm{d}\mathbf{x})$, where Π is the spherical measure (i.e. uniform over $\mathbb{S}^d$ with $\int_{\mathbb{S}^d} \Pi(\mathrm{d}\mathbf{x}) = 1$).

BQ is applied to compute global illumination integrals used in the rendering of surfaces [Pharr and Humphreys, 2004], and we therefore focus on the case where $d = 2$. Uncertainty quantification is motivated by inverse global illumination [e.g. Yu et al., 1999], where the task is to make inferences from noisy observation of an object via computer-based image synthesis; a measure of numerical uncertainty could naturally be propagated in this context. Below, to limit scope, we restrict

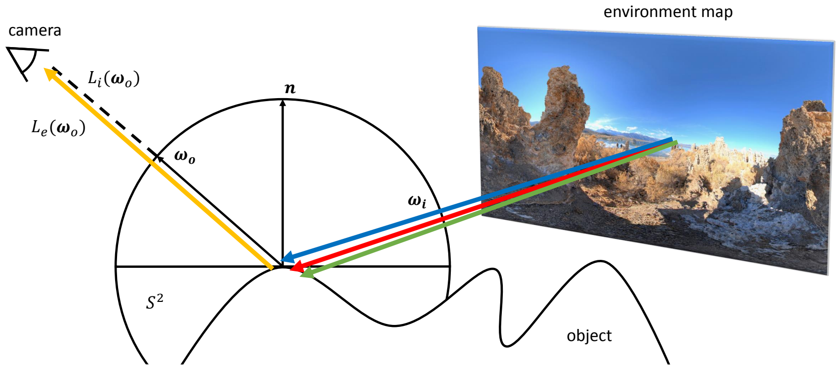

Figure 3.11: *Global illumination integrals in computer graphics.* The California lake environment map, shown, was used in our experiment.

attention to uncertainty quantification in the forward problem.

The models involved in global illumination are based on three main factors: a geometric model for the objects present in the scene, a model for the reflectivity of the surface of each object and a description of the light sources provided by an environment map. The light emitted from the environment will interact with objects in the scene through reflection. This can be formulated as an illumination integral[3]:

$$L_o(\omega_o) = L_e(\omega_o) + \int_{\mathbb{S}^2} L_i(\omega_i)\rho(\omega_i,\omega_o)[\omega_i \cdot \mathbf{n}]_+ \Pi(\mathrm{d}\omega_i). \tag{3.11}$$

The quantity $L_o(\omega_o)$ is called the outgoing radiance and represents the outgoing light in the direction ω_o. $L_e(\omega_o)$ represents the amount of light emitted by the object itself (which we will assume to be known) and $L_i(\omega_i)$ is the light hitting the object from direction ω_i. The term $\rho(\omega_i,\omega_o)$ is the bidirectional reflectance distribution function, which models the fraction of light arriving at the surface point from direction ω_i and being reflected towards direction ω_o. Here $\mathbf{n}$ is a unit vector normal to the surface of the object. A sketch of the problem is provided in Figure 3.11. Our investigation is motivated by strong empirical results for BQMC in this context obtained by Marques et al. [2013].

To assess the performance of BQMC we consider a typical illumination integration problem based on a California lake environment. The goal here is to compute intensities for each of the three red, green and blue colour channels corresponding to observing a virtual object from a fixed direction ω_o. We consider the case of an object directly facing the camera ($\mathbf{w}_o = \mathbf{n}$). For the bidirectional reflectance

[3] Although the integrand is only positive on part of the sphere, we have extended the integral to the entire sphere in order to be able to use a QMC point set defined for integration on the sphere.

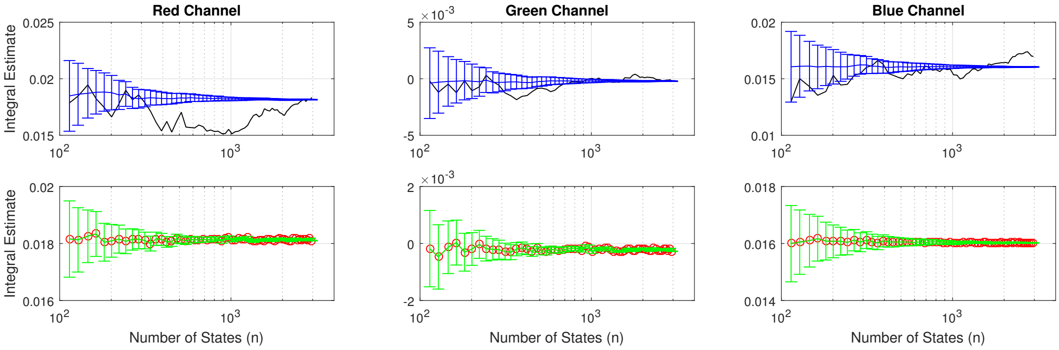

Figure 3.12: *Bayesian quadrature estimates of the red, green and blue colour intensities for the California lake environment.* [Error bars for BMC (blue) and BQMC (green) represent 95% credible intervals. MC estimates (black) and QMC estimates (red) are shown for reference.]

distribution function we took $\rho(\omega_i, \omega_o) = (2\pi)^{-1} \exp(\omega_i \cdot \omega_o - 1)$. The integrand $f(\omega_i) = L_i(\omega_i)\rho(\omega_i, \omega_o)[\omega_i \cdot \omega_o]_+$ was modelled in a Sobolev space of low smoothness. In contrast, Marques et al. [2013] viewed Equation 3.11 as an integral with respect to $\pi(\omega_i) \propto \rho(\omega_i, \omega_o)$ and posited a space of smooth integrands restricted to the hemisphere. The approach that we propose has two possible advantages; (i) it provides a closed-form expression for the kernel mean, (ii) a rougher kernel may be more appropriate in the context of illumination integrals, as pointed out by Brouillat et al. [2009]. The specific function space that we consider is the Sobolev space $\mathcal{H}_\alpha(\mathbb{S}^d)$ for $\alpha = 3/2$ (defined below).

Experimental Setup

The function spaces that we consider are Sobolev spaces $\mathcal{H}_\alpha(\mathbb{S}^d)$ for $\alpha > d/2$, obtained using the reproducing kernel $c(\mathbf{x}, \mathbf{x}') = \sum_{l=0}^{\infty} \lambda_l P_l^{(d)}(\mathbf{x}^\top \mathbf{x}')$, $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^d$, where $\lambda_l \asymp (1+l)^{-2\alpha}$ and $P_l^{(d)}$ are normalised Gegenbauer polynomials [Brauchart et al., 2014]. A particularly simple expression for the kernel in $d = 2$ and Sobolev space $\alpha = 3/2$ can be obtained by taking $\lambda_0 = 4/3$ along with $\lambda_l = -\lambda_0 \times (-1/2)_l/(3/2)_l$ where $(a)_l = a(a+1)\ldots(x+l-1) = \Gamma(a+l)/\Gamma(a)$ is the Pochhammer symbol. Specifically, these choices produce $c(\mathbf{x}, \mathbf{x}') = 8/3 - \|\mathbf{x} - \mathbf{x}'\|_2$, $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^2$. This covariance function is associated with a tractable kernel mean $\Pi[c(\mathbf{x}, \mathbf{x}')] = \int_{\mathbb{S}^2} c(\mathbf{x}, \mathbf{x}')\Pi(\mathrm{d}\mathbf{x}') = 4/3$ and hence the initial error is also available $\Pi\Pi[c] = \int_{\mathbb{S}^2} \Pi[c(\mathbf{x}, \mathbf{x}')]\Pi(\mathrm{d}\mathbf{x}) = 4/3$.

The states $\{\mathbf{x}_i\}_{i=1}^n$ could be generated with MC. In that case, analogous results to those obtained in Section 3.3.2 can be obtained. Specifically, from Theorem 7 of Brauchart et al. [2014], classical MC leads to slow convergence $e(\hat{\Pi}_{\mathrm{MC}}; \Pi, \mathcal{H}_k) = O_P(n^{-1/2})$. Rather than focusing on MC methods, we may also be interested in re-

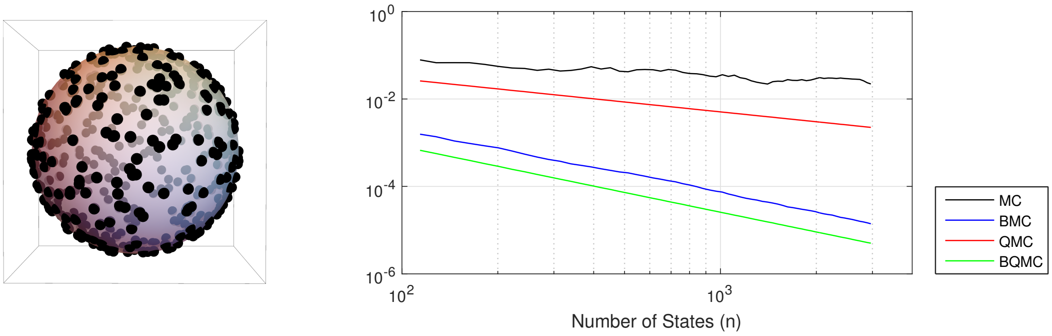

Figure 3.13: *Worst-case error of Bayesian Monte Carlo and Bayesian quasi-Monte Carlo for global illumination integrals. Left:* A spherical t-design over $\mathbb{S}^2$. *Right* The worst-case error, for Monte Carlo (MC), Bayesian MC (BMC), Quasi MC (QMC) and Bayesian QMC (BQMC).

sults based on spherical QMC point sets. We briefly introduce the concept of a spherical t-design [Bondarenko et al., 2013] which is define as a set $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{S}^d$ satisfying $\int_{\mathbb{S}^d} f(\mathbf{x})\Pi(\mathrm{d}\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$ for all polynomials $f : \mathbb{S}^d \to \mathbb{R}$ of degree at most t. (i.e. f is the restriction to $\mathbb{S}^d$ of a polynomial in the usual Euclidean sense $\mathbb{R}^{d+1} \to \mathbb{R}$). We now provide rates for Bayesian estimators in both of these cases.

Proposition 6 (**Consistency of BMC and BQMC for integration on the sphere**)**.** *Suppose that $\mathcal{X} = \mathbb{S}^d$ and Π is a uniform measure on this sphere. Assume that $\mathcal{H}_c$ is norm equivalent to a Sobolev space of smoothness $\alpha = \frac{3}{2}$ on $\mathcal{X}$. Then, if $d = 2$: $e(\hat{\Pi}_{BMC}; \Pi, \mathcal{H}_c) = O_P(n^{-3/4})$.*

Furthermore, $\forall d \geq 2$ there exists C_d such that for all $n \geq C_d t^d$ there exists a spherical t-design on $\mathbb{S}^d$ with n states. When $d = 2$, the use of a spherical t-design leads to a rate of $e(\hat{\Pi}_{\mathrm{BQMC}}; \Pi, \mathcal{H}_c) = O(n^{-3/4})$.

The rate in Proposition 6 is best-possible for a deterministic method in $\mathcal{H}_{\frac{3}{2}}(\mathbb{S}^2)$ [Brauchart et al., 2014]. Although both BMC and BQMC have the same convergence rate, the rate constant will usually be better for the QMC case (although this is not explicit in our theoretical result). Additional theoretical results on point estimates can be found in Fuselier et al. [2014]. Although explicit spherical t-designs are not currently known in closed-form, approximately optimal point sets have been computed numerically to high accuracy. Our experiments were based on such point sets provided on `http://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/sf.html` [Accessed 24 Nov. 2015].

Results

Both BMC and BQMC were tested on an environment map freely available at: `http://www.hdrlabs.com/sibl/archive.html` [Accessed 23 May 2017]. To ensure fair comparison, identical kernels were taken as the basis for both methods.

Figure 3.12 shows performance in red/green/blue-space. For this particular test function, the BQMC point estimate was almost identical to the QMC estimate at all values of n. Overall, both BMC and BQMC provided sensible quantification of uncertainty for the value of the integral at all values of n that were considered. In Figure 3.13, the value of the WCE is plotted for each of the four methods considered (MC, QMC, BMC, BQMC) as the number of states increases. Both BMC and BQMC appear to attain the same rate for $\mathcal{H}_{3/2}(\mathbb{S}^2)$, although BQMC provides a constant factor improvement over BMC. Note that $O(n^{-3/4})$ was shown by Brauchart et al. [2014] to be best possible for a deterministic method in the space $\mathcal{H}_{3/2}(\mathbb{S}^2)$.

Chapter 4

Bayesian Numerical Integration: Advanced Methods

> “The most popular option, however, is to drown our sorrows in alcohol, get punch drunk, and stumble around all night. The technical term for this is Markov chain Monte Carlo, or MCMC for short”
>
> Pedro Domingos, The Master Algorithm

The previous chapter introduced BQ and focused mostly on its variants based on MC, IS, MCMC and QMC point sets. We then provided theory on their asymptotic properties and showed that these optimally-weighted quadrature rules can provide significant improvements in convergence rate under smoothness assumptions on the integrand. We also demonstrated the coverage properties of these Bayesian estimators on some toy problems and a range of applied statistical inference problems.

An important takeaway is that Bayesian numerical methods can provide useful uncertainty quantification for problems in numerical analysis. This will however usually come with some significant additional computational cost. In the case of BQ with conjugate Gaussian or Student-t models, the worst-case computational cost will be $O(n^3)$, whereas for non-conjugate models, the cost could be significantly greater due to the additional necessity of approximating the posterior with MCMC. Before deciding whether to use a Bayesian probabilistic numerical method, one should therefore balance this additional cost with the value provided by the uncertainty quantification.

What is clear is that BQ will be most useful for models where evaluating the integrand f is expensive. In these cases, it is most likely that the numerical error remaining will be large, and a model of this error will hence be useful. The fast convergence rates provided by BQ are also useful here, since it will allow us to reduce the error with fewer function evaluations.

In this chapter, we focus on several extensions of BQ which could help further reduce the number of integrand evaluations to attain a fix error threshold.

- Section 4.1 proposes an extension of BQ to model several integrands simultaneously or sequentially. This formulation allows us to obtain a joint posterior on the integral of these functions. On the technical side, this requires the use of vector-valued reproducing kernel Hilbert spaces, which we will use to encode the correlation across integrands. This extension of BQ is useful as it can improve the speed of convergence of the associated estimators and can provide more accurate quantification of uncertainty.

The rest of the chapter will then focus on efficient point selection schemes for BQ. Even though obtaining the optimal set of points for BQ is an intractable problem, we demonstrate that several point selection schemes can significantly improve on the use of simple MC, MCMC or QMC point sets.

- Section 4.2 proposes a sampling scheme closely related to the idea of experimental design and which is based on the Frank-Wolfe algorithm. Points are chosen sequentially to minimise the posterior variance on $\mathbb{V}[\Pi[g_n]]$. We can also prove convergence results in this case, although this is limited to kernels corresponding to finite-dimensional RKHSs.

- Section 4.3 proposes a sequential MC sampler which aims to approximate the optimal BQ importance sampling distribution for a fixed number of integrand n. Once again this will be particularly useful when the integrand is expensive and so the choice of where to evaluate the function is of great importance.

4.1 Bayesian Quadrature for Multiple Related Integrals

We have already discussed several advantages of probabilistic numerical methods such as quantification of the uncertainty associated with numerical error. However, one property which has not been studied so far is the possibility of jointly inferring several quantities of interest (although briefly mentioned in Hennig et al. [2015] for linear algebra). In this section, we study the problem of numerically integrating a sequence of functions $f_1, \ldots, f_P$, which are correlated to one another, with respect to some probability measure Π. In many applications where we are faced with this type of problem, we also have prior knowledge about correlations between the individual

f_p. However, this information is often ignored and the integrals are approximated individually. This is not principled from a Bayesian point of view since it means we are not conditioning on all available information. In this section, we extend the BQ algorithm to solve this problem by building a joint model of $f_1, \ldots, f_P$ in order to obtain a joint posterior on the integrals $\Pi[f_1], \ldots, \Pi[f_P]$. Such a joint model allows for better finite-sample performance, and can also lead to more refined posterior distributions on each of the individual integrals.

4.1.1 Multi-output Bayesian Quadrature

Suppose we have a sequence of functions $f_p : \mathcal{X} \to \mathbb{R}$ $(p = 1, \ldots, P)$ for which we are interested in numerically computing integrals of the form $\Pi[f_p]$. For notational convenience, we will restrict ourselves to the case where all of the input domains are identical and denoted $\mathcal{X}$, all of the probability measures are identical and denoted Π, and the input sets $\mathbf{X} = \{\mathbf{X}_p\}_{p=1}^{P}$ consist of n points $\mathbf{X}_p = (\mathbf{x}_{p1}, \ldots, \mathbf{x}_{pn})$ per output function f_p. This setup can be made more general if necessary, but these assumptions will significantly simplify presentation. We reframe the integration problem as that of integrating some vector-valued function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^P$ such that $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_P(\mathbf{x}))^\top$. In other words. we want to estimate the vector $\Pi[\mathbf{f}] = (\Pi[f_1], \ldots, \Pi[f_P])^\top$. In this multiple-integrals setting, we could consider generalised quadrature rules of the form:

$$\hat{\Pi}[f_p] \quad = \quad \sum_{p'=1}^{P} \sum_{i=1}^{n} (\mathbf{W}_i)_{pp'} f_{p'}(\mathbf{x}_{p'i}),$$

where $\mathbf{W}_i \in \mathbb{R}^{P \times P}$ are weight matrices and $(\mathbf{W}_i)_{pp'}$ gives the influence of the value of $f_{p'}$ at $\mathbf{x}_{p'i}$ on the estimate of $\Pi[f_p]$. The quadrature rule for $\mathbf{f}$ can be rewritten in compact form as $\hat{\Pi}[\mathbf{f}] = \mathbf{W}^\top \mathbf{f}(\mathbf{X})$ for some weight matrix $\mathbf{W} \in \mathbb{R}^{nP \times P}$ (a concatenation of the weight $\{\mathbf{W}_i\}_{i=1}^{n}$) and function-evaluations vector $\mathbf{f}(\mathbf{X}) = (f_1(\mathbf{x}_{11}), \ldots, f_1(\mathbf{x}_{1n}), \ldots, f_P(\mathbf{x}_{P1}), \ldots, f_P(\mathbf{x}_{Pn}))^\top$.

These generalised quadrature rules encompass popular MC methods such as control variates or functionals [Glasserman, 2004; Oates et al., 2017c], multi-level Monte Carlo Giles [2015] and multi-fidelity Monte Carlo [Peherstorfer et al., 2016b]. However, it is important to point out that these MC methods can only deal with very specific relations between integrands, usually requiring $\left(\int_{\mathcal{X}} (f_p(\mathbf{x}) - f_{p'}(\mathbf{x}))^2 \Pi(\mathrm{d}\mathbf{x})\right)^{\frac{1}{2}}$ to be small for all pairs of integrands $f_p, f_{p'}$. Our method will be able to make use of much more complex relations between functions.

We propose to approach this problem using an extended version of BQ, where we impose a prior stochastic process $\mathbf{g} : \mathcal{X} \times \Omega \to \mathbb{R}^P$ which is a GP with mean a zero vector of size P and covariance function $\mathbf{C} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{P \times P}$. This is often called a multi-output GP or co-kriging model Álvarez and Lawrence [2011]). $\mathbf{C}$ is now matrix-valued and has entries $(C(\mathbf{x}, \mathbf{x}'))_{pp'} = \mathbb{E}_{\mathbb{P}}[g_p(\mathbf{x}, \omega) g_{p'}(\mathbf{x}', \omega)]$. In this case the stochastic process after observing some data is denoted $\mathbf{g}_n$ and is once again a GP. This GP has vector-valued mean $\mathbf{m}_n : \mathcal{X} \to \mathbb{R}^P$ and matrix-valued covariance $\mathbf{C}_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{P \times P}$ given by:

$$\begin{aligned} \mathbf{m}_n(\mathbf{x}) &= \mathbf{C}(\mathbf{x}, \mathbf{X}) \mathbf{C}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}(\mathbf{X}), & (4.1) \\ \mathbf{C}_n(\mathbf{x}, \mathbf{x}') &= \mathbf{C}(\mathbf{x}, \mathbf{x}') - \mathbf{C}(\mathbf{x}, \mathbf{X}) \mathbf{C}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{C}(\mathbf{X}, \mathbf{x}'). & (4.2) \end{aligned}$$

for $\mathbf{C}(\mathbf{x}, \mathbf{X}) = (C(\mathbf{x}, \mathbf{x}_1), \ldots, C(\mathbf{x}, \mathbf{x}_n)) \in \mathbb{R}^{P \times nP}$ and Gram matrix $\mathbf{C}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{nP \times nP}$ is:

$$\mathbf{C}(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} (\mathbf{C}(\mathbf{X}_1, \mathbf{X}_1))_{1,1} & \ldots & (\mathbf{C}(\mathbf{X}_1, \mathbf{X}_P))_{1,P} \\ (\mathbf{C}(\mathbf{X}_2, \mathbf{X}_1))_{2,1} & \vdots & (\mathbf{C}(\mathbf{X}_2, \mathbf{X}_P))_{2,P} \\ \vdots & \vdots & \vdots \\ (\mathbf{C}(\mathbf{X}_P, \mathbf{X}_1))_{P,1} & \ldots & (\mathbf{C}(\mathbf{X}_P, \mathbf{X}_P))_{P,P} \end{bmatrix}.$$

where $(C(\mathbf{X}_p, \mathbf{X}_{p'}))_{p,p'}$ is an $n \times n$ matrix.

Notice the similarity between Equations 4.1 and 4.2 and the equations for the posterior GP in the uni-output case in Chapter 2. The distribution $\Pi[\mathbf{g}_n]$ can also be obtained whenever the kernel mean $\Pi[\mathbf{C}(\cdot, \mathbf{x})]$ and initial error $\Pi\Pi\left[\mathbf{C}\right]$ are available in closed form, and will also be closely related to the result in Proposition 1 in Chapter 3:

Proposition 7 (Multi-output BQ posterior distribution on the solution of the integrals). *Consider multi-output BQ with a GP prior on* $\mathbf{f} = (f_1, \ldots, f_P)^\top$ *which has mean* $\mathbf{0}$ *and covariance function* $\mathbf{C}$*. The distribution of* $\Pi[\mathbf{g}_n]$ *is a* P*-dimensional Gaussian distribution with mean and covariance matrix:*

$$\begin{aligned} \mathbb{E}\left[\Pi[\mathbf{g}_n]\right] &= \Pi[\mathbf{C}(\cdot, \mathbf{X})] \mathbf{C}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}(\mathbf{X}), \\ \mathbb{V}\left[\Pi[\mathbf{g}_n]\right] &= \Pi\Pi\left[\mathbf{C}\right] - \Pi[\mathbf{C}(\cdot, \mathbf{X})] \mathbf{C}(\mathbf{X}, \mathbf{X})^{-1} \Pi[\mathbf{C}(\mathbf{X}, \cdot)]. \end{aligned}$$

The proof is identical to the uni-output case and hence omitted. In this case, we clearly end up with a generalised quadrature rule with weight matrix: $\mathbf{W}^{\mathrm{BQ}} = \mathbf{C}(\mathbf{X}, \mathbf{X})^{-1} \Pi\left[\mathbf{C}(\cdot, \mathbf{X})\right]^\top \in \mathbb{R}^{nP \times P}$. In general, the computational cost for computing the posterior mean and variance is now of order $O(n^3 P^3)$ instead of the $O(n^3)$ for the uni-output setting. However, several choices of covariance functions can reduce this cost significantly, and it is also possible to obtain sparse GP approximations [Álvarez and Lawrence, 2011].

The choice of covariance function $\mathbf{C}$ is of course once again of great importance since it encodes prior knowledge about each of the integrand and their correlation structure and should be made based on the application considered. We also remark that matrix valued covariance functions $\mathbf{C}$ can be described in term of some scalar-valued covariance function r on the extended space $\mathcal{X} \times \{1, \dots, P\}$ as $(\mathbf{C}(\mathbf{x}, \mathbf{x}'))_{pp'} = r((\mathbf{x}, p), (\mathbf{x}', p'))$. We now present two choices of covariance functions which are popular in the literature and will be used in the applications:

- The simplest example are separable covariance functions, which are of the form
$$\mathbf{C}(\mathbf{x}, \mathbf{x}') \quad = \quad \mathbf{B}c(\mathbf{x}, \mathbf{x}'),$$
where $\mathbf{B} \in \mathbb{R}^{P \times P}$ is symmetric and positive definite, and $\mathbf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued covariance function. This treats the covariance as the product of two scalar-valued covariance functions, one defined on $\mathcal{X}$ and the other on $\{1, \dots, P\}$. If all of the elements f_p of the vector-valued function $\mathbf{f}$ are evaluated on the same data set $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, then the Gram matrix can be expressed as $\mathbf{C}(\mathbf{X}, \mathbf{X}) = \mathbf{B} \otimes c(\mathbf{X}, \mathbf{X})$ where $\otimes$ denotes the Kronecker product. Due to properties of the Kronecker, its inverse can then be computed as $\mathbf{C}(\mathbf{X}, \mathbf{X})^{-1} = \mathbf{B}^{-1} \otimes c(\mathbf{X}, \mathbf{X})^{-1}$. It is straightforward to show that similar expressions can be obtained for the multi-output analogues of the kernel mean: $\Pi[\mathbf{C}(\cdot, \mathbf{X})] = \mathbf{B} \otimes \Pi[c(\cdot, \mathbf{X})] = \mathbf{B} \otimes \big(\int_{\mathcal{X}} c(\mathbf{x}, \mathbf{X}) \Pi(\mathrm{d}\mathbf{x})\big)$ and initial error $\Pi\Pi[\mathbf{C}] = \mathbf{B}\, \Pi\Pi[c] \;=\; \mathbf{B} \int_{\mathcal{X} \times \mathcal{X}} c(\mathbf{x}, \mathbf{x}') \Pi(\mathrm{d}\mathbf{x}) \Pi(\mathrm{d}\mathbf{x}')$. These expressions can of course be obtained in closed form whenever the kernel mean and initial error of the scalar-valued covariance function are available in closed form. This type of covariance function can lead to a lower computational cost of order $O(n^3 + P^3)$ when evaluating all f_p on the same input set and using tensor product formulations.

  A particular case of interest is the linear model of co-regionalisation where the matrix is of the form $(\mathbf{B})_{pp'} = \sum_{i=1}^{R} a_p^i a_{p'}^i$ for some $a_p^i \in \mathbb{R}$.

- An alternative is the process convolution covariance function [Ver Hoef and Barry, 1998; Higdon, 2002; Álvarez and Lawrence, 2011], which models integrands $f_1, \dots, f_P$ as blurred transformations of R different underlying functions. It is given by:

$$(\mathbf{C}(\mathbf{x},\mathbf{x}'))_{p,p'} \quad = \quad c_{p,p'}(\mathbf{x},\mathbf{x}') + c_{w_p}(\mathbf{x},\mathbf{x}')\delta_{p,p'},$$

where $\delta_{pp'} = 1$ if $p = p'$ and 0 else. Here there are two parts of the kernel, first $c_{p,p'} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which encodes correlation across integrands and is defined as:

$$c_{p,p'}(\mathbf{x},\mathbf{x}') \quad = \quad \sum_{i=1}^{R} \int_{\mathcal{X}} G_p^i(\mathbf{x}-\mathbf{z}) \int_{\mathcal{X}} G_{p'}^i(\mathbf{x}'-\mathbf{z}') c_i(\mathbf{z},\mathbf{z}') \mathrm{d}\mathbf{z}' \mathrm{d}\mathbf{z}.$$

The second part is $c_{w_p} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and it models the part which is not shared accross integrands. $G_p^i : \mathcal{X} \to \mathbb{R}$ is a blurring kernel which is a continuous function either having compact support or being square integrable. Note that the term "blurring kernel" does not mean the function is a reproducing kernel. Notice that taking $G_p^i(\mathbf{x}-\mathbf{z}) = a_p^i \delta(\mathbf{x}-\mathbf{z})$ (where $\delta(\cdot)$ represents a Dirac function) reduces this covariance function to the linear model of coregionalisation.

It is common to combine covariance functions, by summing them ($\mathbf{C}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q} \mathbf{C}_q(\mathbf{x},\mathbf{x}')$) in order to obtain more flexible models. The use of both of these covariances will be explored in the experiments.

4.1.2 Convergence for Priors with Separable Covariance Functions

In this section, we begin by exploring properties of multi-output BQ as an optimally-weighted quadrature algorithm in some vector-valued RKHS $\mathcal{H}_{\mathbf{C}}$. Denote the norm and inner product of $\mathcal{H}_{\mathbf{C}}$ by $\|\cdot\|_{\mathbf{C}}$ and $\langle\cdot,\cdot\rangle_{\mathbf{C}}$ respectively. Vector-valued RKHSs were extensively studied in Pedrick [1957]; Micchelli and Pontil [2005]; Carmeli et al. [2006, 2010]; De Vito et al. [2013], and generalise the notion of a RKHS to vector-valued functions. In the vector-valued case, there is an extension of the Moore-Aronszajn theorem which guarantees a one-to-one correspondence between the RKHS $\mathcal{H}_{\mathbf{C}}$ and the kernel $\mathbf{C}$. Furthermore, Theorem 3.1 in Micchelli and Pontil [2005] shows that the minimiser of the variational problem:

$$\min_{\mathbf{h}\in\mathcal{H}_{\mathbf{C}}} \left\{ \|\mathbf{h}\|_{\mathbf{C}}^2 \; : \; \mathbf{h} : \mathcal{X} \to \mathbb{R}^P, \mathbf{h}(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i) \; \forall \mathbf{x}_i \in \mathbf{X} \right\}$$

takes the form of $\mathbf{g}_n$, the multi-output GP with mean and covariance given in Proposition 7. We can therefore extend the result from Chapter 2 which shows that $\hat{\Pi}_{\mathrm{BQ}}[f_p]$ is an optimally weighted quadrature rule. In the multi-output case, we give a result for the WCE of individual integrands:

$$e(\hat{\Pi};\Pi,\mathcal{H}_{\mathbf{C}},p) \quad = \quad \sup_{\|\mathbf{f}\|_{\mathbf{C}}\leq 1} \left|\Pi[f_p]-\hat{\Pi}[f_p]\right|.$$

Proposition 8 (**Multi-output BQ is optimally-weighted**). *For a fixed point set* $\mathbf{X}$*, denote by* $\hat{\Pi}[\mathbf{f}]=\mathbf{W}^\top\mathbf{f}(\mathbf{X})$ *any quadrature rule for the vector-valued function* $\mathbf{f}=(f_1,\ldots,f_P)$ *and by* $\mathbf{W}_{BQ}$ *the weights of the multi-output BQ rule with prior mean* $\mathbf{0}$ *and covariance function* $\mathbf{C}$*. Assume that all integrands are evaluated on the same point set* $\mathbf{X}$*. Then,* $\forall p=1,\ldots,P$*:*

$$\mathbf{W}_{BQ} \quad = \quad \operatorname*{arg\,min}_{\mathbf{W}\in\mathbb{R}^{nP\times P}} e(\hat{\Pi};\Pi,\mathcal{H}_{\mathbf{C}},p).$$

In specific cases, it is also possible to characterise the rate of convergence of the worst-case error for each element f_p. This is for example the case when all integrands are evaluated on the same point set $\mathbf{X}$ and the prior is based on a separable covariance function.

Theorem 13 (**Consistency of multi-output BQ with separable covariance function**). *Suppose we want to approximate* $\Pi[\mathbf{f}]$ *for some* $\mathbf{f}:\mathcal{X}\to\mathbb{R}^D$ *and* $\hat{\Pi}_{BQ}[\mathrm{f}]$ *is the multi-output BQ rule with the covariance function* $\mathbf{C}(\mathbf{x},\mathbf{x}')=\mathbf{B}c(\mathbf{x},\mathbf{x}')$ *for some positive definite* $\mathbf{B}\in\mathbb{R}^{D\times D}$ *and scalar-valued kernel* $c:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$*. Then,* $\forall p=1,\ldots,P$*, we have:*

$$e(\hat{\Pi}_{BQ};\Pi,\mathcal{H}_{\mathbf{C}},p) \quad = \quad O\left(e(\hat{\Pi}_{BQ};\Pi,\mathcal{H}_c)\right).$$

A small extension to sums of seperable covariance functions can also be useful in applications.

Proposition 9 (**Consistency for multi-output BQ with sums of covariance functions**). *Suppose that* $\mathbf{C}(\mathbf{x},\mathbf{x}')=\sum_{q=1}^{Q}\mathbf{C}_q(\mathbf{x},\mathbf{x}')$*. Then:*

$$e(\hat{\Pi}_{BQ};\Pi,\mathcal{H}_{\mathbf{C}},p) \quad = \quad \operatorname*{arg\,max}_{q\in\{1,\ldots,Q\}} O\left(e(\hat{\Pi}_{BQ};\Pi,\mathcal{H}_{\mathbf{C}_q},p)\right).$$

It is interesting to note that the point estimator in this case will be the same as that of uni-output BQ. However the posterior variance on each integrands will usually be smaller, but of the same order, in the multi-output BQ. This can be explained intuitively by the fact that, when adding a new integrand, we can only gain by a constant factor since we always evaluate the functions at the same input points. In fact the proof of Theorem 13 provides an expression for this improvement factor (in terms of WCE) for any integrand f_p, and this depends explicitly on its correlation with the other functions: $|\sum_{i,j=1}^{P}(\mathbf{B}^{-1})_{ij}\mathbf{B}_{ip}\mathbf{B}_{jp}|$. From a practitioner's viewpoint, this can clearly be used to balance the value of using several integrands to reduce the uncertainty on the error with the additional computational cost incurred.

We now give a result in the misspecified setting. This result assumes that the points cover the space well and that the function f is assumed to be smoother than it is. In this case, it is still possible to recover the optimal convergence rate:

Theorem 14 (Consistency of multi-output BQ with seperable covariance in misspecified settings). *Let c_α be the kernel of some RKHS norm-equivalent to a Sobolev space on some domain $\mathcal{X}$ with Lipschitz boundary*[1] *and satisfying an interior cone condition. Consider the BQ rule $\hat{\Pi}_{BQ}[\mathbf{f}]$ based on a separable covariance function $\mathbf{C}_\alpha(\mathbf{x},\mathbf{x}') = \mathbf{B}c_\alpha(\mathbf{x},\mathbf{x}')$. Assume all integrands are evaluated on the same point set $\mathbf{X}$ corresponding to a quasi-uniform grid on $\mathcal{X}$, and suppose that $\mathbf{f} \in \mathcal{H}_{\mathbf{C}_\beta}$ where $\frac{d}{2} \leq \beta \leq \alpha$. Then, $\forall p = 1,\ldots,P$:*

$$\left|\Pi[f_p] - \hat{\Pi}_{BQ}[f_p]\right| \quad = \quad O\left(n^{-\frac{\beta}{d}+\epsilon}\right),$$

for some $\epsilon > 0$.

This last theorem demonstrate that the method is rate adaptive as long as we choose a covariance function which is too smooth. This however also demonstrates a drawback of the method: if one of the integrands is rough but all other are smooth, then the worst-case error could potentially converge slowly for all of them

Before moving on to the numerical experiments, it is important to highlight some limitations of our theoretical analysis. Most notably, the assumption that all integrands are evaluated at the same points is a very strong requirement, which will often not hold in practice. In fact, it may not even be desirable, since in this case the individual estimates of the integrals are identical to using a uni-output BQ rule. The only advantage therefore come from a reduced WCE, and hence a refined estimate of our epistemic uncertainty regarding the value of these integrals.

4.1.3 Numerical Experiments

We now proceed to illustrate the performance of multi-output BQ on a range of toy problems and real-world applications in order to illustrate the advantages, but also the limitations, of the methodology.

[1] Domains with Lipschitz boundaries are formally introduced in Appendix A.1.

Prior Specification

One of the main challenges with multi-output BQ is the selection of appropriate hyperparameters. In this section, we consider multi-output BQ with covariance function $\mathbf{C}$ which is parameterised by $\gamma = (\gamma_1, \ldots, \gamma_l) \in \mathbb{R}^l$. To optimise these parameters, we propose to use an empirical Bayes approach and maximise the log-marginal likelihood:

$$l\left(\gamma\right) \;=\; -\frac{1}{2}\mathbf{f}(\mathbf{X})^\top \mathbf{C}(\mathbf{X},\mathbf{X})^{-1}\mathbf{f}(\mathbf{X}) - \frac{1}{2}\log|\mathbf{C}(\mathbf{X},\mathbf{X})| - \frac{nP}{2}\log(2\pi).$$

This can be efficiently optimised by making use of gradients, given by:

$$\frac{\partial l\left(\gamma\right)}{\partial \gamma_i} \;=\; \frac{1}{2}\mathbf{f}(\mathbf{X})^\top \mathbf{C}(\mathbf{X},\mathbf{X})^{-1}\frac{\partial \mathbf{C}(\mathbf{X},\mathbf{X})}{\partial \gamma_i}\mathbf{C}(\mathbf{X},\mathbf{X})^{-1}\mathbf{f}(\mathbf{X}) - \frac{1}{2}\mathrm{Tr}\left(\mathbf{C}(\mathbf{X},\mathbf{X})^{-1}\frac{\partial \mathbf{C}(\mathbf{X},\mathbf{X})}{\partial \gamma_i}\right).$$

for all $i = 1, \ldots, l$. Clearly, this is just one option for parameter selection, and the reader is referred to Chapter 3 for alternatives to empirical Bayes.

Multi-fidelity modelling

Consider the problem of integrating some function $f^{\text{high}} : \mathcal{X} \to \mathbb{R}$ representing some complex engineering model of interest. We may be interested in such integrals for a variety of tasks, including statistical inference or optimisation. These models usually require the simulation of underlying physical systems, which can make each evaluation prohibitively expensive and will therefore limit the number of integrand evaluations n to the order of tens or hundreds.

To tackle this issue, multi-fidelity modelling proposes to build cheap, but less accurate, approximations $f_1^{\text{low}}, \ldots, f_{P-1}^{\text{low}} : \mathcal{X} \to \mathbb{R}$ to the model of interest f^{high}. The cheaper models can then be used to accelerate computation for the task of interest. Several approaches are possible. One could for example use surrogate models (e.g. support vector machines, GPs or neural networks), projection-based models (Krylov subspace or reduced basis methods) or a models where the underlying physics is simplified; see Peherstorfer et al. [2016a] for an overview.

In this section, we consider the problem of numerical integration in such a multi-fidelity setup. Two related methods for MC estimation are the multi-fidelity MC estimator [Peherstorfer et al., 2016a] and the multilevel MC of Giles [2015], both of which are based on control variate identities.

We approach this problem with multi-output BQ on the vector-valued function $\mathbf{f} = (f^{\text{high}}, f_1^{\text{low}}, \ldots, f_{P-1}^{\text{low}})^\top$. Note that multi-output GPs were already proposed for multi-fidelity modelling in [Perdikaris et al., 2016; Raissi and Karniadakis, 2016; Parussini et al., 2017], and we extend their methodologies to the task of numerical integration. We consider two toy problems from the work of Raissi and Karniadakis [2016] to highlight some of the advantages and disadvantages of our methodology:

1. A step function on $\mathcal{X} = [0, 2]$:

$$f_1^{\text{low}}(x) \;=\; \begin{cases} 0, \; 0 \leq x \leq 1 \\ 1, \; 1 < x \leq 2 \end{cases} \qquad f^{\text{high}}(x) \;=\; \begin{cases} -1, \; 0 \leq x \leq 1 \\ 2, \; 1 < x \leq 2 \end{cases}$$

2. The Forrester function with Jump on $\mathcal{X} = [0, 1]$:

$$\begin{aligned} f_1^{\text{low}}(x) &= \begin{cases} (\frac{3}{2}x - \frac{1}{2})^2 \sin(12x - 4) + 10(x - 1), \; x \leq \frac{1}{2} \\ 3 + (\frac{3}{2}x - \frac{1}{2})^2 \sin(12x - 4) + 10(x - 1), \; x > \frac{1}{2} \end{cases} \\ f^{\text{high}}(x) &= \begin{cases} \;\; 2f^{\text{low}}(x) - 20(x - 1), \; x \leq \frac{1}{2} \\ 4 + 2f^{\text{low}}(x) - 20(x - 1), \; x > \frac{1}{2} \end{cases} \end{aligned}$$

Of course, the theory developed in the previous section does not apply to this case since we are interested in evaluating the low-fidelity integrand more frequently than the high-fidelity integrand. An extension of the theory which fits this setting is reserved for future work.

The functions considered and the corresponding posteriors with credible intervals are given in Figure 4.1. The uni-output and multi-output BQ estimates for integration of these functions against a uniform measure Π are given in the table in Figure 4.2. In both cases, 20 equidistant points are used, with point number $4, 10, 11, 14$ and 17 used to evaluate the high fidelity model and the others used for the low fidelity model. The choice of hyperparameters was made using empirical Bayes for both the seperable and process convolution covariances.

Note that both of these problems are challenging for several reasons. Firstly, due to their discontinuity, the integrands are not in the RKHS $\mathcal{H}_{\mathbf{C}}$ corresponding to the covariance function $\mathbf{C}$ used in the multi-output BQ prior. More concerningly, the problems are misspecified in the sense that the true function is not even in the support of the prior. It is therefore difficult to interpret the posterior distribution on $\Pi[\mathbf{f}]$, and we end up with credible intervals which are too wide. This is for example illustrated in the values of the posterior variance for the high-fidelity Forrester function.

Secondly, in each case, the high and low-fidelity models are defined on dif-

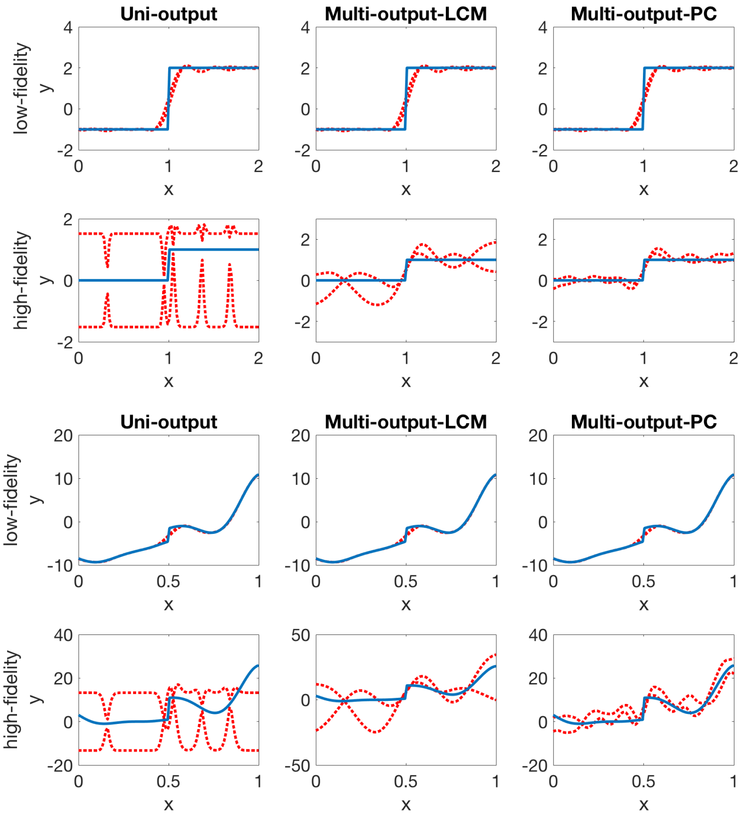

Figure 4.1: *Test functions and Gaussian process interpolants in multi-fidelity modelling.* Plot of the Step function (top) and Forrester function (bottom) in blue with GP 95% credible intervals in red. The plots on the left correspond to uni-output BQ, the plots in the middle to multi-output BQ with the linear co-regionalisation model and the plots on the right to multi-output BQ with process convolution covariance.

ferent scales and so require tuning of several kernel hyperparameters. This can of course make it challenging for multi-output BQ since the number of function evaluations n is small and the empirical Bayes performance will tend to be inefficient in those cases.

Despite these two issues, it is interesting to note that both of the multi-output BQ methods manage to significantly outperform uni-output BQ in terms of point estimate, as the sharing of data allows the multi-output models to better represent the main trends in the functions. Furthermore, the multi-output BQ does not suffer from the issues of overconfident posterior credible intervals present in uni-output BQ. To see this, contrast for example the posterior variances for the high-fidelity step function. The process convolution prior allows for much more

Model	BQ	LMC-BQ	PC-BQ
Step (l)	0.024 (0.223)	0.021 (0.213)	0.016 (0.516)
Step (h)	0.405 (0.03)	0.09 (0.091)	0.036 (0.155)
For. (l)	0.076 (4.913)	0.076 (4.951)	0.075 (33.954)
For. (h)	3.962 (3.984)	2.856 (27.01)	1.063 (63.801)

Figure 4.2: *Uni-output and multi-output Bayesian quadrature estimates for multi-fidelity modelling.* Performance of uni-output BQ and multi-output BQ with linear model of co-regionalisation kernel (LMC-BQ) and process convolution kernel (PC-BQ) on the step function (Step) and the Forrester function with jump (For.) in the low fidelity (l) and high fidelity (h) cases. The values given are absolute errors with the variance in brackets.

complex functions, which likely explains that it provides significant gains over the linear co-regionalisation model.

Global illumination

Our second application of multi-output BQ revisits the global illumination example from Chapter 3. We follow the setup previously described and consider the problem as $\Pi[f^{\omega_0}] = \int_{\mathbb{S}^2} f^{\omega_0}(\omega_i)\Pi(\mathrm{d}\omega_i)$ where Π is the uniform measure on $\mathbb{S}^2$, and $f^{\omega_0}(\omega_i) = L_i(\omega_i)\rho(\omega_i, \omega_0)[\omega_i \cdot \omega_0]_+$ is a function which can be evaluated by making a call to an environment map (which we consider to be a black box). One scenario which is common in these type of problems is to look at an object from different angles ω_0, with the camera moving. In this case, it is reasonable to assume that the different integrands f^{ω_0} will be very similar when the difference in the angle ω_0 is small, and it is therefore natural to consider jointly estimating their integrals. In the experiments we consider $f_1, \ldots, f_5$ on a great circle of the sphere at intervals determined by an angle of 0.005π.

We therefore consider two-output and five-output BQ with different IID realisations $\mathbf{X}_1, \ldots, \mathbf{X}_5$ from the uniform measure Π. We propose to use a separable covariance with scalar-valued RKHS $\mathcal{H}_c$ being a Sobolev space of smoothness $\frac{3}{2}$ over $\mathbb{S}^2$: $c(\mathbf{x}, \mathbf{x}') = \frac{8}{3} - \|\mathbf{x} - \mathbf{x}'\|_2^2$. For the matrix B representing the covariance between outputs, we propose to make this covariance proportional to the difference in angle at which the camera looks at the object. In particular we choose $(B)_{ij} = \exp(\omega_i \cdot \omega_j - 1)$ for simplicity. This could be generalised to include a lengthscale and amplitude hyperparameter inferred by empirical Bayes, however this would most likely require a larger value of n.

Results for integration error are given in Figure 4.3. As noticed, the inte-

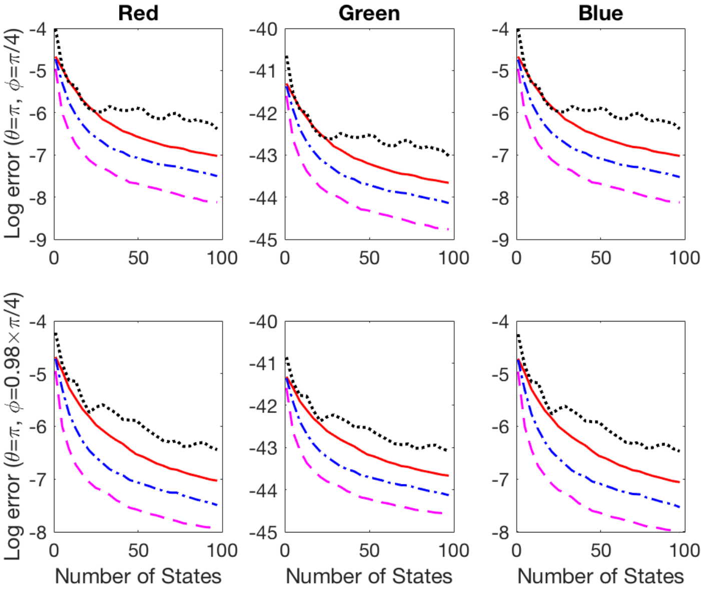

Figure 4.3: *Uni-output and multi-output Bayesian quadrature estimates in the global illumination problem.* Plot of error estimates for f_1 (top) and f_2 (bottom), in the case of the red, green and blue channels. The log-error is plotted for uni-output BQ (red), two-output and five-output BQ based on the linear model of co-regionalisation (blue and magenta respectively) and standard MC (dotted black).

gration error (for a fixed number of evaluations n of each integrand) is significantly reduced by increasing the number of outputs P. Since the experiments use different point sets for each integrand, it is reasonable to assume that the convergence rate obtained in practice will be at least as good as that for identical point sets.

Proposition 10 (**Consistency of multi-output BMC with separable covariance function on the sphere**)**.** *Let $\mathcal{X}$ be the sphere $\mathbb{S}^2$ and suppose all integrands are evaluated on the same point set $\mathbf{X}$ which consists of IID realisations from the uniform measure on $\mathcal{X}$. Furthermore, assume $\mathbf{C}$ is a separable kernel with c defined above. Then:*

$$e(\hat{\Pi}_{BQ}; \Pi, \mathcal{H}_{\mathbf{C}}, p) \;=\; O_P\left(n^{-\frac{3}{4}}\right).$$

The same rate with improved rate constant was observed in Chapter 3 when using QMC point sets, and similar gains could be obtained in this multi-output case.

Before concluding this section, we note that there a significant potential further gains for the use of multi-output BQ for this application. Similar integration problems need to be computed for three colors in every pixel of an image, and for every image in a video. This is challenging computationally and limits the use of

MC methods to a few dozen points. Designing specific matrix-valued kernels for this application could provide enormous gains since we usually end up with thousands of correlated integrands. Furthermore, the weights only depend on the choice of kernel and not on function values. They could therefore be precomputed off-line and later used in real-time in parallel at no more computational cost than MC weights.

Conclusion and Future work

There are several potential extensions of multi-output BQ which we reserve for future work. One important question remaining is that of the choice of sampling distribution. In the multi-output case, the problem is even more complex than in the uni-output case due to the interaction between the different integration problems. However, the literature on the design of experiments for co-kriging/multi-output GPs may provide some useful algorithm, and the use of more advanced sampling distributions will certainly provide significant gains.

The multi-output BQ methodology has the potential to impact a wide range of applications domains, the most obvious being areas where co-kriging/multi-output GPs are already being used. Other areas also include multivariate time series analysis and time-evolving computer models Conti and O'Hagan [2010], model comparison in Bayesian statistics or even the development of new probabilistic numerical methods.

4.2 Efficient Point Selection Methods I: The Frank-Wolfe Algorithm

The remainder of this chapter studies efficient sampling strategies for uni-output BQ. In particular, this section studies BQ from the point of view of experimental design, which has been shown to be promising in previous work Osborne et al. [2012]; Huszár and Duvenaud [2012]; Gunter et al. [2014]. Design of experiments for GP models is an active area of research [Krause et al., 2008; Beck and Guillas, 2016] with much relevance to the problem at hand. In this section, we propose two novel algorithms specialised to the numerical integration setting. These are based on optimisation routines to sequentially minimise the posterior variance on $\mathbb{V}[\Pi[g_n]]$.

Our starting point is recent work by Chen et al. [2010]; Bach et al. [2012], who cast the design of quadrature rules as a problem in convex optimisation that can be solved using the Frank-Wolfe (FW) algorithm. This algorithm is combined with the optimal weights of BQ, and we prove that exponential rates hold for posterior consistency under a finite-dimensional RKHS assumption. The methodology is explored in simulations and also applied to a challenging model selection problem from cellular biology, where numerical error could lead to misallocation of expensive resources.

4.2.1 Frank-Wolfe Bayesian Quadrature

The Frank-Wolfe (FW) algorithm, also called the conditional gradient algorithm, is a convex optimisation method introduced in Frank and Wolfe [1956] designed for problems of the form $\arg\min_{g \in \mathcal{G}} J(g)$ where the function $J : \mathcal{G} \to \mathbb{R}$ is convex and continuously differentiable. A particular case of interest in this section will be when the domain $\mathcal{G}$ is a compact and convex space of functions, as recently investigated in Jaggi [2013].

At each iteration i, the FW algorithm computes a linearisation of the objective function J at the previous state $g_{i-1} \in \mathcal{G}$ along its gradient $(DJ)(g_{i-1})$ and selects an 'atom' $\bar{g}_i \in \mathcal{G}$ that minimises the inner product taken between a state g and $(DJ)(g_{i-1})$. The new state $g_i \in \mathcal{G}$ is then a convex combination of the previous state g_{i-1} and of the atom $\bar{g}_i$. This convex combination depends on a step size ρ_i which is predetermined and different versions of the algorithm may have different step size sequences. The various steps of this algorithm will be formalised below.

Quadrature Rules from the Frank-Wolfe Algorithm

Recall from the previous chapter that approximating the kernel mean $\Pi[c(\cdot, \mathbf{x})]$ is equivalent to choosing a quadrature rule which will minimise the WCE in the RKHS $\mathcal{H}_c$. Recently, this insight led Bach et al. [2012] to frame integration as a FW optimisation problem whose objective function is minimised when $\mu(\Pi)$ is perfectly approximated. In particular, the optimisation domain $\mathcal{G} \subseteq \mathcal{H}_c$ is a space of functions and the objective function is given by half the WCE squared:

$$J(g) \;=\; \frac{1}{2}\big\|g - \Pi[c(\cdot, \mathbf{x})]\big\|^2_{\mathcal{H}_c}. \tag{4.3}$$

In this functional approximation setting, minimisation of J is carried out over $\mathcal{G} = \mathcal{M}$, where $\mathcal{M}$ denotes the marginal polytope of the RKHS $\mathcal{H}_c$. The marginal polytope $\mathcal{M}$ is defined as the closure of the convex hull of $\{c(\cdot, \mathbf{x})\}_{\mathbf{x} \in \mathcal{X}}$, so that in particular the kernel mean is an element of $\mathcal{M}$. Assuming as in Lacoste-Julien et al. [2015] that $c(\cdot, \mathbf{x})$ is uniformly bounded in feature space (i.e. $\exists R > 0 : \forall \mathbf{x} \in \mathcal{X}$, $\|c(\cdot, \mathbf{x})\|_{\mathcal{H}_c} \leq R$), then $\mathcal{M}$ is a closed and bounded set and can be optimised over.

In order to formalise the algorithm, we introduce the Fréchet derivative of J, denoted DJ, such that for $\mathcal{H}_c^*$ being the dual space of $\mathcal{H}_c$, we have the unique map $DJ : \mathcal{H}_c \to \mathcal{H}_c^*$ such that for each $g \in \mathcal{H}_c$, $(DJ)(g)$ is the function mapping $h \in \mathcal{H}_c$ to $(DJ)(g)(h) = \big\langle g - \Pi[c(\cdot, \mathbf{x})], h \big\rangle_{\mathcal{H}_c}$. We also introduce the bilinear map $\langle \cdot, \cdot \rangle_\times : \mathcal{H}_c \times \mathcal{H}_c^* \to \mathbb{R}$ which, for $F \in \mathcal{H}_c^*$ given by $F(g) = \langle g, f \rangle_{\mathcal{H}_c}$, is the rule giving $\langle h, F \rangle_\times = \langle h, f \rangle_{\mathcal{H}_c}$. The FW algorithm at iteration i can now be summarised in the two steps below:

1. Compute a new atom: $\bar{g}_i \;=\; \text{argmin}_{g \in \mathcal{G}} \big\langle g, (DJ)(g_{i-1}) \big\rangle_\times$.
2. Move in the direction of the new atom: $g_i \;=\; (1 - \rho_i) g_{i-1} + \rho_i \bar{g}_i$.

A particular advantage of the FW algorithm is that it returns 'sparse' solutions which are linear combinations of the atoms $\{\bar{g}_i\}_{i=1}^n$ [Bach et al., 2012]. This property wouldn't necessarily hold for any optimisation method and, as shown below, is particularly convenient as it leads to a weighted estimate for the kernel mean:

$$g_n \;=\; \sum_{i=1}^n \Big(\prod_{j=i+1}^n \big(1 - \rho_{j-1}\big) \rho_{i-1} \Big) \bar{g}_i \;:=\; \sum_{i=1}^n w_i^{\text{FW}} \bar{g}_i \;\;=\;\; \hat{\Pi}_{\text{FW}}[c(\cdot, \mathbf{x})],$$

where by default $\rho_0 = 1$ which leads to all $w_i^{\text{FW}} \in [0,1]$ when $\rho_i = 1/(i+1)$. Since minimisation of a linear function can be restricted to extreme points of the domain, the atoms will be of the form $\bar{g}_i = c(\cdot, \mathbf{x}_i^{\text{FW}})$ for some $\mathbf{x}_i^{\text{FW}} \in \mathcal{X}$. The minimisation in g over $\mathcal{G}$ therefore becomes a minimisation in $\mathbf{x}$ over $\mathcal{X}$ and this algorithm therefore provides us with quadrature points. Using the reproducing property, we can show that the FW estimate is indeed a quadrature rule:

$$\hat{\Pi}_{\text{FW}}[f] \;=\; \Big\langle f, \sum_{i=1}^n w_i^{\text{FW}} \bar{g}_i \Big\rangle_{\mathcal{H}_c} \;=\; \sum_{i=1}^n w_i^{\text{FW}} \big\langle f, c(\cdot, \mathbf{x}_i^{\text{FW}}) \big\rangle_{\mathcal{H}_c} \;=\; \sum_{i=1}^n w_i^{\text{FW}} f(\mathbf{x}_i^{\text{FW}}).$$

As a side effect, the FW algorithm also provides a weighted empirical measure $\hat{\Pi}_{\text{FW}} = \sum_{i=1}^n w_i^{\text{FW}} \delta(\mathbf{x}_i^{\text{FW}})$.

In summary, at each iteration i, the FW algorithm hence selects a design point $\mathbf{x}_i^{\text{FW}} \in \mathcal{X}$ which induces an atom $\bar{g}_i$ and gives us an approximation $\hat{\Pi}_{\text{FW}}[c(\cdot, \mathbf{x})]$ of the kernel mean $\Pi[c(\cdot, \mathbf{x})]$.

Step 1: Selection of new quadrature points

We now highlight in detail the first step, which at iteration i, consists of choosing a new point $\mathbf{x}_i^{\text{FW}}$. Let $\{w_l^{(i)}\}_{l=1}^{i-1}$ denote the FW weights assigned to each of the previous design points $\{\mathbf{x}_l^{\text{FW}}\}_{l=1}^{i-1}$ at the previous iteration. The choice of new design point is done by computing the derivative of the objective function $J(g_{i-1})$ and finding the point $\mathbf{x}^*$ which minimises the inner product $\arg\min_{g \in \mathcal{G}} \big\langle g, (DJ)(g_{i-1}) \big\rangle_\times$.

To do so, we need to obtain an equivalent expression of the minimisation of the linearisation of J (denoted DJ) in terms of kernel values and evaluations of the kernel mean $\Pi[c(\cdot, \mathbf{x})]$. Since minimisation of a linear function can be restricted to extreme points of the domain, we have that

$$\begin{aligned}
\arg\min_{g \in \mathcal{G}} \big\langle g, (DJ)(g_{i-1}) \big\rangle_{\times} &= \arg\min_{\mathbf{x} \in \mathcal{X}} \big\langle c(\cdot, \mathbf{x}), (DJ)(g_{i-1}) \big\rangle_{\times} \\
&= \arg\min_{\mathbf{x} \in \mathcal{X}} \big\langle c(\cdot, \mathbf{x}), g_{i-1} - \Pi[c(\cdot, \mathbf{x})] \big\rangle_{\mathcal{H}_c} \\
&= \arg\min_{\mathbf{x} \in \mathcal{X}} \Big\langle c(\cdot, \mathbf{x}), \sum_{l=1}^{i-1} w_l^{(i-1)} c(\cdot, \mathbf{x}_l) - \Pi[c(\cdot, \mathbf{x})] \Big\rangle_{\mathcal{H}_c} \\
&= \arg\min_{\mathbf{x} \in \mathcal{X}} \sum_{i=1}^{i-1} w_l^{(i-1)} \big\langle c(\cdot, \mathbf{x}), c(\cdot, \mathbf{x}_l) \big\rangle_{\mathcal{H}_c} \\
&\qquad\qquad - \big\langle c(\cdot, \mathbf{x}), \Pi[c(\cdot, \mathbf{x})] \big\rangle_{\mathcal{H}_c} \\
&= \arg\min_{\mathbf{x} \in \mathcal{X}} \sum_{l=1}^{i-1} w_l^{(i-1)} c(\mathbf{x}, \mathbf{x}_l) - \Pi[c(\cdot, \mathbf{x})](\mathbf{x}).
\end{aligned}$$

Our new design point $\mathbf{x}_i^{\mathrm{FW}}$ is therefore the point $\mathbf{x}^*$ which minimises this expression. Note that the total computational cost is $O(n^2)$ since we need to loop through all previous samples at each iteration. Note also that this equation may not be convex and may require us to make use of approximate methods to find the minimum $\mathbf{x}^*$. To do so, we sample M points (where M is large) and pick the sample which minimises the expression above. From Lacoste-Julien et al. [2015] this introduces an additive error term of size $O(M^{-1/4})$, which does not impact our convergence analysis provided that M vanishes sufficiently quickly as a function n. In all experiments at the end of this section we took n to be a few hundreds and M between $10,000$ and $50,000$ so that this error will be negligible.

It is important to note that sampling from Π is likely to be suboptimal for optimising this expression. One may be better off using another optimisation method which does not require convexity (for example, Bayesian optimisation). This would however lead to larger computational costs at each iteration.

Step 2: Selection of a step size sequence

Several choices are possible for the step size sequence, the most common of which is to have a decreasing sequence $\rho_i = 1/(i+1)$ since this will lead to equally-weighted points. However, the step size can also be chosen adaptively. An extension of the FW algorithm known as Frank Wolfe with line-search (FWLS) uses a line-search method to find the optimal step size ρ_i at each iteration:

$$\rho_i^* \quad = \quad \text{argmin}_{\rho \in [0,1]} J\left((1-\rho) g_{i-1} + \rho\, \bar{g}_i\right).$$

This leads to improved solutions, but comes at a higher computational cost. Once again, the approximation obtained by FWLS has a sparse expression as a convex combination of all the previously visited states and we obtain an associated quadrature rule. For the problem of computing integrals, this optimisation step can actually be obtained analytically.

Proposition 11 (**Optimal Frank-Wolfe line-search step size for quadrature rules**)**.** *The optimal step size sequence $\{\rho_i^*\}_{i \in \mathbb{N}}$ for minimising the objective function $J(g)$ as given in Equation 4.3 is given by:*

$$\rho_i^* \;=\; \frac{\sum_{l=1}^{i-1}\sum_{m=1}^{i-1} w_l^{(i-1)} w_m^{(i-1)} c(\mathbf{x}_l, \mathbf{x}_m) - \sum_{l=1}^{i-1} w_l^{(i-1)} \Big[c(\mathbf{x}_l, \mathbf{x}_i) + \Pi[c(\mathbf{x}_l, \cdot)] \Big] + \Pi[c(\mathbf{x}_i, \cdot)]}{\sum_{l=1}^{i-1}\sum_{m=1}^{i-1} w_l^{(i-1)} w_m^{(i-1)} c(\mathbf{x}_l, \mathbf{x}_m) - 2\sum_{l=1}^{i-1} w_l^{(i-1)} c(\mathbf{x}_l, \mathbf{x}_i) + c(\mathbf{x}_i, \mathbf{x}_i)}.$$

FWLS has theoretical convergence rates that can be stronger than standard versions of FW but has computational cost which is $O(n^3)$. The authors in Garber and Hazan [2015] provide a survey of FW-based algorithms and their convergence rates under different regularity conditions on the objective function and domain of optimisation.

Frank-Wolfe Quadratures with Optimal Weights

To combine the advantages of a Bayesian method with the efficient point-selection of the FW algorithm, we propose Frank-Wolfe Bayesian Quadrature (FWBQ) and Frank-Wolfe line-search Bayesian Quadrature (FWLSBQ):

$$\hat{\Pi}_{\text{FWBQ}}[f] := \sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{FW}}), \quad \hat{\Pi}_{\text{FWLSBQ}}[f] := \sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{FWLS}}),$$

where $\{\mathbf{x}_i^{\text{FW}}\}_{i=1}^n$ and $\{\mathbf{x}_i^{\text{FWLS}}\}_{i=1}^n$ are FW and FWLS point sets respectively. These two algorithms will combine the efficient point selection strategy of the FW (or FWLS) algorithm with the RKHS-optimal weights provided by BQ, and the uncertainty quantification properties of the Bayesian interpretation.

4.2.2 Consistency and Contraction in Finite-Dimensional Spaces

An important question is whether it is possible to provide consistency results for this algorithm. Indeed, to the best of our knowledge, there has never been any consistency result for BQ based on experimental design selection of point sets. The answer is yes, but this will be under the assumption that the integrand f belongs to a finite-dimensional RKHS $\mathcal{H}_c$. This assumption is in line with recent literature on the FW algorithm [Bach et al., 2012; Garber and Hazan, 2015; Jaggi, 2013], but is unfortunately rather restrictive for the quadrature application. Unfortunately, there are no general results for the FW or FWLS in infinite-dimensional RKHSs. See Bach et al. [2012] and Grunewalder [2018] for a detailed discussion and intuitive explanation of the issues encountered in infinite-dimensional spaces. The result below follows from the Bayesian reweighting bound (Lemma 1) and a result of Bach et al. [2012]:

Theorem 15 (**Consistency of FWBQ and FWLSBQ in finite-dimensional RKHSs**). *Assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and that $\pi(\mathbf{x}) > 0 \ \forall \mathbf{x} \in \mathcal{X}$. Let c be a reproducing kernel corresponding to a finite-dimensional RKHS $\mathcal{H}_c$, and denote by $\hat{\Pi}_{FWBQ}[f]$ and $\hat{\Pi}_{FWLSBQ}[f]$ the BQ estimators with prior covariance c based on FW (with step size $\rho_i = 1/(i+1)$ for all i) and FWLS point sets. Then $\exists C > 0$ such that:*

$$e(\hat{\Pi}_{FWBQ}; \Pi, \mathcal{H}_c) \ = \ O(n^{-1}), \qquad e(\hat{\Pi}_{FWLSBQ}; \Pi, \mathcal{H}_c) \ = \ O\left(\exp\left(-Cn\right)\right).$$

In the case of FWBQ, if $f \in \mathcal{H}_c$ and $\delta > 0$,

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \quad = \quad 1 - O_P(\exp(-C_\delta n^2)),$$

where $C_\delta > 0$ depends on δ. Similarly, for FWLSBQ, if $f \in \mathcal{H}_c$ and $\delta > 0$,

$$\mathbb{P}\{\Pi[f] - \delta < \Pi[g_n] < \Pi[f] + \delta\} \quad = \quad 1 - O_P(\exp(-C_{1,\delta} n^2 - C_{2,\delta} \exp(C_{1,\delta} n))),$$

where $C_{1,\delta}, C_{2,\delta} > 0$ depends on δ.

Even though this is not explicit in our result above, the choice of covariance function affects the convergence of the FWBQ and FWLSBQ methods. Clearly, we expect faster convergence if the function we are integrating is 'close' to the space of functions induced by our covariance function. Indeed, the covariance function specifies the geometry of the marginal polytope $\mathcal{M}$, that in turn directly influences the rate constant associated with FW optimisation.

We note that FWBQ and FWLSBQ will have a convergence rate that is atleast as good as that of FW and FWLS, but there is no guarantee in our theory on how much better this rate will be. This will all boil down to how close the

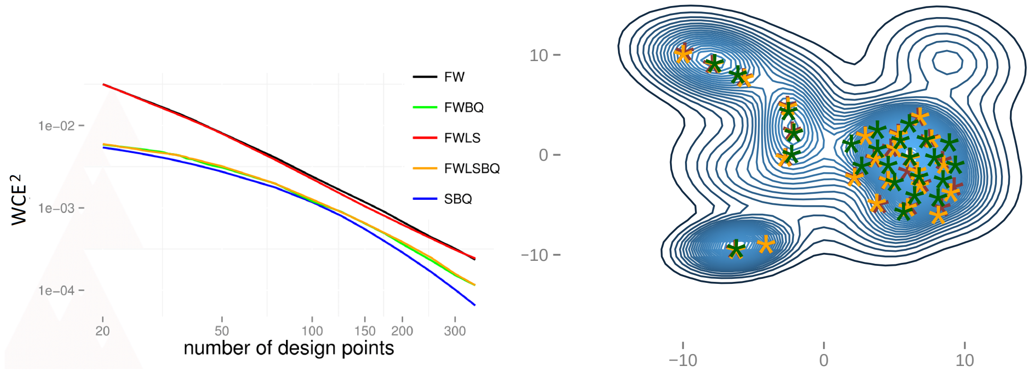

Figure 4.4: *The Frank-Wolfe algorithm for integration of test functions against a mixture of Gaussian distribution. Left:* Worst-case error for several quadrature rules. Both FWBQ and FWLSBQ are seen to outperform FW and FWLS, with sequential Bayesian quadrature (SBQ) performing best overall. *Right:* Point sets obtained from the Frank-Wolfe algorithm for a mixture of Gaussian distribution. Density of a mixture of 20 Gaussian distributions, displaying the first $n = 25$ design points chosen by FW (red), FWLS (orange) and sequential Bayesian quadrature (green).

FW/FWLS weights will be to the FWBQ/FWLSBQ weights. This would be an interesting topic of research for future work.

4.2.3 Numerical Experiments

Simulation Study

To facilitate the experiments in this section we employed a Gaussian RBF covariance function $c(\mathbf{x}, \mathbf{x}') := \lambda^2 \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|_2^2)$. This corresponds to an infinite-dimensional RKHS which is not covered by Theorem 15. Gaussian RBF covariance functions are convenient since the kernel mean $\Pi[c(\cdot, \mathbf{x})]$ is analytically tractable when Π is a mixture of Gaussian distributions (see Table 3.1).

For this simulation study, we took Π to be a 20-component mixture of 2D-Gaussian distributions. MC is often used for such distributions but has a slow convergence rate of $O_P(n^{-1/2})$. FW and FWLS are known to converge more quickly and are in this sense preferable to MC [Bach et al., 2012]. In our simulations (Figure 4.4, left), both our novel methods FWBQ and FWLSBQ decreased the WCE much faster than the FW/FWLS methods of Bach et al. [2012]. All methods use the same hyperparameters for the covariance function: an amplitude parameter of $\lambda = 1$ and a lengthscale of $\sigma = 0.8$.

The principle advantage of our proposed methods is that they reconcile the-

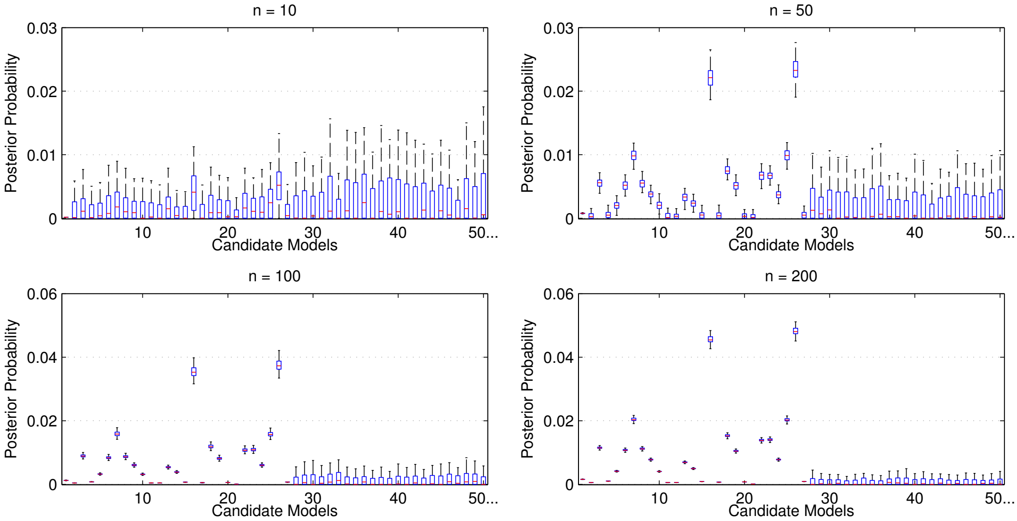

Figure 4.5: *Quantifying numerical error in a model selection problem using Frank-Wolfe Bayesian Quadrature.* FWBQ was used to model the numerical error of each integral $p(\mathcal{M}_i|\mathbf{X})$ explicitly. For integration based on $n = 10$ design points, FWBQ tells us that the computational estimate of the model posterior will be dominated by numerical error. When instead $n = 50$ or $n = 100$ design points are used, uncertainty due to numerical error becomes much smaller but not yet small enough to determine the maximum a-posteriori estimate. This only occurs for $n = 200$.

oretical tractability with a Bayesian estimator based on the sequential optimisation of sample locations. An interesting remark is that sequential Bayesian quadrature seems to give even better performance as n increases. An intuitive explanation is that sequential Bayesian quadrature picks points to minimise the WCE whereas FWBQ and FWLSBQ only minimise an approximation of the WCE (its linearisation along DJ). In addition, the sequential Bayesian quadrature weights are optimal at each iteration, which is not true for FWBQ and FWLSBQ. We conjecture that Theorem 15 provides upper bounds on the rates of sequential Bayesian quadrature. This conjecture is partly supported by Figure 4.4 (right), which shows that sequential Bayesian quadrature selects similar design points to FW/FWLS (but weights them optimally). Note also that both FWBQ and FWLSBQ give very similar result. This is not surprising as FWLS has no guarantees over FW in infinite-dimensional RKHSs [Jaggi, 2013].

Proteomic Model Selection Problem

A topical bioinformatics application that extends recent work by Oates et al. [2014] is presented. The objective is to select among a set of candidate models $\{\mathcal{M}_i\}_{i=1}^{M}$

for protein regulation. This choice is based on a dataset $\mathbf{X}$ of protein expression levels, in order to determine a 'most plausible' biological hypothesis for further experimental investigation. Each $\mathcal{M}_i$ is specified by a vector of kinetic parameters θ_i in some Euclidean space Θ_i (full details in the supplementary materials of [Briol et al., 2015a]). The goal of this experiment is very closely related to the model selection problem for logistic regression which was presented in Chapter 3.

Recall that Bayesian model selection requires that these parameters are integrated out against a prior $p(\theta)$ to obtain marginal likelihood (or model evidence) terms $p(\mathbf{X}|\mathcal{M}_i) = \int_{\theta_i \in \Theta_i} p(\mathbf{X}|\theta, \mathcal{M}_i)p(\theta_i)\mathrm{d}\theta_i$. Our integration problem therefore consists of integrating the function $f(\theta) = p(\mathbf{X}|\theta)$ against the prior measure on parameters. In this experiment, we assume a priori that all models are equally likely ($p(\mathcal{M}_i) = 1/M$ for all $i = 1, \ldots, M$), so that the posterior over each model is given by $p(\mathcal{M}_i|\mathbf{X}) = p(\mathbf{X}|\mathcal{M}_i)/M \sum_{j=1}^{M} p(\mathbf{X}|\mathcal{M}_j) \propto p(\mathbf{X}|\mathcal{M}_i)$. Our focus here is on obtaining the maximum a-posteriori model, defined as the maximiser of the posterior model probability $p(\mathcal{M}_i|\mathbf{X})$. Numerical error in the computation of each term $p(\mathcal{M}_i|\mathbf{X})$, if unaccounted for, could cause us to return a model that is different from the true maximum a-posterior estimate and lead to the misallocation of valuable experimental resources.

The problem is quickly exaggerated when the number M of models increases, as there are more opportunities for one of the $p(\mathcal{M}_i|\mathbf{X})$ terms to be too large due to numerical error. In Oates et al. [2014], the number m of models was combinatorial in the number of protein kinases measured in a high-throughput arrays (currently on the order of 10^2 but in principle up to the order of 10^4). This led Oates et al. [2014] to deploy substantial computing resources to ensure that numerical error in each estimate of $p(\mathcal{M}_i|\mathbf{X})$ was individually controlled. As previously highlighted, the use of BQ in this setting allows for quantification of our uncertainty over the value of each of the integrals $p(\mathcal{M}_i|\mathbf{X})$. As such we can determine, on-line, the precise point in the computational pipeline when numerical uncertainty near the maximum a-posteriori estimate becomes acceptably small, and cease further computation.

The FWBQ methodology was applied to one of the model selection tasks in Oates et al. [2014]. In Figure 4.5 (top left) we display posterior model probabilities for each of $M = 352$ candidates models, where a low number ($n = 10$) of samples were used for each integral. (For display clarity only the first 50 models are shown.) In this low-n regime, numerical error introduces a second level of uncertainty that we quantify by combining the FWBQ error models for all integrals in the computational pipeline; this is summarised by a box plot (rather than a single point) for each of the models. These box plots reveal that our estimated posterior model probabilities are

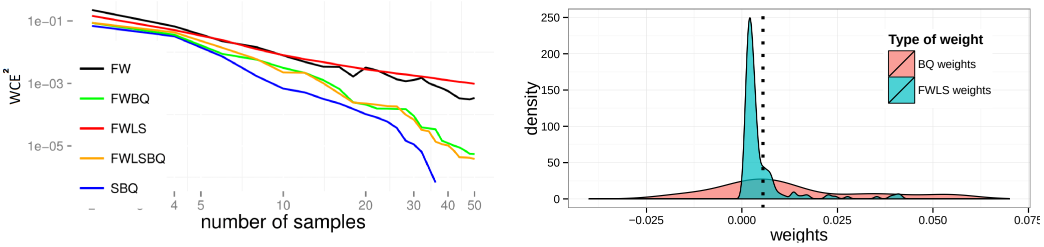

Figure 4.6: *Comparison of experimental design-based quadrature rules on the proteomics application. Left:* Value of the WCE^2 for FW (black), FWLS (red), FWBQ (green), FWLSBQ (orange) and sequential Bayesian quadrature (blue). *Right:* Empirical distribution of weights. The dotted line represent the weights of the FWLS, which has all weights $w_i = 1/n$. Note that the distribution of BQ weights ranges from -17.39 to 13.75 whereas all versions of FW have weights limited to $[0, 1]$ and have to sum to 1.

completely dominated by numerical error. In contrast, when n is increased through 50, 100 and 200 (again, see Figure 4.5), the uncertainty due to numerical error becomes negligible. At $n = 200$ we can conclude that model 26 is the true maximum a-posteriori estimate and further computations can be halted. Correctness of this result was confirmed using the more computationally intensive methods in Oates et al. [2014].

In Figure 4.6 (left) we compared the relative performance of FWBQ, FWLSBQ and sequential Bayesian quadrature on this problem. The figure shows that the BQ weights reduced the WCE by orders of magnitude relative to FW and FWLS and that sequential Bayesian quadrature converged more quickly than methods based on the FW algorithm. This is partly explained by the fact that the BQ weights are not limited to non-negative value, as seen in Figure 4.6 (right).

4.3 Efficient Point Selection Methods II: A sequential Monte Carlo sampler

The FW algorithm is clearly a useful tool for point selection in BQ. The algorithm selects points one-by-one in an adaptive manner and was shown to give good results for any value of n at which one may wish to stop. There is however no guarantee that this will be the best one can do if the total number of points n is known a-priori.

Our goal in this section will be to focus on this particular problem, and to do so we propose an extension of BIS. In Chapter 3, we showed that BMC and BIS converge at a rate determined by the ratio α/d, where α and d encode the smoothness and dimension of the integrand (see Theorem 9). However, this section highlights that the rate constant C is highly sensitive to the distribution of the random points and the choice of prior on the integrand. More importantly, it is also dependent on the number of realisations n. This section proposes a novel algorithm to approximate an optimal importance measure which takes the form of an SMC sampler with adaptive tempering.

4.3.1 Limitations of Bayesian Importance Sampling

For the default MC estimator, we have a root-mean-squared error bound:

$$\sqrt{\mathbb{E}[\hat{\Pi}_{\mathrm{MC}}[f] - \Pi[f]]^2} \quad \leq \quad \frac{C_{\mathrm{MC}}(f;\Pi)}{\sqrt{n}}, \tag{4.4}$$

where $C_{\mathrm{MC}}(f;\Pi)$ is the standard deviation of the integrand f under Π, and the expectation is with respect to the joint distribution of the points. When MCMC methods are used instead; the rate-constant $C_{\mathrm{MCMC}}(f;\Pi)$ is then related to the asymptotic variance of f under the Markov chain sample path. Considerations of computational cost place emphasis on methods to reduce the rate constant $C_{\mathrm{MC}}(f;\Pi)$. For the MC estimator, this rate constant can be made smaller via IS, where an optimal choice of importance distribution Π' is one that minimises $C_{\mathrm{MC}}(f\pi/\pi';\Pi')$, and its density is available in explicit closed-form: $\pi^*(\mathbf{x}) = |f(\mathbf{x})|\pi(\mathbf{x})/\int_{\mathcal{X}}|f(\mathbf{x})|\pi(\mathbf{x})\mathrm{d}\mathbf{x}$; see Theorem 3.3.4 in Robert and Casella [2004]. However, the root-mean-squared error remains asymptotically gated at $O(n^{-1/2})$.

Similar issues arise for BQ estimators based on Monte Carlo point sets. In particular, a trivial modification of the consistency result for BIS (with importance distribution Π') in Theorem 9 gives the following root-mean-squared error bound:

$$\sqrt{\mathbb{E}[\hat{\Pi}_{\mathrm{BIS}}[f] - \Pi[f]]^2} \quad \leq \quad \frac{C(f;\Pi')}{n^{\alpha/d-\epsilon}},$$

when $\alpha > \frac{d}{2}$ and where both the integrand f and each argument of the covariance function c admit continuous mixed weak derivatives of order α and $\epsilon > 0$ can be arbitrarily small.

One notable disadvantage of BIS is that little is known about how the rate constant $C(f;\Pi')$ depends on the choice of sampling distribution Π'. In contrast to IS, no general closed-form expression has been established for an optimal importance distribution Π' for BQ (the technical meaning of 'optimal' is defined below). Moreover, limited practical guidance is available on the selection of the sampling distribution. An exception is in [Bach, 2017], but the distribution proposed in this

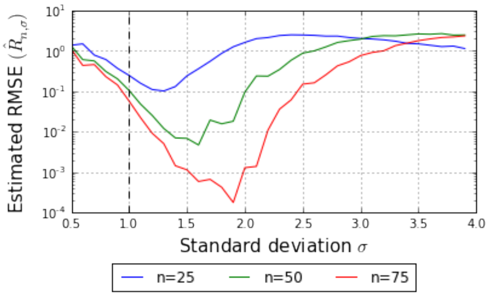

Figure 4.7: *Influence of the importance distribution in Bayesian importance sampling.* Performance of BIS with covariance function $c(x, x') = \exp(-(x - x')^2)$ on the test function $f(x) = 1 + \sin(2\pi x)$, where the target measure was $\mathcal{N}(0, 1)$, while n samples were generated from $\mathcal{N}(0, \sigma^2)$.

paper can usually not be obtained in closed-form. In applications, it is therefore usual to take $\Pi' = \Pi$. This choice is convenient but leads to estimators that are not efficient, as highlighted below.

Consider the toy problem with state space $\mathcal{X} = \mathbb{R}$, target distribution Π which is a $\mathcal{N}(0, 1)$, a single test function $f(x) = 1 + \sin(2\pi x)$ and covariance function $c(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2)$. For this problem, consider a range of sampling distributions Π' of the form $\mathcal{N}(0, \sigma^2)$ for $\sigma \in (0, \infty)$. In this case $\Pi[f] = 1$ is available in closed-form. Figure 4.7 plots an empirical estimate for the root-mean-squared error given by

$$\hat{R}_{n,\sigma} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{\Pi}_{n,m,\sigma}[f] - \Pi[f])^2},$$

where $\hat{\Pi}_{n,m,\sigma}[f]$ is the mth of M independent BIS estimates for $\Pi[f]$ based on n samples drawn from the distribution Π' with standard deviation σ. It is seen that the choice of $\sigma = 1$ which corresponds to BMC (i.e. $\Pi' = \Pi$) is suboptimal. Notice that the values of σ that minimise the root-mean-squared error are uniformly greater than $\sigma = 1$ (dashed line) and depend on the number n of samples. The intuition here is that samples from the tails of the distribution are rather informative for building the interpolant $\hat{f}$ underlying BQ. We should therefore over-sample these values via a heavier-tailed Π'. The same intuition is used for column sampling and to construct leverage scores [Mahoney, 2011; Drineas et al., 2012].

Another problem is that the integrand f will in general belong to an infinitude of Hilbert spaces, while for BIS (and in fact any BQ algorithm) a single covariance function c must be selected. This choice will in general significantly affect the performance of the BIS estimator. We extend the toy problem above based on a class

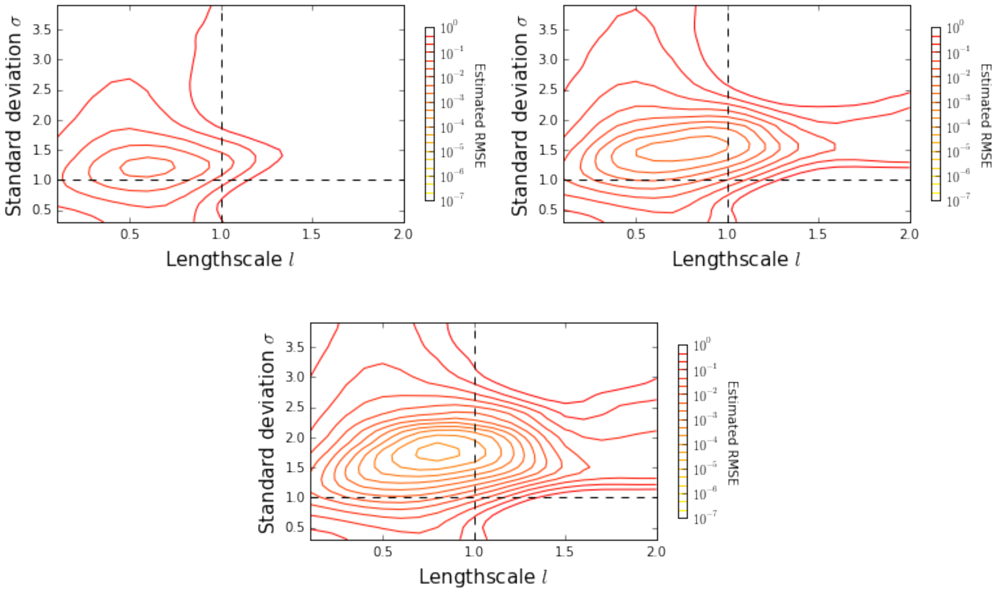

Figure 4.8: *Sensitivity of Bayesian importance sampling to the choice of both the covariance function and importance distribution.* Here the same setup as Figure 4.7 was used with $n = 25$ (top left), $n = 50$ (top right) and $n = 75$ (bottom). The Gaussian RBF covariance function $c(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/\ell^2)$ was used for various choices of parameter $\ell \in (0, \infty)$. The root-mean-squared error (over $M = 300$ repetitions) is sensitive to choice of ℓ for all choices of σ, suggesting that on-line kernel learning could be used to improve over the default choice of $\ell = 1$ and $\sigma = 1$ (dashed lines).

of Gaussian RBF covariance functions $c(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/\ell^2)$ parameterised by $\ell \in (0, \infty)$. Results showed that, for all choices of the sampling parameter σ, the root-mean-squared error of BQ is sensitive to choice of ℓ and the default choice of $\ell = 1$ is not optimal.

Results, shown in Figure 4.8, demonstrate two principles that guided the methodological development in this section. Firstly, length scales ℓ that are ‘too small’ to learn from n samples do not permit good approximations $\hat{f}$ and lead in practice to high root-mean-squared error. At the same time, if ℓ is taken to be ‘too large’ then efficient approximation at size n will also be sacrificed. This is of course well understood from a theoretical perspective and is borne out in our empirical results.

Secondly, the ‘sweet spot’, where σ and ℓ lead to minimal root-mean-squared error, will in general be quite small. However, the problem of optimal choice for σ and ℓ does not seem to become more or less difficult as n increases. This suggests that a method for selection of σ (and possibly also of ℓ) ought to be effective regardless of the number n of states that will be used.

4.3.2 Robustness of Bayesian Quadrature to the Choice of Kernel

One question which is of course of interest is whether the same issues also arise for BQ methods based on experimental design. For example, in the previous section, the selection of points was approached as a greedy optimisation problem where the WCE was minimised given the location of the previous points. This approach has demonstrated considerable success in applications, but the WCE is strongly dependent on the choice of covariance function c and the sequential optimisation approach is hence vulnerable to prior misspecification. For this reason, experimental design approaches tend to be less robust to prior misspecification than alternative approaches where the sampling mechanism does not depend on the prior.

To demonstrate this lack of robustness to misspecified priors, we once again considered integration against some measure Π which is a $\mathcal{N}(0,1)$. We focused on functions that can be well approximated using BQ rules with the covariance function $c(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|_2^2/\ell^2)$. We studied sequential Bayesian quadrature where the length scale was fixed at $\ell = 0.01$ and we consider a more regular integrand, such as that shown in Figure 4.9 (left). The location of the states obtained using sequential Bayesian quadrature and BMC are shown in Figure 4.9 (right). It is clear that the greedy selection of points is not an efficient use of computation for integration of the integrand against $\mathcal{N}(0,1)$. Of course, a bad choice of length scale parameter ℓ can in principle be alleviated by kernel learning, but this will not be robust in the case when n is very small.

More work will be required to better understand when methods such as sequential Bayesian quadrature or FWBQ can be reliable in the presence of unknown covariance function hyperparameters. Related work on subsample selection, such as leverage scores [Bach, 2013], can also be non-robust to misspecified covariance functions. The partial solution of online kernel learning requires a sufficient number n of data and is not always practicable in small-n regimes that motivate BQ.

The next section consider the selection of good importance sampling distribution for BIS estimators. Although our method also makes use of c to select Π', it reverts to $\Pi' = \Pi$ in the limit as the length scale of c is made small. In this sense, our algorithm offers more robustness to covariance function misspecification than optimisation methods.

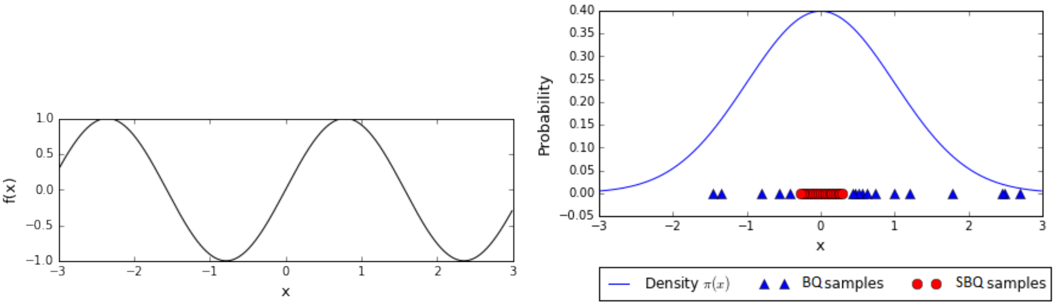

Figure 4.9: *Lack of robustness of experimental-design based quadrature rules. Left:* Toy integrand. *Right:* Sequential Bayesian quadrature does not lead to adequate placement of points when the covariance function is misspecified. Here the length scale of the covariance function was fixed to $\ell = 0.01$, points selected by sequential Bayesian quadrature are represented as red whereas points drawn from Π, as used in BIS, are shown in blue.

4.3.3 Sequential Monte Carlo Bayesian Quadrature

The main contributions of this section are twofold. First, we formalise the problem of optimal sampling for BIS as an important and open challenge in computational statistics. To be precise, our target is an optimal sampling distribution for BIS, defined as

$$\Pi^* \in \arg\min_{\Pi'} \sup_{\|f\|_{\mathcal{H}} \leq 1} \sqrt{\mathbb{E}[\hat{\Pi}_{\mathrm{BIS}}[f] - \Pi[f]]^2}. \tag{4.5}$$

for some functional class $\mathcal{H}$ to be specified and where Π' denotes the sampling distribution of $\hat{\Pi}_{\mathrm{BIS}}[f]$. In general a (possibly non-unique) optimal Π^* will depend on $\mathcal{H}$ and, unlike for IS, also on the covariance function c and the number of samples n used in the quadrature. It is also not possible to obtain it in closed form.

Second, we propose a novel and automatic method for selection of Π' that is rooted in approximation of the unavailable Π^*. The overall approach is facilitated with an efficient SMC sampler and called sequential Monte Carlo Bayesian quadrature (SMC-BQ). In brief, our method considers candidate sampling distributions of the form $\Pi' = \Pi_0^{1-t}\Pi^t$ for $t \in [0,1]$ and Π_0 a reference distribution on $\mathcal{X}$.

Our SMC sampler has several features to enable automation of the method: (i) it chooses a discretisation for t in an adaptive manner, and (ii) it uses a stopping criterion based on estimates of the root-mean-squared error. Finally, an extension is proposed for the case where kernel learning is required. Although we do not provide formal guarantees on the quality of the resulting approximation, our algorithm is motivated through several ansatzs and later shown to perform well in applications.

Similar results to those presented in Chapter 3 would not make sense in this case since we are interested in performance for a fixed value of n.

A Sequential Monte Carlo Sampler

To begin, consider the integrand f, covariance function c and number of evaluations n as fixed. The following ansatz is central to our proposed SMC-BQ method: An optimal distribution Π^* (in the sense of Equation 4.5) can be well-approximated by a distribution of the form

$$\Pi_t = \Pi_0^{1-t}\Pi^t, \quad t \in [0,1] \tag{4.6}$$

for a specific (but unknown) 'inverse temperature' parameter $t = t^*$. Here Π_0 is a reference distribution to be specified and which should be chosen to be uninformative in practice. It is assumed that all Π_t have densities who can be normalised. The motivation for this ansatz stems from the toy problem in the previous subsection, where Π is a $\mathcal{N}(0,1)$ and Π_t is a $\mathcal{N}(0,\sigma^2)$ cast with $t = \sigma^{-1}$ and Π_0 an (improper) uniform distribution on $\mathbb{R}$. In general, tempering generates a class of distributions which over-represent extreme events relative to Π. This property has the potential to improve performance for BIS, as was once again demonstrated with the toy example.

The ansatz of Equation 4.6 reduces the nonparametric sampling problem for BQ to the one-dimensional parametric problem of selecting a suitable $t \in [0,1]$. The problem can be further simplified by focusing on a discrete temperature ladder $\{t_i\}_{i=0}^T$ such that $t_0 = 0$, $t_i < t_{i+1}$ and $t_T = 1$. This reduced problem, where we seek an optimal index $i^* \in \{0,\ldots,T\}$, is still non-trivial as no closed-form expression is available for the root-mean-squared error at each candidate t_i. To construct our proposed SMC-BQ algorithm, we require a second ansatz, namely that the root-mean-squared error is convex in t and possesses a global minimum in the range $t \in (0,1)$. This second ansatz (borne out in numerical results in Figure 4.7) motivates an algorithm that begins at $t_0 = 0$ and tracks the root-mean-squared error until an increase is detected, say at t_i; at which point the index $i^* = i-1$ is fixed and used within a BQ algorithm.

To realise such an algorithm, we propose to exploit SMC samplers [Chopin, 2002; Del Moral et al., 2006], already briefly introduced in Chapter 1. Here, a set of weighted particles $\{(w_j, \mathbf{x}_j)\}_{j=1}^N$ is first obtained where $\{\mathbf{x}_j\}_{j=1}^N$ are IID realisations from Π_0 and $w_j = \frac{1}{N}$ for $j = 1,\ldots,N$. Note that we take the number of particles N to be greater than the desired number of quadrature points n. Then, at iteration i, the particle approximation to $\Pi_{t_{i-1}}$ is reweighted, resampled and subject to a

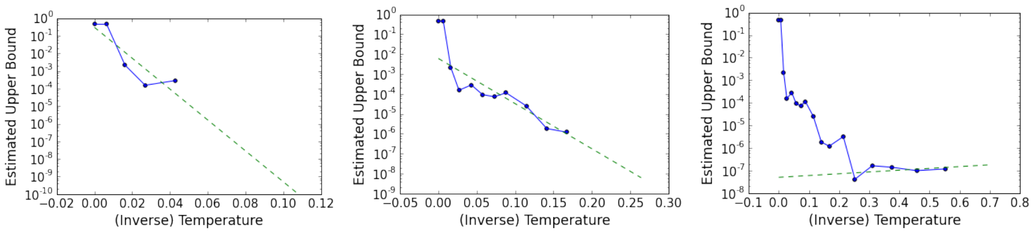

Figure 4.10: *Implementation of the stopping criterion for sequential Monte Carlo Bayesian quadrature.* A linear smoother (dashed line) was based on 5 consecutive (inverse) temperature parameters $t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}, t_i$. To begin it is required that 5 temperatures are considered (left panel). The algorithm terminates on the first occasion when the linear smoother takes a positive gradient (right panel).

Markov transition, to deliver a particle approximation $\{(w_j', \mathbf{x}_j')\}_{j=1}^N$ to Π_{t_i}. Resampling occurs when the effective sample size, $\|\mathbf{w}\|_2^{-2}$ drops below a fraction ρ of the total number N of particles. In this work we took $\rho = 0.95$ which is a common default.

At iteration i, a subset of size n is drawn (without replacement) from the unique elements in $\{\mathbf{x}_j'\}_{j=1}^N$, from the particle approximation to Π_{t_i}, and proposed for use in BQ. This ensures that covariance matrices have full rank. It does not introduce bias into BQ, since in general Π' need not equal Π. A criterion, defined below, is used to determine whether the resultant BQ error has increased relative to $\Pi_{t_{i-1}}$. If this is the case, then the distribution $\Pi_{t_{i-1}}$ from the previous iteration is taken for use in BQ. Otherwise the algorithm proceeds to t_{i+1} and the process repeats. In the degenerate case where the root-mean-squared error has a minimum at t_T, the algorithm defaults to standard BQ with $\Pi' = \Pi$.

Stopping criterion for the Sequential Monte Carlo sampler

The SMC-BQ algorithm is designed to track the root-mean-squared error as t is increased. However, the root-mean-squared error is not available in closed form. We now derive a tight upper bound on the root-mean-squared error that is used as a stopping criterion. Recall the Cauchy-Schwarz upper bound on the integration error given in Equation 3.5 in Chapter 3. At each iteration of the SMC algorithm, it can be adapted to obtain: $|\hat{\Pi}[f] - \Pi[f]| \leq e(\hat{\Pi}_{\text{BQ}}; \Pi, \mathcal{H}_c)\|f\|_{\mathcal{H}_c}$ where $\hat{\Pi}_{\text{BQ}}[f] = \sum_{j=1}^n w_j^{\text{BQ}} f(\mathbf{x}_j)$. This motivates the following upper bound on the mean-squared error:

$$\mathbb{E}[(\hat{\Pi}_{\text{BQ}}[f] - \Pi[f])^2] \leq \underbrace{\mathbb{E}[e(\hat{\Pi}_{\text{BQ}}; \Pi, \mathcal{H}_c)^2]}_{(*)} \underbrace{\|f\|_{\mathcal{H}_c}^2}_{(**)}. \tag{4.7}$$

The term $(*)$ can be estimated with the bootstrap approximation

$$R^2 \quad = \quad \sum_{m=1}^{M} \frac{e(\hat{\Pi}_{\mathrm{BQ}}^{m}; \Pi, \mathcal{H}_c)^2}{M},$$

where $\hat{\Pi}_{\mathrm{BQ}}^{m}[f] = \sum_{j=1}^{n} w_{m,j}^{\mathrm{BQ}} f(\tilde{\mathbf{x}}_{m,j})$ is a BQ rule based on quadrature points $\tilde{\mathbf{x}}_{m,j}$ which are independent draws from $\{\mathbf{x}_j\}_{j=1}^{N}$. In SMC-BQ the term $(**)$ is an unknown constant and the statistic R, an empirical proxy for the root-mean-squared error, is monitored at each iteration. The algorithm terminates once an increase in this statistic occurs.

The problem with the naive approach of comparing R estimated at t_{i-1} directly with R estimated at t_i is that MC error can lead to an incorrect impression that R is increasing, when it is in fact decreasing, and cause the algorithm to terminate when estimation is poor (see Figure 4.10 and note the jaggedness of the estimated R curve as a function of inverse temperature t). Our solution was to apply a least-squares linear smoother to the estimates for R over 5 consecutive temperatures. This approach, illustrated in Figure 4.10, determines whether the gradient of the linear smoother is positive or negative, and in this way we are able to provide robustness to MC error in the termination criterion. In particular, the algorithm requires at least 5 temperature evaluations before termination is considered (Figure 4.10; left) and terminates when the gradient of the linear smoother becomes positive for the first time (Figure 4.10; right).

Adaptive Selection of Temperature Ladder

We conclude by noting that this whole procedure will be highly dependent on the choice of temperature ladder. The choice of temperature schedule $\{t_i\}_{i=0}^{T}$ influences several aspects of SMC-BQ: (i) The SMC approximation to Π_{t_i} is governed by the "distance" (in some appropriate metric) between $\Pi_{t_{i-1}}$ and Π_{t_i}, (ii) The speed at which the minimum t^* can be reached is linear in the number of temperatures between 0 and t^*, and (iii) The precision of BQ depends on the approximation $t^* \approx t_{i^*}$.

Factors (i,iii) motivate the use of a fine schedule with T large, while (ii) motivates a coarse schedule with T small. For this work, a temperature schedule was used that is well suited to both (i) and (ii), while a strict constraint $t_i - t_{i-1} \leq \Delta$ was imposed on the grid spacing to acknowledge (iii). The specific schedule used in this work was determined based on the conditional effective sample size (CESS) of the current particle population, as proposed in the recent work of Zhou et al.

[2016] and previously discussed in Chapter 1. The construction for the temperature schedule makes use of a sequential least squares programming algorithm and consists of two steps. Given the current temperature t_{i-1}, these steps are given by:

1. Perform a binary search in $[t_{i-1}, 1]$ by solving $\text{CESS}(\{(w_j, \mathbf{x}_j)\}_{j=1}^N, t) = N \cdot \rho$.
2. Select t_i as the solution to $\min\{t_{i-1} + \Delta, t\}$.

Sequential Monte Carlo Bayesian Quadrature

Putting all of the above together, our SMC-BQ algorithm can be summarised with the following steps:

1. Initialise the N particles using IID samples from Π_0.
2. Compute the current value of the stopping criterion.
3. Whilst the stopping criterion hasn't increased:
   (a) Select a new temperature value t_i adaptively using the conditional effective sample size criterion.
   (b) Move the particles towards Π_{t_i} using an SMC step.
   (c) Compute the value of the stopping criterion.
4. Return a BQ estimator based on n samples from the final distribution.

Note that the above algorithm assumes that the covariance function is fixed a-priori. If this is note the case, we propose to estimate kernel parameters γ via an empirical Bayes approach. This algorithm is then called sequential Monte Carlo Bayesian quadrature with kernel learning (SMC-BQ-KL). In this extended algorithm, the function evaluations are obtained at the first n (of N) states $\{\mathbf{x}_j\}_{j=1}^n$ and the parameters γ are updated in each iteration of the SMC. this can be summarised in the following step:

3. (d) Update the parameters of the covariance functions using empirical Bayes.

Note that for SMC-BQ-KL the term $(**)$ is non-constant as it depends on the kernel hyperparameters; then $(**)$ can in addition be estimated as $\|\hat{f}\|_{\mathcal{H}_c}^2 = \mathbf{w}^\top \mathrm{C}_\gamma \mathbf{w}$ and we monitor the product of R and $\|\hat{f}\|_{\mathcal{H}_c}$, with termination when an increase is observed.

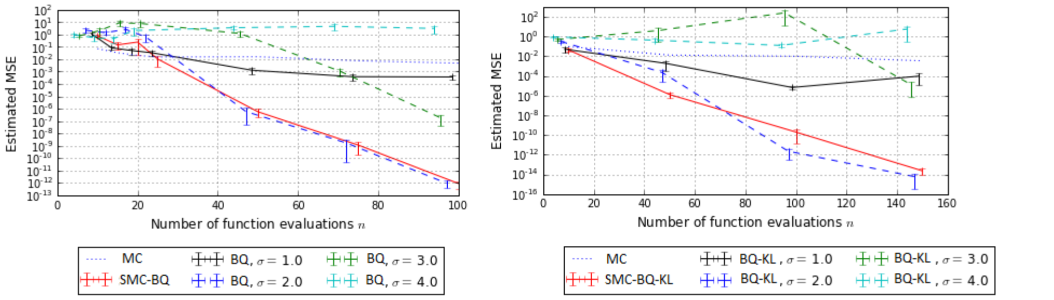

Figure 4.11: *Performance of Sequential Monte Carlo Bayesian quadrature on the running illustration.* The left plot shows SMC-BQ against BQ, whilst the right plot illustrates the versions with kernel learning.

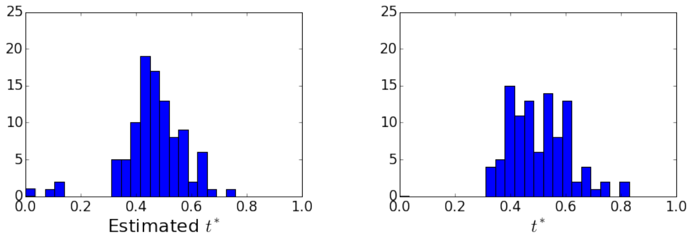

Figure 4.12: *Histograms for the optimal (inverse) temperature parameter t^*. Left:* Estimate of t^* provided under the termination criterion. *Right:* Estimate of t^* obtained by estimating R over a grid for $t \in [0, 1]$ and returning the global minimum. The similarity of these histograms is supportive of the convexity ansatz.

4.3.4 Numerical Experiments

To summarise, we have developed a novel procedure, SMC-BQ (and an extension SMC-BQ-KL), designed to approximate the optimal BQ estimator based on the unavailable optimal distribution in Equation 4.5 where the supremum over the unit ball of some RKHS $\mathcal{H}_c$. Empirical results in the previous sections suggest that SMC-BQ has the potential to provide a powerful and general algorithm for numerical integration. The additional computational cost of optimising the sampling distribution does however have to be counterbalanced with the potential reduction in numerical error, and so this method will mainly be of practical interest for problems with expensive integrands or complex target distributions. The following section reports experiments designed to test this claim.

Simulation Study

To continue our illustration from the previous section, we investigated the performance of SMC-BQ and SMC-BQ-KL for integration of $f(x) = 1 + \sin(2\pi x)$ against a $\mathcal{N}(0,1)$ measure. Here the reference measure Π_0 was taken to be $\mathcal{N}(0,8^2)$. All experiments employed SMC with $N = 300$ particles, random walk Metropolis transitions for the MCMC steps, the resample threshold was taken to be $\rho = 0.95$ and the maximum grid size $\Delta = 0.1$.

Figure 4.11 (left) reports results for SMC-BQ against BQ, for fixed length-scale $\ell = 1$. Corresponding results for SMC-BQ-KL against BQ-KL are shown in the plot on the right. It was observed that SMC-BQ (respectively SMC-BQ-KL) outperformed BQ (resp. BQ-KL) in the sense that, on a per-function-evaluation basis, the mean-squared error achieved by the proposed method was lower than for the standard method. The largest reduction in mean-squared error achieved was about 8 orders of magnitude (correspondingly 4 orders of magnitude in root-mean-squared error). A fair approximation to the $\sigma = 2$ method, which is approximately optimal for $n = 75$ (c.f. results in Figure 4.7), was observed. As an aside, we note that the mean-squared error was gated at 10^{-16} for all methods to avoid numerical ill-conditionning of the Gram matrix $\mathbf{C}$.

To understand whether the termination criterion was suitable (and, by extension, to examine the validity of the convexity ansatz, in Figure 4.12 we presented histograms for both estimated and actual optimal (inverse) temperature parameter t^*. Figure 4.13 (left) reports the dependence on the choice of initial distribution Π_0. There was relatively little influence on the root-mean-squared error obtained by the method for this wide range of initial distribution, which supports the purported robustness of the method.

We also test the method on more complex integrands in Figure 4.14: $f(x) = 1+\sin(4\pi x)$ and $f(x) = 1+\sin(8\pi x)$. These are more challenging for BQ since they are more difficult to interpolate due to their higher periodicity. However, SMC-BQ still manages to adapt to the complexity of the integrand and performs as well as the best importance sampling distribution ($\sigma = 2$).

As an extension, we also study the robustness to the dimensionality of the problem. We consider the generalisation of our main test function to $f : \mathbb{R}^d \to \mathbb{R}$ given by $f(\mathbf{x}) = 1 + \prod_{j=1}^{d} \sin(2\pi x_j)$. Notice that the integral can still be computed analytically and equals 1. We present results for $d = 2$, $d = 3$ and $d = 10$ in Figure 4.15. These two cases are more challenging for both the BQ and SMC-BQ methods, since the higher dimension implies a slower convergence rate. Once again, we notice that SMC-BQ manages to adapt to the complexity of the problem at hand, and provides improved performance on simpler sampling distributions.

Finally, we considered replacing the IID samples from Π with samples drawn from a quasi-random point sequence. Figure 4.13 (right) reports results where draws

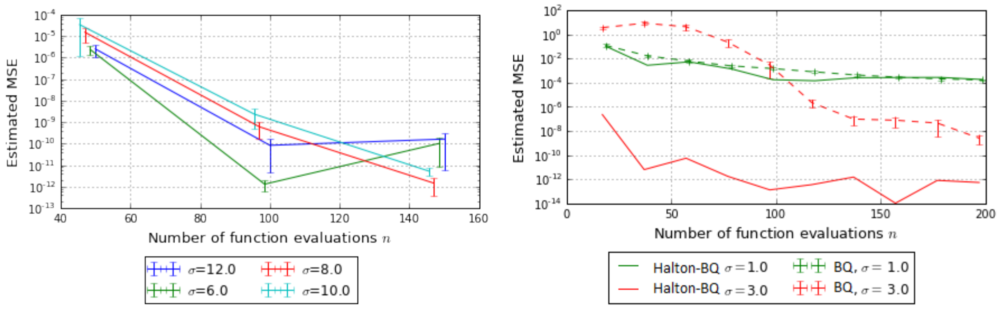

Figure 4.13: *Sensitivity of sequential Monte Carlo Bayesian quadrature to the choice of initial distribution and to the random number generator. Left:* Comparison of the performance of SMC-BQ on the running illustration of Figures 4.7 and 4.8 for varying initial distribution $\Pi_0 = \mathcal{N}(0, \sigma^2)$. *Right:* Performance of sequential quasi-Monte Carlo samplers for Bayesian quadrature. Comparison between BIS and BQ with $x_j = \Phi^{-1}(u_j)$ where the $\{u_j\}_{j=1}^n$ are the first n terms in the Halton sequence and Φ is the standard Gaussian cumulative density function.

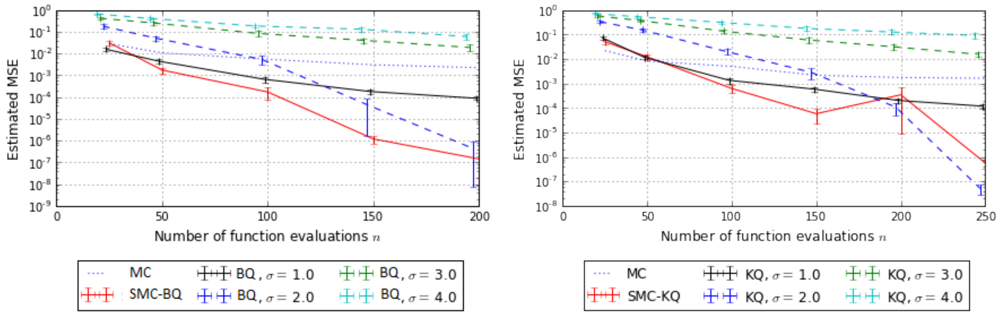

Figure 4.14: *Performance of sequential Monte Carlo Bayesian quadrature for synthetic problems of increasing complexity.* BQ and SMC-BQ are use to integrate $f(x) = 1 + \sin(4\pi x)$ (top) and $f(x) = 1 + \sin(8\pi x)$ (bottom) against $\mathcal{N}(0, 1)$. The SMC sampler was initiated with a $\mathcal{N}(0, 8^2)$ distribution. The covariance function used was a Gaussian RBF with length scales $\ell = 0.25$ (top) and $\ell = 0.15$ (bottom) each chosen to reflect the complexity of the functions.

from $\mathcal{N}(0, 1)$ were produced based on a Halton sequence. In this case, the performance is improved by up to 10 orders of magnitude in mean-squared error when the sampling is done with respect to a range of tempered sampling distribution (here $\mathcal{N}(0, 3^2)$). This suggests that a SQMC approach [Gerber and Chopin, 2015] could provide further improvement.

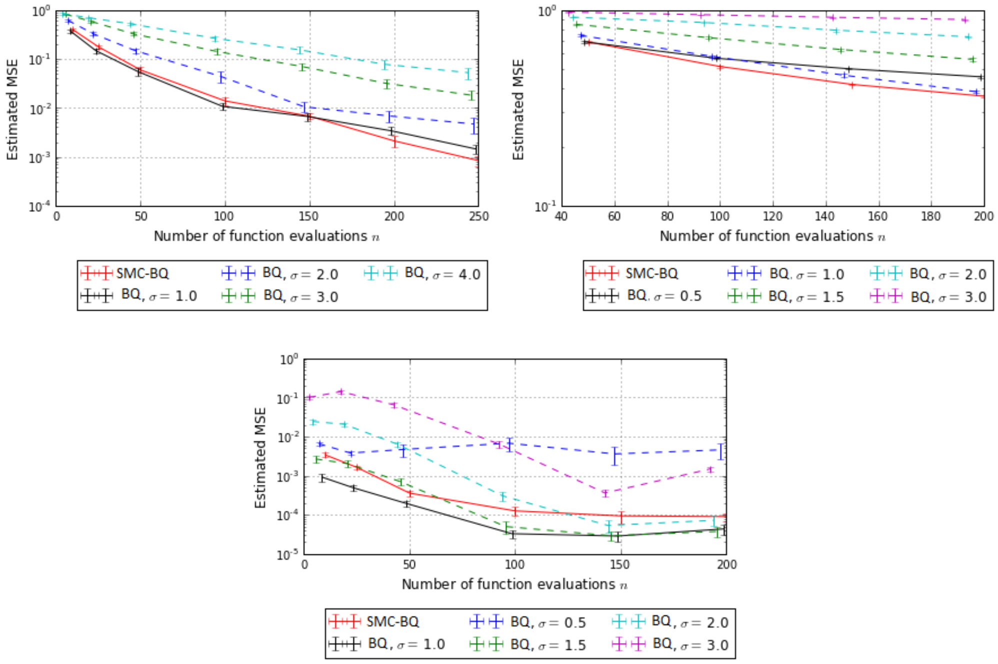

Figure 4.15: *Performance of Sequential Monte Carlo Bayesian quadrature on the running illustration in increasing dimensions.* BQ and SMC-BQ are used to integrate $f(\mathbf{x}) = 1 + \prod_{j=1}^{d} \sin(2\pi x_j)$ against a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution for $d = 2$ (top left), $d = 3$ (top right) and $d = 10$ (bottom). The SMC sampler was initiated with a $\mathcal{N}(\mathbf{0}, 8^2\mathbf{I})$ distribution. The covariance function used was a Gaussian RBF $c(\mathbf{x}, \mathbf{y}) = \exp(-\sum_{j=1}^{d}(x_j - y_j)^2/\ell_j^2)$ with the length scales $\ell_1 = \cdots = \ell_d = 0.25$ were used.

Inference for Differential Equations

Consider the model given by $\mathrm{d}u/\mathrm{d}t = f(t|\theta)$ with solution $u(t|\theta)$ depending on unknown parameters θ. Suppose we can obtain observations through the following noise model (likelihood): $y(t_i) = u(t_i|\theta) + e_i$ at times $0 = t_1 < \ldots < t_n$ where we assume $e_i \sim N(0, \sigma^2)$ for known $\sigma > 0$. Our goal is to estimate $u(T|\theta)$ for a fixed (potentially large) $T > 0$. To do so, we will use a Bayesian approach and specify a prior $p(\theta)$, then obtain samples from the posterior $\pi(\theta) := p(\theta|y)$ using MCMC. The posterior predictive mean is then defined as: $\Pi[u(T|\cdot)] = \int_\Theta u(T|\theta)\pi(\theta)\mathrm{d}\theta$, and this can be estimated using an empirical average from the posterior samples. This type of integration problem is particularly challenging as the integrand requires simulating from the differential equation at each iteration. Furthermore, with large T or the fine grid, the computational cost will be large.

For a tractable test-bed, we considered Hooke's law, given by the following

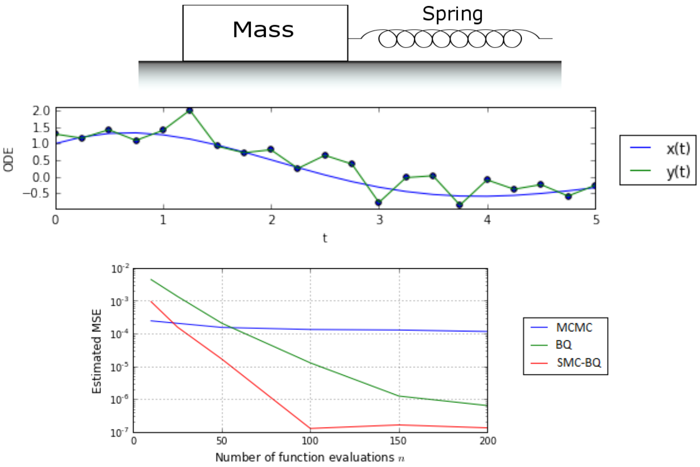

Figure 4.16: *Performance of sequential Monte Carlo Bayesian quadrature for an inverse problem based on an ordinary differential equation.* The top plot illustrates the physical system, the middle plot shows observations of the differential equation, whilst the bottom plot illustrates the superior performance of SMC-BQ against BQ.

second order homogeneous ordinary differential equation given by $\theta_5(\mathrm{d}^2u/\mathrm{d}t^2) + \theta_4(\mathrm{d}u/\mathrm{d}t) + \theta_3 u = 0$, with initial conditions $u(0) = \theta_1$ and $\frac{\mathrm{d}u}{\mathrm{d}t}(0) = \theta_2$. This equation represents the evolution of a mass on a spring with friction [Robinson, 2004, Chapter 13]. More precisely, θ_3 denotes the spring constant, θ_4 the damping coefficient representing friction and θ_5 the mass of the object. Since this differential equation is an overdetermined system we fixed $\theta_5 = 1$. In this case, if $\theta_4^2 \leq 4\theta_3$, we get a damped oscillatory behaviour as presented in Figure 4.3.4 (top). Data were generated with $\sigma = 0.4$, $(\theta_1, \theta_2, \theta_3, \theta_4) = (1, 3.75, 2.5, 0.5)$ and with log-normal priors with scale equal to 0.5 were selected for all parameters.

To implement BQ under an unknown normalisation constant for Π, we followed Oates et al. [2017c] and made use of a Gaussian RBF covariance function that was adapted with Stein's method. This will be discussed at length in Chapter 5. More precisely, we considered a kernel of the form

$$c(\theta,\theta') \;=\; 1 + \sum_{j=1}^{d} \frac{\partial^2 c_b(\theta,\theta')}{\partial\theta_j \partial\theta_j'} + s_j(\theta)\frac{\partial c_b(\theta,\theta')}{\partial\theta_j'} + s_j(\theta')\frac{\partial c_b(\theta,\theta')}{\partial\theta_j} + s_j(\theta)s_j(\theta')c_b(\theta,\theta'),$$

where c_b is a Gaussian RBF covariance function $c_b(\theta,\theta') = \exp(-\sum_{j=1}^{d}(\theta_j - \theta_j')^2/\ell_j^2)$ and $s_j(\theta) = (\nabla \log \pi(\theta))_j$ is the score function. Using integration by parts, we can easily check that $\Pi[c(\cdot,\theta)] = 1$ and $\Pi\bar{\Pi}[c] = 1$. We can also obtain the derivatives in closed form: $\partial c(\theta,\theta')/\partial\theta_j = -2\ell_j^{-2}(\theta_j - \theta_j')c(\theta,\theta')$, $\partial c(\theta,\theta')/\partial\theta_j' = 2\ell_j^{-2}(\theta_j -$

$\theta_j')c(\theta,\theta')$ and $\partial^2 c(\theta,\theta')/\partial\theta_j\partial\theta_j' = (2\ell_j^2 - 4(\theta_j - \theta_j')^2\ell_j^{-4})c(\theta,\theta')$. Furthermore, we can obtain expressions for the score function for posterior densities as follows $s_j(\theta) = \partial \log \pi(\theta)/\partial\theta_j + \partial \log \pi(\mathbf{y}|\theta)/\partial\theta_j$.

The reference distribution Π_0 was an wide uniform prior on the hypercube $[0, 10]^4$. Brute force computation was used to obtain a benchmark value for the integral. For the SMC algorithm, an independent log-normal transition kernel was used at each iteration with parameters automatically tuned to the current set of particles. Results in Figure 4.3.4 demonstrate that SMC-BQ outperforms BQ for these integration problems. These results improve upon those reported in Oates et al. [2018] for a similar integration problem based on parameter estimation for differential equations.

Chapter 5

Statistical Inference and Computation with Intractable Models

> "Despite the progress made over the last 30 years, the reasons for the effectiveness of Stein's method still remain something of a mystery."
>
> Barbour and Chen [2005]

This final chapter moves on from Bayesian probabilistic numerical methods and focuses on our second challenge. As highlighted in Chapter 1, modern statistical inference needs to cope with increasingly complex models, and in particular models with intractable densities. Two cases were highlighted: unnormalised models, where the densities can only be evaluated up to some unknown normalisation constant, and generative models, where the densities cannot be evaluated but it is nevertheless possible to obtain realisations from the model for any given parameter value. In this section, we study two notions of distance between probability measures with the useful property that they can be easily estimated for distributions for which evaluation (exact, or approximate) of densities is not possible.

The case of unnormalised models will be discussed in Section 5.1. The chapter will begin with an introduction to Stein's method, which originates in probability theory as an analytical tool to prove the asymptotic convergence of sequences of random variables, and has lately been used across computational statistics. We will then discuss how Stein's method can be combined with reproducing kernels to create a useful notion of distance between an empirical measure and a posterior measure whose density is unnormalised. This distance is called kernel Stein discrepancy (KSD) and closely relates to the WCE studied in previous chapters. We then highlight two algorithms making use of KSDs to create quadrature rules for posterior integrals as well as efficient samplers for complex posterior distributions. These will be closely related to the BQ and FW algorithms, but will allow us to by-pass the greatest drawback of these algorithms: intractable kernel means.

In the remainder of the Chapter, we will then discuss the use of kernel-based notions of discrepancy as statistical estimators. First, Section 5.2.2 will introduce a novel statistical inference algorithm for unnormalised models with KSD, then Section 5.2 will discuss a similar approach for the case of generative models using the WCE in some RKHS. In both cases, we will connect the estimators to proper scoring rules and use notions from information geometry to derive efficient numerical optimisation routines for practical implementation.

5.1 Stein's Method and Reproducing Kernels

5.1.1 Distances on Probability Measures

As discussed, we would like to have an easily computable notion of distance between two complex probability measures, such as statistical divergences. Let $\mathcal{X}$ be a metric space, and denote by $\mathcal{P}(\mathcal{X})$ be the set of Borel probability measures on this space. Statistical divergences are functions of the form $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ that satisfy $D(\mathbb{P}_1||\mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$ for all $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathcal{X})$. Divergences are usually not symmetric and do not satisfy the triangle inequality. Divergences have many uses in statistical computation including, amongst other examples, inference in statistical models [Kass and Vos, 1997] and the construction of novel variational inference schemes [Jordan et al., 1999; Blei et al., 2017], numerical optimisation algorithms [Amari, 1998; Karakida et al., 2016] or robust inference [Knoblauch et al., 2018]. As highlighted below, there exists many divergences with useful "principled" properties, but a common drawback is that they are hard or impossible to compute for most complex models.

The most commonly used divergence is the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(\mathbb{P}_1||\mathbb{P}_2) \quad := \quad \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}\right) \mathrm{d}\mathbb{P}_1, \tag{5.1}$$

where $\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_2$ is the Radon-Nikodym derivative of $\mathbb{P}_1$ with respect to $\mathbb{P}_2$. The KL divergence is closely linked to the field of information complexity (where it is often called the information gain or relative entropy), and is also popular due to its invariance to transformations of the coordinates of $\mathcal{X}$ and its convexity in the first argument. In fact, the KL divergence is a special case of two important classes of divergences: the f-divergences and the Bregman divergences [Amari, 2016]. The former is a class of divergences of the form $D_f(\mathbb{P}_1||\mathbb{P}_2) = \int_{\mathcal{X}} f(\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_2)\mathrm{d}\mathbb{P}_2$ for some convex function f satisfying $f(1) = 0$, which includes the Hellinger distance $(f(x) = (\sqrt{x}-1)^2)$ and the total-variation distance $(f(x) = 1/2(x-1))$.

Instead of using statistical divergences, it is also common to directly work with metrics or pseudo-metrics on probability measures. Pseudo-probability metrics are functions $d_{\mathcal{H}} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ which satisfy (i) $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_1) = 0$, (ii) symmetry: $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = d_{\mathcal{H}}(\mathbb{P}_2, \mathbb{P}_1)$, and (iii) the triangle inequality: $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_3) \leq d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) + d_{\mathcal{H}}(\mathbb{P}_2, \mathbb{P}_3)$ for all probability measures $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3 \in \mathcal{P}(\mathcal{X})$. Furthermore, probability metrics are pseudo-probability metrics which satisfy (iv) $d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$. Clearly, all probability metrics are divergences, but the converse does not necessarily hold. The most common pseudo-probability metrics are the integral (pseudo-)probability metrics [Müller, 1997; Sriperumbudur et al., 2010b, 2012; Sriperumbudur, 2016]:

$$d_{\mathcal{H}}(\mathbb{P}_1, \mathbb{P}_2) \quad := \quad \sup_{f \in \mathcal{H}} \left| \int_{\mathcal{X}} f(\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}) \right|. \tag{5.2}$$

Equation 5.2 should of course be familiar, since it corresponds to the definition of WCE for integration in $\mathcal{H}$. Familiar examples of integral (pseudo-)probability metrics include the following:

(i) The total variation distance, obtained using the unit ball of the set of bounded functions $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} : \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| \leq 1\}$,

(ii) The 1−Wasserstein distance (or Kantorovich metric or earth mover's distance), obtained by the unit-ball of 1-Lipschitz functions: $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} : \sup_{\mathbf{x} \neq \mathbf{y} \in \mathcal{X}} |f(\mathbf{x}) - f(\mathbf{y})|/\|\mathbf{x} - \mathbf{y}\| \leq 1\}$,

(iii) The Dudley probability metric, obtained by considering the set of bounded Lipschitz functions: $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} : \sup_{\mathbf{x} \neq \mathbf{y} \in \mathcal{X}} |f(\mathbf{x}) - f(\mathbf{y})|/\|\mathbf{x} - \mathbf{y}\| + \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| \leq 1\}$,

(iv) The maximum mean discrepancy for which $\mathcal{H}$ is taken to be the unit ball of some RKHS $\mathcal{H}_k$: $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_{\mathcal{H}_k} \leq 1\}$.

Under rather weak conditions on $\mathcal{X}$, examples (i), (ii) and (iii) are all probability metrics, but (iv) is only a probability metric under certain conditions on the kernel (and otherwise is a pseudo-probability metric). Any kernel which makes (iv) a probability metric is called a characteristic kernel [Sriperumbudur et al., 2010b]. Other examples of integral probability metrics can also be found in [Müller, 1997; Sriperumbudur et al., 2010b, 2012; Sriperumbudur, 2016].

Taking a step back to our objective of finding a statistical distance which can be computed for intractable models, it should be obvious that all of the divergences and metrics highlighted above are somewhat inadequate for our purpose. The KL divergence requires access to densities in normalised form, whilst the integral probability metrics require computation of a supremum over $\mathcal{H}$. Computing these notions of distance will hence usually be impossible whenever the model is in an unnormalised or generative form.

In the next section, we will derive a distance between probability measures called kernel Stein discrepancy (KSD) [Chwialkowski et al., 2016; Liu et al., 2016], which bypasses these issues for unnormalised models. KSDs can be recovered from maximum mean discrepancies (MMDs) by specific choice of kernels, and under several assumptions can be shown to be statistical divergences. MMDs were extensively discussed in previous chapters and correspond to the WCEs in some RKHSs. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}_k$ of functions $\mathcal{X} \to \mathbb{R}$. From Proposition 2 in Chapter **??** we have that the MMD has a straightforward expression in term of integrals of the kernel k. Furthermore, recall from Equation 3.7 in Chapter 3 that given an empirical measure $\mathbb{Q}^n = \sum_{i=1}^n w_i \delta(\mathbf{x}_i)$, where $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$, and a target measure $\mathbb{P}$, the MMD is given by[1]:

$$\begin{aligned}\mathrm{MMD}\left(\mathbb{Q}^n, \mathbb{P}\right)^2 \quad := & \quad \int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x}, \mathbf{x}')\mathbb{P}(\mathrm{d}\mathbf{x})\mathbb{P}(\mathrm{d}\mathbf{x}') - 2\sum_{i=1}^n w_i \int_{\mathcal{X}} k(\mathbf{x}_i, \mathbf{x})\mathbb{P}(\mathrm{d}\mathbf{x}) \\ & \quad + \sum_{i,j=1}^n w_i w_j k(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

As we have already clearly highlighted in Chapter 4, there are very few cases where we can actually compute this expression in closed form. Certainly, this will in general not be possible whenever the density p of $\mathbb{P}$ is unnormalised.

5.1.2 Kernel Stein Discrepancies

In this section, we introduce a divergence based on MMD where the underlying RKHS has a kernel with certain properties which allow us to avoid intractability issues in the case of unnormalised densities. Our method is based on Stein's method

[1] Note that we changed the notation from Chapter 3 to emphasise that we now see the MMD as a function of two probability measures.

[Stein, 1972], which was first used as a tool for constructing a central limit theorem for dependent variables.

Stein Discrepancies

Stein's method is based on three components: a probability measure $\mathbb{Q}$, a function space $\mathcal{G}$ (called Stein space), and an operator $\mathcal{T}_{\mathbb{Q}}$ (called Stein operator), which together satisfy the following equation called Stein's identity:

$$\int_{\mathcal{X}} \mathcal{T}_{\mathbb{Q}}[g](\mathbf{x})\mathbb{P}(\mathrm{d}\mathbf{x}) \;=\; 0 \quad \forall g \in \mathcal{G} \quad \Leftrightarrow \quad \mathbb{P} = \mathbb{Q}. \tag{5.3}$$

In this case, it is said that the Stein operator characterises the measure $\mathbb{Q}$. Stein's method has mostly been developed for analytic convergence results in probability theory; see the reviews by Barbour and Chen [2005]; Chen et al. [2011]; Barbour and Chen [2014]; Ross [2011]. More recently, it has also been used for several tasks in statistics: the analysis of maximum likelihood estimators [Anastasiou and Reinert, 2017, 2018; Anastasiou, 2017], the comparison of prior distributions in Bayesian inference [Ley et al., 2017; Ghaderinezhad and Ley, 2018] and goodness-of-fit testing [Gaunt et al., 2017]. Later in this section, we will also discuss applications to numerical integration [Oates et al., 2018, 2017c] and approximation of posterior measures [Chen et al., 2018, 2019]

Of course, finding triplets of probability measures, operators and function space which satisfy Stein's identity (Equation 5.3) can be challenging. Under regularity conditions on q (the density of $\mathbb{Q}$), a common choice of operator when $\mathcal{X} = \mathbb{R}^d$ is linked to the generator of an overdamped Langevin equation [Barbour and Chen, 2005; Gorham et al., 2016] and hence referred to as Langevin Stein operator:

$$\mathcal{L}_{\mathbb{Q}}[g](\mathbf{x}) \;=\; \frac{\langle \nabla, q(\mathbf{x})g(\mathbf{x})\rangle}{q(\mathbf{x})} \;=\; \langle g(\mathbf{x}), \nabla \log q(\mathbf{x})\rangle + \langle \nabla, g(\mathbf{x})\rangle, \tag{5.4}$$

where $\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d})^\top$ and $\langle \nabla, g(\mathbf{x})\rangle = \sum_{j=1}^d \frac{\partial g_j(\mathbf{x})}{\partial x_j}$. This operator must operate on a Stein class $\mathcal{G}$ of vector-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}^d$. We can also choose operators based on infinitesimal generators of other diffusions, see for example the following generator of an Itô diffusion process:

$$\mathcal{I}_{\mathbb{Q}}[g](\mathbf{x}) \;=\; \frac{\langle \nabla, q(\mathbf{x})\nabla g(\mathbf{x})\rangle}{q(\mathbf{x})} \;=\; \langle \nabla g(\mathbf{x}), \nabla \log q(\mathbf{x})\rangle + \Delta g(\mathbf{x}), \tag{5.5}$$

which can be used with Stein classes of scalar-valued functions on the domain $\mathcal{X}$, and where $\Delta g(\mathbf{x}) = \langle \nabla, \nabla g(\mathbf{x})\rangle = \sum_{j=1}^d \frac{\partial^2 g_j(\mathbf{x})}{\partial x_j^2}$ is called the Laplacian of g. There are many other such operators; for example a generalised version of the above two is studied by Gorham et al. [2016]:

$$\mathcal{S}_{\mathbb{Q}}[g](\mathbf{x}) \quad = \quad \frac{\langle \nabla, q(\mathbf{x})(a(\mathbf{x}) + c(\mathbf{x}))g(\mathbf{x})\rangle}{q(\mathbf{x})}. \tag{5.6}$$

where $g : \mathcal{X} \to \mathbb{R}^d$ is a vector-valued function, $a : \mathcal{X} \to \mathbb{R}^{d\times d}$ is a positive semi-definite matrix-valued function and $c : \mathcal{X} \to \mathbb{R}^{d\times d}$ is a skew symmetric matrix-valued function. Note that the three Stein operators above can be evaluated without knowledge of the normalisation constant of q. They are also all based on the generator of a diffusion process, and can be derived using the generator approach to Stein's method, which was introduced in Barbour [1988]. The importance of the particular choice of Stein operator is unclear for the applications of interest in this thesis. The main property of interest here comes from the Stein identity which allows us to construct zero-mean functions.

Kernel Stein Discrepancies

It turns out that the Stein identity (Equation 5.3) can be extremely useful to simplify the expression of integral probability metrics. In particular, it allows us to remove the problem of integration against one of the measures (which may have had an unnormalised density). Taking the function class of the IPM to be the image of functions in the Stein class through the corresponding Stein operator leads to a general class of divergences, called Stein discrepancy, and first proposed by Gorham and Mackey [2015]:

$$\begin{aligned} D_{\text{Stein}}\left(\mathbb{P}_1 || \mathbb{P}_2\right) &= \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{T}_{\mathbb{P}_2}[g](\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} \mathcal{T}_{\mathbb{P}_2}[g](\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}) \right| \\ &= \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{T}_{\mathbb{P}_2}[g](\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}) \right|. \end{aligned} \tag{5.7}$$

where $\mathcal{T}_{\mathbb{P}_2}$ is a Stein operator adapted to $\mathbb{P}_2$ and we can hence use Equation 5.3 to obtain the second identity. Note that this expression will only be a divergence under regularity conditions on the function class $\mathcal{G}$. Intuitively, we want the function class to be large enough to differentiate the two measures well. When this is the case, we clearly will have the property that whenever $\mathbb{P}_1$ is equal to $\mathbb{P}_2$, then $\int_{\mathcal{X}} \mathcal{T}_{\mathbb{P}_2}[g](\mathbf{x})\mathbb{P}_1(\mathbf{x}) = 0$ so the Stein divergence will have value zero. The general notion of Stein discrepancy with an underlying RKHS $\mathcal{H}_k$ as Stein class leads to the kernel Stein discrepancy (KSD):

$$\mathrm{KSD}\left(\mathbb{P}_1||\mathbb{P}_2\right) \quad := \quad \sup_{\|g\|_{\mathcal{H}_k}\leq 1}\left|\int_{\mathcal{X}} \mathcal{T}_{\mathbb{P}_2}[g](\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x})\right|. \tag{5.8}$$

Note that the choice of base RKHS could also be optimised, as proposed in Jitkrittum et al. [2017]. Alternative choices of Stein classes are also possible; see for example the complete graph Stein discrepancies and spanner Stein graph discrepancies of Gorham and Mackey [2015] or the random feature Stein discrepancies of Huggins and Mackey [2018]. Larger function classes could also be used, but they will tend to make the Stein discrepancy intractable.

If the Stein operator maps scalar-valued functions to other scalar-valued functions, we will take the function class $\mathcal{G}$ to be a RKHS $\mathcal{H}_k$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Alternatively, if the Stein operator maps vector-valued functions to scalar-valued functions, we will take the function class $\mathcal{G}$ to be the unit ball of some vector-valued RKHS which takes the form of the tensor product space $\mathcal{H}_k \otimes \ldots \otimes \mathcal{H}_k$ (also sometimes written as $\mathcal{H}_k^d$ where $d \in \mathbb{N}$ is the number of elements in the tensor). In either case, under regularity conditions, the image of $\mathcal{G}$ under a Stein operator $\mathcal{T}_{\mathbb{P}}$ is a scalar-valued RKHS, denoted $\mathcal{H}_{k_{\mathbb{P}}}$. When this is the case, the kernel $k_{\mathbb{P}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of $\mathcal{H}_{k_{\mathbb{P}}}$ is called a Stein reproducing kernel and takes the form $k_{\mathbb{P}}(\mathbf{x}, \mathbf{x}') = \mathcal{T}_{\mathbb{P}}\bar{\mathcal{T}}_{\mathbb{P}}k(\mathbf{x}, \mathbf{x}')$, where k is called a base kernel. Here, $\bar{\mathcal{T}}_{\mathbb{P}}$ correspond to the operator $\mathcal{T}_{\mathbb{P}}$ but acting on the second argument of the function. Note that we emphasise the distribution $\mathbb{P}$ to which the Stein kernel is adapted to in the notation $k_{\mathbb{P}}$. The KSD can alternatively be obtained from the MMD with a Stein kernel adapted to the second argument of the discrepancy, and can hence be expressed as:

$$\begin{aligned}\mathrm{KSD}\left(\mathbb{P}_1||\mathbb{P}_2\right)^2 \quad &= \quad \int_{\mathcal{X}\times\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_1(\mathrm{d}\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}') - 2\int_{\mathcal{X}\times\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_1(\mathrm{d}\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}') \\ &\quad + \int_{\mathcal{X}\times\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_2(\mathrm{d}\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}') \\ &= \quad \int_{\mathcal{X}\times\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_1(\mathrm{d}\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}'). \end{aligned} \tag{5.9}$$

The expression above was simplified using the fact that Stein reproducing kernels are elements of a Stein class corresponding to a Stein operator $\mathcal{T}_{\mathbb{P}_2}$, and hence possess the useful property that the kernel mean satisfies $\int_{\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_2(\mathrm{d}\mathbf{x}) = 0$ and hence $\int_{\mathcal{X}\times\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_2(\mathrm{d}\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}') = 0$ and $\int_{\mathcal{X}\times\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x}, \mathbf{x}')\mathbb{P}_1(\mathrm{d}\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}') = 0$. This is the main property of interest from the point of view of computational statistics.

Clearly, the expression above may not be a metric anymore since it might not be symmetric as the kernel depends on one of the arguments. However, under regularity assumptions on the base kernel, the expression above will be a statistical divergence. Recall that a kernel is called characteristic if and only if the corresponding MMD is a probability metric. To parallel this notion, we will call a Stein kernel a characteristic Stein reproducing kernel if and only if the corresponding KSD is a statistical divergence. This will be a strong assumption on the Stein kernel which will need to be checked on a case-by-case basis.

When the first argument is an empirical measure $\mathbb{Q}^n = \sum_{i=1}^n w_i \delta(\mathbf{x}_i)$ approximating some measure $\mathbb{Q}$, the expression further simplifies to:

$$\text{KSD}(\mathbb{Q}^n||\mathbb{P}) = \sqrt{\sum_{i,j=1}^n w_i w_j k_{\mathbb{P}}(\mathbf{x}_i, \mathbf{x}_j)}.$$

The equation above can be seen as an exact expression for the KSD between $\mathbb{Q}^n$ and $\mathbb{P}$, or an approximation of the KSD between $\mathbb{Q}$ and $\mathbb{P}$.

Langevin Kernel Stein Discrepancies

We will now focus on the case where $\mathcal{G}$ is a vector-valued RKHS $\mathcal{H}_k \otimes \ldots \otimes \mathcal{H}_k$ and where the operator is the Langevin Stein operator in Equation 5.4 adapted to some measure $\mathbb{P}$. In this case, we have a Stein reproducing kernel of the form [Oates and Girolami, 2016; Oates et al., 2017c, 2018]:

$$\begin{aligned} k_{\mathbb{P}}(\mathbf{x}, \mathbf{x}') &= \langle \nabla_1, \nabla_2 k(\mathbf{x}, \mathbf{x}') \rangle + \langle \nabla_1 k(\mathbf{x}, \mathbf{x}'), \nabla \log p(\mathbf{x}') \rangle \\ &\quad + \langle \nabla_2 k(\mathbf{x}, \mathbf{x}'), \nabla \log p(\mathbf{x}) \rangle + k(\mathbf{x}, \mathbf{x}') \langle \nabla \log p(\mathbf{x}), \nabla \log p(\mathbf{x}') \rangle. \end{aligned} \tag{5.10}$$

where $\nabla_1 k(\mathbf{x}, \mathbf{y}) = (\partial k(\mathbf{x}, \mathbf{y})/\partial x_1, \ldots, \partial k(\mathbf{x}, \mathbf{y})/\partial x_d)^\top$ and $\nabla_2 k(\mathbf{x}, \mathbf{y}) = (\partial k(\mathbf{x}, \mathbf{y})/\partial y_1, \ldots, \partial k(\mathbf{x}, \mathbf{y})/\partial y_d)^\top$. We now have a kernel which depends on the measure $\mathbb{P}$, but notice that it only depends on it through $\nabla \log p$, which itself can be evaluated without access to the normalisation constant of p. The KSD between two measures $\mathbb{P}_1$ and $\mathbb{P}_2$ with continuously differentiable densities is hence:

$$\begin{aligned} \text{KSD}\left(\mathbb{P}_1||\mathbb{P}_2\right) &= \left\| \int_{\mathcal{X}} \left[\langle k(\mathbf{x}, \cdot), \nabla \log p_2(\mathbf{x}) \rangle + \langle \nabla_{\mathbf{x}}, k(\mathbf{x}, \cdot) \rangle \right] \mathbb{P}_1(\mathrm{d}\mathbf{x}) \right\|_{\mathcal{H}_k} \\ &= \int_{\mathcal{X} \times \mathcal{X}} \langle \nabla \log p_2(\mathbf{x}) - \nabla \log p_1(\mathbf{x}), \nabla \log p_2(\mathbf{x}') - \nabla \log p_1(\mathbf{x}') \rangle \\ &\qquad \times k(\mathbf{x}, \mathbf{x}') \mathbb{P}_1(\mathrm{d}\mathbf{x}) \mathbb{P}_1(\mathrm{d}\mathbf{x}'), \end{aligned} \tag{5.11}$$

which can be seen either as Stein discrepancy with Stein space $\mathcal{H}_k \otimes \ldots \otimes \mathcal{H}_k$ or as the MMD with underlying Langevin Stein kernel as given in Equation 5.10, but adapted to $\mathbb{P}_2$. The KSD with Langevin Stein operator is a statistical divergence whenever it is based on a characteristic Stein kernel, which will impose certain regularity conditions on the base reproducing kernel k and the densities of the two measures. We now present several sufficient conditions for the property to hold (in all cases $\mathcal{X} \subseteq \mathbb{R}^d$):

- Theorem 2.2 in [Chwialkowski et al., 2016] shows that the Langevin KSD is a divergence if the kernel k is C_0-universal, $\mathbb{P}_1$ and $\mathbb{P}_2$ both admit continuously differentiable densities p_1 and p_2, and $\int_{\mathcal{X}} \|\nabla \log p_2(\mathbf{x}) - \nabla \log p_1(\mathbf{x})\|_2^2 \mathbb{P}_1(d\mathbf{x}) < \infty$ and $\int_{\mathcal{X}} k_{\mathbb{P}_2}(\mathbf{x},\mathbf{x})\mathbb{P}_1(\mathbf{x}) < \infty$.
- Proposition 3.3 in Liu et al. [2016] shows that the Langevin KSD is a divergence if the kernel k is integrally strictly positive definite, $\mathbb{P}_1$ and $\mathbb{P}_2$ admit continuous densities p_1 and p_2, and $\int_{\mathcal{X}} \|\nabla \log p_2(\mathbf{x}) - \nabla \log p_1(\mathbf{x})\|_2^2 \mathbb{P}_1(d\mathbf{x}) < \infty$.

This Langevin kernel Stein discrepancy was recently used for several tasks across statistics, including hypothesis testing [Chwialkowski et al., 2016; Liu et al., 2016], sampling [Liu and Wang, 2016; Liu and Lee, 2017; Liu, 2017] and convergence of sampling methods [Gorham and Mackey, 2017]. For the remainder of this section, we will highlight two more applications: the approximation of posterior measures, using a method called Stein points [Chen et al., 2018, 2019], and the construction of control variates in MC and MCMC integration [Oates et al., 2017c, 2018].

5.1.3 Stein Reproducing Kernels for Approximating Measures

We have already seen in previous chapters how quadrature estimators need efficient point selection methods for enhanced performance. KSDs can be useful for this task, especially in cases where the integrals of interest are taken against measures with densities known only up to normalisation constants (as is usually the case in Bayesian statistics).

This subsection briefly discusses one approach, called Stein points [Chen et al., 2018, 2019]. The philosophy behind Stein points is to see the problem of approximating a target measure Π (against which we would like to integrate) as an optimisation problem. More precisely, we propose to select points $\{\mathbf{x}_i\}_{i=1}^n$ to form an empirical measure $\hat{\Pi}_n = \frac{1}{n}\sum_{i=1}^n \delta(\mathbf{x}_i)$ which approximates Π well. This is done by minimising the KSD between these two measures.

$$\operatorname*{arg\,min}_{\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}} \ \mathrm{KSD}(\hat{\Pi}_n || \Pi). \tag{5.12}$$

This can equivalently be seen as selecting the optimal states with respect to the WCE in $\mathcal{H}_{k_\Pi}$ for an equally-weighted quadrature rule. For the remainder of Section 5.1, we will use the notation $D_{k_\Pi}(\{\mathbf{x}_i\}_{i=1}^n)$ to denote the kernel Stein discrepancy with Langevin Stein operator. This choice is made to make the dependence on the point set explicit.

Note that point-selection algorithms based on optimisation of statistical divergences already exist in the literature. These include the minimum energy designs of Joseph et al. [2015, 2017], which minimise the energy distance, the Stein variational gradient descent algorithm of Liu and Wang [2016]; Liu [2017], which minimises the KL divergence, and the kernel herding and FW algorithms of Chen et al. [2010]; Bach et al. [2012], which minimise MMD.

Obviously, the problem in Equation 5.12 is a highly non-convex optimisation problem, which will be high-dimensional in the case where we want a high number of points n. To reduce the complexity of this problem, we propose two different point-sequences.

The first and simplest algorithm that we consider follows a greedy strategy and is hence called Stein greedy points. The initial point $\mathbf{x}_1$ is taken to be a global maxima of the density π of Π, then each subsequent point $\mathbf{x}_n$ is taken to be a global minima of $D_{k,\pi}(\{\mathbf{x}_i\}_{i=1}^n)$, with the objective function being viewed as a function of $\mathbf{x}_n$ with $\{\mathbf{x}_i\}_{i=1}^{n-1}$ being fixed. This is equivalent to selecting:

$$\mathbf{x}_n \quad \in \quad \arg\min_{\mathbf{x}\in\mathcal{X}} \quad \sum_{i=1}^{n-1} k_\Pi(\mathbf{x}_i,\mathbf{x}) + \frac{k_\Pi(\mathbf{x},\mathbf{x})}{2}. \tag{5.13}$$

As seen in Chapter 4, another approach is to use a FW algorithm, which boils down to solving the problem: $\arg\min_{g\in\mathcal{M}} \frac{1}{2}\|g-\Pi[k_\Pi(\cdot,\mathbf{x})]\|^2_{\mathcal{H}_{k_\Pi}}$, where $\mathcal{M}$ is the marginal polytope of the RKHS $\mathcal{H}_{k_\Pi}$ (see Equation 4.3 in Chapter 4). As might be expected, the objective function is closely related to KSD; for $g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n k_\Pi(\mathbf{x}_i,\mathbf{x})$:

$$D_{k_\Pi}(\{\mathbf{x}_i\}_{i=1}^n) \quad = \quad \|g-\Pi[k_\Pi(\cdot,\mathbf{x})]\|_{\mathcal{H}_{k_\Pi}}.$$

This leads us to our second algorithm, where the initial point $\mathbf{x}_1$ is once again taken to be a global maximum of the density π; which in the context of this algorithm corresponds to an element $g_1(\mathbf{x}) = k_\Pi(\mathbf{x}_1,\mathbf{x})$. Then, at iteration $n > 1$, the convex combination $g_n = \frac{n-1}{n}g_{n-1} + \frac{1}{n}\bar{g}_n$ is constructed where the element $\bar{g}_n$ encodes a direction of steepest descent. Given that minimisation of a linear objective over a convex set can be restricted to the boundary of that set, it follows that $\bar{g}_n(\mathbf{x}) = k(\mathbf{x}_n,\mathbf{x})$ for some $\mathbf{x}_n \in \mathcal{X}$ (see step 1 of the algorithm in 4.2.1).

The second algorithm, called Stein herding, can hence be concisely summarised as follows. First select $\mathbf{x}_1 \in \arg\max_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})$, then at iteration $n > 1$:

$$\mathbf{x}_n \quad \in \quad \arg\min_{\mathbf{x}\in\mathcal{X}} \quad \sum_{i=1}^{n-1} k_\Pi(\mathbf{x}_i, \mathbf{x}). \tag{5.14}$$

The Stein greedy and Stein herding updates (Equations 5.13 and 5.14 respectively) are very similar to one another. First, the Stein greedy update can be seen as a regularised version of the Stein herding update, with regulariser $\frac{1}{2}k_\Pi(\mathbf{x}, \mathbf{x})$. The two updates coincide if $k_\Pi(\mathbf{x}, \mathbf{x})$ is a constant. This is true for most reproducing kernels used in practice as these tend to be isotropic, however, this is typically not true for a Stein reproducing kernel such as the Langevin Stein kernel in Equation 5.10.

The Stein greedy and Stein herding algorithms both require solving a global (non-convex) optimisation problem over $\mathcal{X}$ at each iteration. In practice, this will be infeasible, and the use of numerical methods such as a grid search, MC search or Nelder-Mead search will be required. Both algorithm will also have roughly the same computational cost, which will be $O(n^2)$ in addition to any computational cost of the global optimisation routine. We thus anticipate applications in which the evaluation of π (or its gradient) constitutes the principal computational bottleneck.

We now highlight the performance of Stein points on a synthetic example popular in the sampling literature: the Rosenbrock density. The Rosenbrock target has density of the form: $\log \pi(\mathbf{x}) \propto -100(x_2 - x_1^2)^2 - (1 - x_1)^2$, which tends to be challenging since the region of high density is narrow and has high curvature (see Figure 5.1). We demonstrate the performance of the Stein greedy algorithm on this target, where a Monte Carlo search is performed at iteration, using a high number of IID uniform points on $[-4, 4] \times [-1, 10]$. The KSD used in this example used a base kernel which was an inverse-multiquadric kernel $k(\mathbf{x}, \mathbf{x}') = (\|\mathbf{x} - \mathbf{x}'\|_2^2 + 1)^{-l}$ with parameter $l = 0.7$. As seen in Figure 5.1, the Stein greedy algorithm is able to select representative points from this target. This required a large number of Monte Carlo points due to the fact that the region of high density is very narrow.

Further applications to problems in Bayesian computation were also presented in [Chen et al., 2018, 2019], including approximating the posterior distribution over parameters of a GP model, and the posterior distribution over parameters of an integrated generalised autoregressive conditional heteroskedasticity model.

On the theoretical side, under regularity conditions, it is in fact possible to show that both the Stein greedy and Stein herding algorithms will minimise KSD asymptotically. One such condition is that k_Π is Π-sub-exponential, which means that $\mathbb{P}_{Z\sim\Pi}[k_\Pi(Z, Z) \geq t] \leq c_1 e^{-c_2 t}$ for some constants $c_1, c_2 > 0$ and all $t \geq 0$.

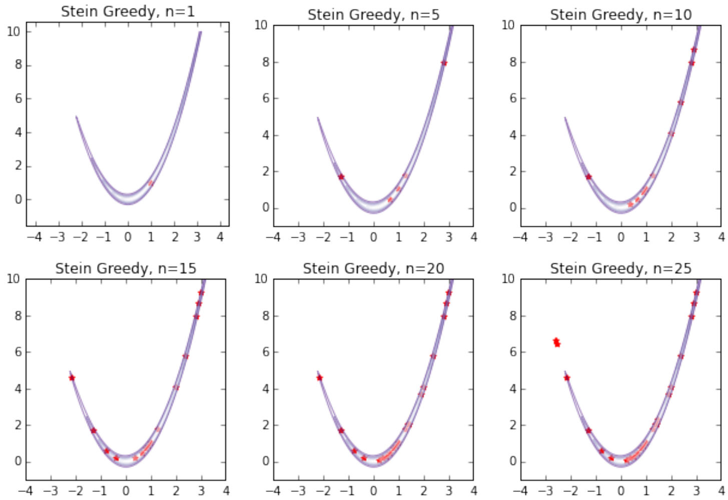

Figure 5.1: *Stein greedy points for the Rosenbrock density.* The algorithm starts at the global maximum $\mathbf{x}_0 = (1, 1)$ of the density, then greedily add points to minimise the Langevin KSD. In this case, the inner-optimisation loops where performed using a Monte Carlo search with IID uniform random variables on $[-4, 4] \times [-1, 10]$.

Theorem 16 (**Consistency of Stein greedy points**). *Suppose that the Stein reproducing kernel k_Π is a Π-sub-exponential reproducing kernel. Then $\exists c_1, c_2 > 0$ such that for all $\{\mathbf{x}_i\}_{i=1}^n$ satisfying*

$$\frac{k_\Pi(\mathbf{x}_j, \mathbf{x}_j)}{2} + \sum_{i=1}^{j-1} k_\Pi(\mathbf{x}_i, \mathbf{x}_j) \;\leq\; \frac{\delta}{2} + \min_{\mathbf{x} \in \mathcal{X} : k_\Pi(\mathbf{x},\mathbf{x}) \leq R_j^2} \frac{k_\Pi(\mathbf{x}, \mathbf{x})}{2} + \sum_{i=1}^{j-1} k_\Pi(\mathbf{x}_i, \mathbf{x})$$

with $\sqrt{2\log(j)/c_2} \leq R_j \leq \infty$ for each $j = 1, \ldots, n$, we have

$$e(\hat{\Pi}_n; \Pi, \mathcal{H}_{k_\Pi}) \;=\; D_{k_\Pi}(\{\mathbf{x}_i\}_{i=1}^n) \;\leq\; e^{\pi/2}\sqrt{\frac{2\log(n)}{c_2 n} + \frac{c_1}{n} + \frac{\delta}{n}}.$$

The proof of this result can be found in the supplementary material of Chen et al. [2018], and a similar theorem for the herding case can also be found in this paper. We note a particular strength of this theorem: the rate holds even when the global optimisation routine at each iteration has not converged. Indeed, the $\delta/2$ term allows for error at each iteration. Another advantage is that we do not require the kernel to be bounded, but weaken this condition to $\Pi-$sub-exponential. We note that the theorem gives a convergence rate of $O_P(n^{-\frac{1}{2}+\epsilon})$ for functions in $\mathcal{H}_{k_\Pi}$. This is not particularly fast when compared with rates for optimally-weighted quadrature rules in Chapter 3. However, Stein points have the significant advantage that they can be used without access to a kernel mean. The result in this theorem also does not seem to match the impressive approximation properties highlighted in Figure 5.1 or Chen et al. [2018], indicating that there is most likely a gap between empirical results and the theory available for these algorithms.

To summarise, we have now proposed two algorithms, called Stein greedy and Stein herding, for the approximation of measures whose densities are only known up to normalisation constants. This is particularly useful in the case of Bayesian statistics, where the posterior often includes an intractable integral which is hard to approximate. In this section, we illustrated how these algorithms can be particularly efficient at this task, and given theoretical backing for this performance.

In terms of theory, one question remains: is minimising a KSD a sensible objective for obtaining a point set? Or in other words, is the RKHS $\mathcal{H}_{k_\Pi}$ large enough to differentiate two measures? The answer to this question can be shown to be affirmative under several conditions on the base kernel and target measure. Gorham and Mackey [2017] (Section 3.2 and 3.3) and Chen et al. [2018] (Section 5.2) provided sufficient conditions to guarantee convergence in distribution of $\hat{\Pi}_n$ to the target measure Π for several heavy-tail kernels (such as the inverse-multiquadric). This was later extended to pre-conditioned kernels in Chen et al. [2019].

5.1.4 Stein Reproducing Kernels for Numerical Integration

Recall our main challenge of numerically approximating integrals $\Pi[f] = \int_{\mathcal{X}} f(\mathbf{x})\Pi(\mathrm{d}\mathbf{x})$, and assume the measure Π admits a continuously differentiable density π with respect to the Lebesgue measure. We will show in this section that Stein's method can be extremely useful in creating efficient quadrature rules which can be used as control variates for MC and MCMC integration.

Assume that we have access to a set of points $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ such that the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta(\mathbf{x}_i)$ is a good approximation of the target Π. These points might be Π-distributed, realisations of a Markov chain with invariant distribution Π, or even obtained with deterministic methods, such as the Stein points algorithms from the previous subsection. For the sake of simplicity, we will limit ourselves to MC and MCMC methods.

We have already seen how these point sets lead to quadrature rules, and have discussed/studied their performance through several error criterion. For example, in the MC or MCMC case, recall that the central limit theorem states that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \Pi[f]\right) \ \xrightarrow{D} \ \mathcal{N}(0,\sigma^2),$$

In the MC case, the variance of the central limit theorem is $\sigma^2_{\text{MC}} = \text{Var}_\pi[f]$, which corresponds to the variance of f under Π. On the other hand, in the MCMC case, the central limit theorem has variance: $\sigma^2_{\text{MCMC}} = \text{Var}_\pi[f] + 2\sum_{k=1}^{\infty} \text{Cov}_\pi[f(X_0), f(X_k)]$ [Jones, 2004]. Direct MC or MCMC estimation of $\Pi[f]$ would hence be prohibitive whenever f had high variance with respect to the target Π. To reduce the error of these schemes, it is common to use control variates, which are functions $\tilde{f}_{\text{CV}} : \mathcal{X} \to \mathbb{R}$ such that the integral $\Pi[\tilde{f}_{\text{CV}}]$ is known analytically. In this case, we can rewrite the integral of interest as

$$\Pi[f] \ = \ \Pi[f] - \Pi[\tilde{f}_{\text{CV}}] + \Pi[\tilde{f}_{\text{CV}}] \ = \ \Pi[f - \tilde{f}_{\text{CV}}] + \Pi[\tilde{f}_{\text{CV}}],$$

where now the second term is known in closed form and the first term needs to be estimated using some quadrature rule:

$$\Pi[f] \ \approx \ \hat{\Pi}[f - \tilde{f}_{\text{CV}}] + \Pi[\tilde{f}_{\text{CV}}]. \tag{5.15}$$

If $\tilde{f}_{\text{CV}}$ is chosen such that $\text{Var}_\pi[f - \tilde{f}_{\text{CV}}]$ is much smaller than $\text{Var}_\pi[f]$, the error in approximating $\Pi[f]$ via Equation 5.15 will be lower than when using direct MC or MCMC integration.

In general, such a function $\tilde{f}_{\text{CV}}$ may be directly available through domain-specific knowledge [Newton, 1994; Henderson and Glynn, 2002], but this is rarely the case in general. Alternatively, control variate can sometimes be built using known properties of the method used for obtaining samples. See Andradóttir et al. [1993]; Hammer and Tjelmeland [2008]; Dellaportas and Kontoyiannis [2012] for control variates based on the proposal densities of MCMC samplers, and Hickernell et al. [2005] for control variates specialised to QMC. An obvious drawback is that these approaches cannot be used in general settings where properties of $\{\mathbf{x}_i\}_{i=1}^n$ are unknown. A more general and applicable approach is the following. First, separate $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ into two sets $\mathbf{X}_1 = \{\mathbf{x}_i\}_{i=1}^m$ and $\mathbf{X}_2 = \{\mathbf{x}_i\}_{i=m+1}^n$. Then:

1. Use $\mathbf{X}_1$ to build an approximation $\tilde{f}_{\text{CV}}$ of f in some space $\mathcal{H}$ such that $\forall h \in \mathcal{H}, \exists c \in \mathbb{R}$ such that $\Pi[h] = c$ is known in closed form.

2. Approximate $\Pi[f - \tilde{f}_{\text{CV}}]$ using a quadrature rule based on $\mathbf{X}_2$.

In this case, if the integrand can be approximated at a fast rate in m, then $\text{Var}_\pi[f - \tilde{f}_{\text{CV}}]$ will decrease at a fast rate which may reduce the integration error at a faster rate than the Monte Carlo rate.

Clearly the first step will be the most challenging as finding a function space $\mathcal{H}$ with the property that the integral of all functions is known in closed-form will be non-trivial. An example is given in Paisley et al. [2012]; Wang et al. [2013], who use a Taylor expansion of the integrand. Unfortunately, this will only be a feasible approach when integrating against simple probability measures, like a Gaussian or uniform, but we would like a general methodology which can be applied to any measure with density known up to normalisation constant.

However, the first step is clearly amenable to the use of Stein's method. Any function of the form $\tilde{f}_{\text{CV}} = \mathcal{T}_\Pi[g] + c$ for $g \in \mathcal{G}$ and $c \in \mathbb{R}$, where $\mathcal{T}_\Pi$ and $\mathcal{G}$ are a pair of Stein operator and Stein class, is a possible choice of control variate. In this case, step 1 reduces to finding a function of this form leading to the greatest reduction in numerical integration error. It is common to select g from a parametric family of functions $\{g_\theta\}_{\theta \in \Theta}$, in which case the search in $\mathcal{G}$ is replaced by a search over the parameter space Θ. This problem can be solved by considering a general discrepancy loss function, which given a value in Θ, returns a value describing the suitability of g_θ. We now highlight two examples.

The first example is to choose θ by interpolation, which can be done by solving numerically $\tilde{f}_{\text{CV}}(\mathbf{x}) = \mathcal{T}_\Pi[g_\theta](\mathbf{x}) + c = f(\mathbf{x})$ in terms of (c, θ). Of course, selecting $\tilde{f}_{\text{CV}}$ by interpolation will indirectly minimise the variance $\text{Var}_\pi[f - \tilde{f}_{\text{CV}}]$, and the variance will take value zero if we interpolate the function exactly. A second option would be to select g_θ to minimise the asymptotic variance $\text{Var}_\pi[f - \tilde{f}_{\text{CV}}] = \text{Var}_\pi[f - \mathcal{T}_\Pi[g_\theta] - c]$ directly. In this case, the term c is not needed and can be set to zero by default. This is because the variance is not affected by constants: $\text{Var}_\pi[f - \mathcal{T}_\Pi[g_\theta]] = \text{Var}_\pi[f - \mathcal{T}_\Pi[g_\theta] - c]$.

Our proposed strategy for building control variates is therefore the following. First, separate $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ into two sets $\mathbf{X}_1 = \{\mathbf{x}_i\}_{i=1}^m$ and $\mathbf{X}_2 = \{\mathbf{x}_i\}_{i=m+1}^n$ and fix a Stein operator $\mathcal{T}_\Pi$ and parametric Stein class $\mathcal{G}$. Then:

1. Use $\mathbf{X}_1$ to select a control variate of the form $\tilde{f}_{\text{CV}} = \mathcal{T}_\Pi[g_\theta] + c$ by minimising a loss function in θ.

2. Compute a quadrature approximation of $\Pi[f - \tilde{f}_{\text{CV}}]$ using $\mathbf{X}_2$.

It turns out that many existing control variates methodologies available in the literature can be recovered as special cases of this approach. We now highlight a few examples.

1. Motivated by a specific Hamiltonian differential operator from statistical physics, Assaraf and Caffarel [1999]; Mira et al. [2013] proposed to perform step 1 using functions:

$$\mathcal{T}_\Pi[g_\theta](\mathbf{x}) = -\frac{\Delta[P(\mathbf{x}|\theta)\sqrt{\pi(\mathbf{x})}]}{2\sqrt{\pi(\mathbf{x})}} + \frac{P(\mathbf{x}|\theta)\Delta[\sqrt{\pi(\mathbf{x})}]}{2\sqrt{\pi(\mathbf{x})}}$$

   where $P(\mathbf{x}|\theta)$ is a class of polynomials of order $p \in \mathbb{N}$ with coefficients summarised in the vector θ. They estimate the coefficients θ by minimising $\text{Var}_\pi[f - \tilde{f}_{\text{CV}}]$. Full implementation details can be found in Papamarkou et al. [2014]. This can be shown to be equivalent to using the Itô Stein operator in Equation 5.5 together with a Stein class consisting of polynomials of order p.

2. Another example, called control functionals, was recently proposed in Oates et al. [2017a,c, 2018]; Oates and Girolami [2016]. These control variates are based on interpolants in a RKHS of the form $\mathcal{L}_\Pi[g_\theta](\mathbf{x}) = \sum_{i=1}^m \theta_i k_\Pi(\mathbf{x}, \mathbf{x}_i)$ with $\mathbf{x}_i \in \mathcal{X}$ and $\theta_i \in \mathbb{R}\ \forall i = 1, \ldots, m$, and where the kernel k_Π is the Langevin Stein reproducing kernel previously defined in Equation 5.10. Finding the optimal θ for interpolation can be solved in closed form as a least-squares problem. Control functionals were shown to be effective for variance reduction and can lead to faster convergences rates than direct MCMC integration [Oates et al., 2018].

3. Zhu et al. [2018] also approached this problem using neural networks and used functions of the form: $\mathcal{L}_\Pi[g_\theta](\mathbf{x}) = \langle \nabla_\mathbf{x}, g(\mathbf{x}|\theta)\rangle + \langle g(\mathbf{x}|\theta), \nabla_\mathbf{x} \log \pi(\mathbf{x})\rangle$, where $g(\mathbf{x}|\theta)$ are vector-valued neural networks with weights θ. Zhu et al. [2018] then propose to use an estimate of the mean-squared error to optimise the parameters. Neural networks have been shown to be particularly effective at approximating high dimensional functions which can be written as a composition of low-dimensional functions [Poggio et al., 2017].

Before concluding this section, we note that the integral of a control variate can itself be used as a stand-alone quadrature estimator. Indeed, we can simply disregard the estimator of $\Pi[f - \tilde{f}_{\text{CV}}]$ and use $\Pi[\tilde{f}_{\text{CV}}]$ as an estimate of $\Pi[f]$ in Equation 5.15. For example, when considering the control functionals approach of Oates et al. [2018, 2017c], we notice that the integral $\Pi[\tilde{f}_{\text{CV}}]$ can be obtained in closed form, and actually corresponds to the BQ estimator of $\Pi[f]$ obtained when using the kernel k_+. This is in fact what was done in Chapter 4 for the differential equation example. Stein's method therefore provides us with an alternative to the methodologies developed in Chapter 3 for BQ with intractable kernel means.

Unfortunately, one drawback of this approach is that the estimators will be biased.

5.2 Kernel-based Estimators for Intractable Models

We have now completed our discussion of the use of Stein's method for statistical computation. Clearly Stein's identity is a very useful tool to construct novel methodology for numerical integration, and there is scope for much further work in this area. In the second part of this chapter, we will highlight another use of Stein's method and kernel methods: statistical inference for models with intractable likelihoods. We will focus on both unnormalised models, which will be tackled using KSDs, and generative models, for which we will use MMDs.

5.2.1 Minimum Distance Estimators

Our kernel-based estimators for intractable models fall within the class of minimum distance estimators, which are introduced below together with the related field of information geometry. Information geometry [Amari, 1987, 2016; Barndorff-Nielsen, 1978] is concerned with the geometry of statistical manifolds. These are manifolds for which each point corresponds to a Borel probability measure $\mathbb{P} \in \mathcal{P}(\mathcal{X})$. Commonly, these manifolds correspond to parametric families $\mathcal{P}_\Theta(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ which are classes of probability measures $\mathbb{P}_\theta$ indexed by a parameter $\theta = (\theta_1, \ldots, \theta_p) \in \Theta$. An obvious choice of coordinates on a statistical manifold is given by the parameter θ. The parameter space Θ will be assumed to be a subset of $\mathbb{R}^p$ for some $p \in \mathbb{N}$ for the remainder of this chapter, but could itself be a space of functions.

A common example of statistical manifold is the exponential family, which is a class of probability measures with probability density function of the form:

$$p(\mathbf{x}|\theta) \quad = \quad h(\mathbf{x}) \exp\left(\langle \theta, T(\mathbf{x}) \rangle - c(\theta)\right), \tag{5.16}$$

for some function $h : \mathcal{X} \to \mathbb{R}$ of the form $h(\mathbf{x}) \propto \exp(b(\mathbf{x}))$, which is the density of some base measure, some summary statistic $T : \mathcal{X} \to \mathbb{R}^p$ and some normalisation constant $c : \Theta \to \mathbb{R}$ which guarantees that $p(\mathbf{x}|\theta)$ is a probability density function (i.e. is normalised). In this case, the parameter space is given by $\Theta = \{\theta \in \mathbb{R}^p : \log c(\theta) = \int_{\mathcal{X}} h(\mathbf{x}) \exp(\langle \theta, T(\mathbf{x}) \rangle) \mathrm{d}\mathbf{x} < \infty\}$. The formulation above is in terms of a parameterisation called the natural parameterisation. The exponential family is a large family which includes some classical distributions such as the Gaussian, Poisson, Dirichlet and Gamma distributions. It also includes many more complex models such as graphical models, including pairwise interaction models [Lin et al., 2016], or certain neural networks [Gutmann and Hyvärinen, 2012].

Going back to the concept of statistical manifold, we need to construct a notion of distance on a parametric class of probability models. This will usually be derived from a statistical divergence. Although a divergence does not define a metric on $\mathcal{P}(\mathcal{X})$, it induces a symmetric tensor g whose matrix (g_{ij}) is positive semi-definite: $g_{ij}(\theta) := -(\partial^2/\partial\alpha^i\partial\beta^j)D\big(\mathbb{P}_\alpha||\mathbb{P}_\beta\big)|_{\theta=\alpha=\beta}$. When $g_{ij}(\theta)$ is positive definite for all $\theta \in \Theta$, it defines a function g which maps θ to the matrix $g_{ij}(\theta)$. This is called the metric tensor or information metric, and can be used to define a Riemannian geodesic distance Amari [2016].

Minimum Distance Estimators and Scoring Rules

Consider now the problem of statistical inference for a given statistical model. A common approach is to consider some loss function $L : \Theta \to \mathbb{R}$ based on a divergence between an element of the parametric family $\mathcal{P}_\Theta(\mathcal{X}))$ and an empirical probability measure $\mathbb{Q}^m = \frac{1}{m}\sum_{j=1}^m \delta(\mathbf{y}_j)$ obtained from the IID realisations $\{\mathbf{y}_j\}_{j=1}^m$ available to us from the correct model $\mathbb{Q}$. These estimators are called minimum distance estimators and are given by the solution of the following (usually non-convex) optimisation problem:

$$\hat{\theta}_m \;=\; \underset{\theta\in\Theta}{\arg\min}\, L(\theta) \;=\; \underset{\theta\in\Theta}{\arg\min}\, D(\mathbb{Q}^m||\mathbb{P}_\theta). \tag{5.17}$$

See the books of Pardo [2005] and Basu et al. [2011] for more details, or the recent paper by Jewson et al. [2018] for a Bayesian alternative. In special cases, this optimisation problem can be solved in closed form, but it will generally be necessary to employ numerical optimisation routines. Clearly, this pair of parametric family and statistical divergence directly leads to the notion of statistical manifold, and we will be able to use information geometry to study this problem.

Minimum distance estimators are closely connected to the concept of scoring rules [Gneiting and Raftery, 2007; Dawid, 2007; Parry et al., 2012], although not all scoring rules lead minimum distance estimators[2]. A scoring rule is a function $S : \mathcal{X}\times\mathcal{P}(\mathcal{X}) \to \mathbb{R}$ such that $S(\mathbf{x},\mathbb{P})$ quantifies the accuracy of a model $\mathbb{P}$ upon observing the realisation $\mathbf{x}$. A scoring rule is said to be strictly proper if $\int_\mathcal{X} S(\mathbf{x},\mathbb{P}_2)\mathbb{P}_1(\mathrm{d}\mathbf{x})$ is uniquely minimised when $\mathbb{P}_1 = \mathbb{P}_2$. Any strictly proper scoring rule induces a divergence of the form $D_S(\mathbb{P}_1||\mathbb{P}_2) = \int_\mathcal{X} S(\mathbf{x},\mathbb{P}_2)\mathbb{P}_1(\mathrm{d}\mathbf{x}) - \int_\mathcal{X} S(\mathbf{x},\mathbb{P}_2)\mathbb{P}_2(\mathrm{d}\mathbf{x})$, which by construction will be minimised when $\mathbb{P}_1 = \mathbb{P}_2$. These divergences can then be

[2] We note that the name "scoring rule" is in no way related to the score function $\nabla \log p$, although some scoring rules might depend on $\nabla \log p$.

used as loss functions to get minimum distance estimators of the form:

$$\begin{aligned}\hat{\theta}_m^S &= \underset{\theta \in \Theta}{\arg\min}\, D_S(\mathbb{Q}^m || \mathbb{P}_\theta) &= \underset{\theta \in \Theta}{\arg\min} \int_{\mathcal{X}} S(\mathbf{y}, \mathbb{P}_\theta) \mathbb{Q}^m(\mathrm{d}\mathbf{y}) \qquad (5.18)\\ & &= \underset{\theta \in \Theta}{\arg\min} \frac{1}{m} \sum_{j=1}^{m} S(y_j, \mathbb{P}_\theta).\end{aligned}$$

See the work by Mameli and Ventura [2015] and Dawid et al. [2016] for asymptotic properties of such estimators, and Merkle and Steyvers [2013] for advice on choosing a scoring rule. A popular choice of strictly proper scoring rules are the local strictly proper scoring rules, which only depend on the log-likelihood and its derivatives [Parry et al., 2012; Ehm and Gneiting, 2012; Parry, 2016]. Note that scoring rules can also be defined for discrete domains [Dawid et al., 2012]. Estimators based on scoring rules require finding the solution to the following equations in θ: $\sum_{j=1}^{m} \nabla_\theta S(\mathbf{y}_j, \mathbb{P}_\theta) = \mathbf{0}$, which are called estimating equations and where $\mathbf{0} = (0, \ldots, 0)^\top \in \mathbb{R}^p$. For strictly proper scoring rules, one can easily show that these estimating equations are unbiased (i.e. $\int_{\mathcal{X}} \nabla_\theta S(\mathbf{x}, \mathbb{P}_\theta) \mathbb{P}_\theta(\mathrm{d}\mathbf{x}) = 0$), and as a consequence the associated estimators are consistent (see for example Theorem 1 and Corollary 2 of Dawid [2007]).

There are two scenarios of interest in the context of minimum distance estimator: The M-closed and M-open cases. First, in the M-closed case, we assume that $\mathbb{Q}$ is an instance of the parametric family $\mathcal{P}_\Theta(\mathcal{X})$. The statistical inference problem therefore boils down to finding the value $\theta^* \in \Theta$ such that $\mathbb{P}_{\theta^*}$ corresponds to $\mathbb{Q}$. Alternatively, in the M-open case, $\mathbb{Q}$ can be any probability measure in $\mathcal{P}(\mathcal{X})$, and is not necessarily in the parametric family $\mathcal{P}_\Theta(\mathcal{X})$. In this case, we look for the value θ^* such that $\mathbb{P}_{\theta^*}$ is the closest possible to $\mathbb{Q}$ in terms of some statistical divergence. Obviously, the M-closed case is much more restrictive, but can be more easily understood from a theoretical viewpoint. The M-open case, on the other hand, reflects the practical realities illustrated by George E. P. Box's now famous phrase: "all models are wrong, but some are useful". The M-open case is, however, much harder to analyse from a theoretical viewpoint.

The M-open setting requires us to study the robustness of an estimator, which is concerned with corruptions in the data generating process. For example, in applied statistics, data might be assumed to correspond to IID realisations of some model but might in fact consist of correlated observation. Alternatively, we might be in an M-open setting where our data consists of realisations from a mixture distribution consisting of a model from the parametric family, and of some distribution of outliers. The reader is referred to Huber and Ronchetti [2009] or Chapter 10 in Steinwart and

Christmann [2008] for extensive introductions. Here, the choice of divergence will significantly influence the robustness of the associated estimator. There is usually a trade-off between robustness and efficiency of estimators, and the choice of scoring rule should hence be made with this in mind.

An important concept in robust statistics is that of the influence function $\text{IF}_S : \mathcal{X} \times \mathcal{P}_\Theta(\mathcal{X}) \to \mathbb{R}$ where $\text{IF}_S(\mathbf{z}, \mathbb{P}_\theta)$ measures the impact of an infinitesimal contamination of the data generating model $\mathbb{P}_\theta$ in the direction of a Dirac measure located at some point $\mathbf{z}$. The influence function of a minimum distance estimator based on a scoring rule S is given by [Dawid and Musio, 2014]:

$$\text{IF}_S(\mathbf{z}, \mathbb{P}_\theta) = \left(\int_{\mathcal{X}} \nabla_\theta \nabla_\theta S(\mathbf{x}, \mathbb{P}_\theta) \mathbb{P}_\theta(\mathrm{d}\mathbf{x}) \right)^{-1} \nabla_\theta S(\mathbf{z}, \mathbb{P}_\theta). \tag{5.19}$$

where $(\nabla_\theta \nabla_\theta S(\mathbf{x}, \mathbb{P}_\theta))_{jk} = \partial^2 S(\mathbf{x}, \mathbb{P}_\theta)/\partial\theta_j \partial\theta_k$. The supremum of the influence function over $\mathbf{z} \in \mathcal{X}$ is called the gross-error sensitivity, and if it is finite, we say that an estimator is bias-robust (also called B-robust, or robust in the sense of Hampel) [Hampel, 1971].

Maximum Likelihood Estimation

To illustrate the definitions above, we now consider the most widely studied example of minimum distance estimator. When using the KL divergence, the minimum distance estimator in Equation 5.17 becomes equivalent to maximum likelihood estimators [Fisher, 1922]:

$$\arg\min_{\theta \in \Theta} L(\theta) = \arg\min_{\theta \in \Theta} D_{\text{KL}}(\mathbb{Q}^m || \mathbb{P}_\theta) = \arg\max_{\theta \in \Theta} \frac{1}{m} \sum_{j=1}^m \log p(\mathbf{y}_j | \theta). \tag{5.20}$$

This can be derived as strictly proper scoring rule from the log-score: $S_{\text{KL}}(\mathbf{x}, \mathbb{P}) = -\log p(\mathbf{x})$. Since it is a strictly proper scoring rule, we can trivially show that maximum likelihood estimation is consistent in the M-closed case.

In the case of exponential family models, the problem of maximum likelihood estimation can be simplified significantly. In this case, $\nabla_\theta S(\mathbf{x}, \mathbb{P}_\theta) = -\nabla_\theta \log p(\mathbf{x}|\theta) = -T(\mathbf{x}) + \nabla_\theta c(\theta)$, and so maximum likelihood estimation is equivalent to solving the following estimation equations: $\sum_{j=1}^m T(\mathbf{y}_j) = -\nabla_\theta c(\theta)$. Clearly this requires knowledge of the normalisation constant of the model or, more precisely, of the derivative of the log normalisation constant. Maximum likelihood estimation will hence not be feasible in cases where this constant is not available in closed form.

Since minimum distance estimators are based on parametric families and divergences, the performance of these estimators will be closely interlinked with the geometry of the corresponding statistical manifold. The metric tensor obtained from the KL-divergence is called the Fisher information metric. It corresponds to the covariance of the score vectors of the distribution:

$$g_{jk}^{\mathrm{KL}}(\theta) \;\; = \;\; \int_{\mathcal{X}} \left(\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_j} \right) \left(\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_k} \right) p(\mathbf{x}|\theta) \mathrm{d}\mathbf{x}.$$

Geometric quantities can be useful to understand asymptotic properties of the estimator. The most common example of this is the Cramer-Rao theorem (see for example Amari [2016], Theorem 7.7) which states that for any asymptotically unbiased estimator $\hat{\theta}$ of θ, we have: $\mathbb{E}[(\hat{\theta}_j - \theta_j)(\hat{\theta}_k - \theta_k)] \geq (1/m) g_{jk}^{\mathrm{KL}}$, where the expectation is taken with respect to the distribution of the data-generating process. Since maximum likelihood estimation attains this lower bound, we say that it is efficient. Unfortunately, as previously mentioned, efficiency often has to be traded with robustness and maximum likelihood estimation is not robust. This can be noticed by looking at the influence function (obtained by plugging in S_{KL} into Equation 5.19):

$$\begin{aligned} \mathrm{IF}_{\mathrm{KL}}(\mathbf{z}, \mathbb{P}_\theta) \;\; &= \;\; \left(- \int_{\mathcal{X}} \nabla_\theta \nabla_\theta \log p(\mathbf{x}|\theta) p(\mathbf{x}|\theta) \mathrm{d}\mathbf{x} \right)^{-1} (-\nabla_\theta \log p(\mathbf{z}|\theta)) \\ &= \;\; g^{\mathrm{KL}}(\theta)^{-1} \nabla_\theta \log p(\mathbf{z}|\theta). \end{aligned}$$

Even for simple models such as a Gaussian distribution with unknown standard deviation, the influence function will be $O(\mathbf{z})$ and hence unbounded, clearly demonstrating the lack of bias-robustness of maximum likelihood estimation.

Maximum likelihood methods have nonetheless been widely popular in the past due to the likelihood principle [Young and Smith, 2005], which states that, given a model, all of the evidence in a data set which is relevant to parameter inference is contained in the likelihood function. There are however several limitations to this approach, the most obvious being the requirement to have access to the likelihood (or equivalently the log-likelihood). We will now highlight alternative loss functions for use when the likelihood is not available.

5.2.2 Estimators for Unnormalised Models

A first scenario which is common in statistics is when the likelihood $p(\mathbf{x}|\theta)$ is not available due to an unknown normalisation constant $Z(\theta)$ (which depends on the parameter vector θ). Usually this is due to the high computational cost of evaluating the normalisation constant, or because this constant is itself defined as some intractable integral. In this case, the optimisation problem in Equation 5.20 cannot be solved since the normalisation constant depends on θ but is unavailable, and maximum likelihood estimation is hence not feasible.

In this section, we will discuss classes of estimators which can by-pass the need for normalisation constants. The first estimators discussed are the score-matching estimatorswhich are extensively used in machine learning. These will be formally discussed in the context of minimum distance estimators, will be shown to both originates from the notion of Stein discrepancy. Once this connection is made, we will discuss estimators based on other underlying Stein classes, such as kernel spaces.

Score Matching Estimators

The issue of intractable normalisation constants has led to the development of statistical inference methods based on the score function. This is because the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta)$ does not depend on $Z(\theta)$, and so can be evaluated even when the likelihood is unnormalised. This is a major advantage since it allows us to by-pass the computation of expensive normalisation constants whilst still obtaining an asymptotically exact solution. An example of divergence based on the score function is the score-matching divergence (SM) [Hyvärinen, 2006, 2007], also called the Hyvärinen or Fisher divergence, and which is defined as:

$$D_{\mathrm{SM}}(\mathbb{P}_1||\mathbb{P}_2) \quad := \quad \int_{\mathcal{X}} \|\nabla_{\mathbf{x}} \log p_1(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_2(\mathbf{x})\|_2^2 \, \mathbb{P}_1(\mathrm{d}\mathbf{x}). \tag{5.21}$$

This divergence can also be generalised to include higher-order derivatives of the log-likelihood; see Lyu [2009]. Using integration by parts, Hyvärinen [2006] (Theorem 1) showed that the SM divergence can be rewritten in a convenient form when considered as a function of $\theta \in \Theta \subseteq \mathbb{R}^p$: $D_{\mathrm{SM}}(\mathbb{Q}||\mathbb{P}_\theta) = \int_{\mathcal{X}} \big(\Delta_{\mathbf{x}} \log p(\mathbf{x}|\theta) + \frac{1}{2}\|\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta)\|_2^2\big)\mathbb{Q}(\mathrm{d}\mathbf{x}) + C$ for some $C \in \mathbb{R}$ which does not depend on θ. The SM estimator minimises this divergence over $\theta \in \Theta$ and hence clearly does not depend on the intractable constant C:

$$\hat{\theta}_m^{\mathrm{SM}} \quad = \quad \operatorname*{arg\,min}_{\theta \in \Theta} D_{\mathrm{SM}}(\mathbb{Q}^m||\mathbb{P}_\theta), \tag{5.22}$$

$$D_{\mathrm{SM}}(\mathbb{Q}^m||\mathbb{P}_\theta) \quad = \quad \frac{1}{m}\sum_{l=1}^{d}\sum_{j=1}^{m} \Delta_{\mathbf{y}} \log p(\mathbf{y}_j|\theta) + \frac{1}{2}\|\nabla_{\mathbf{y}} \log p(\mathbf{y}_j|\theta)\|_2^2. \tag{5.23}$$

The SM estimator can also be derived from a strictly proper scoring rule of the form: $S_{\mathrm{SM}}(\mathbf{x}, \mathbb{P}) = \Delta_{\mathbf{x}} \log p(\mathbf{x}) + \frac{1}{2}\|\nabla_{\mathbf{x}} \log p(\mathbf{x})\|_2^2$. A direct implication is that the

SM estimator is consistent in the M-closed case (since the estimating equations are unbiased), although it may not necessarily be efficient. The SM estimator is clearly a local scoring rule, since it depends on $\mathbb{P}$ only through derivatives of $\log p$.

It is however not necessarily bias-robust. Take the case of a one-dimensional Gaussian distribution with mean zero and unknown standard deviation. Then $\nabla_\theta S_{\mathrm{SM}}(z, \mathbb{P}_\theta) = (2 + z^2)/\theta^3$, so the influence function is $\mathrm{IF}_{\mathrm{SM}}(z, \mathbb{P}_\theta) = O(z^2)$ and is hence clearly unbounded.

The metric tensor for the Hyvärinen divergence was derived by [Karakida et al., 2016] and is given by: $g^{\mathrm{SM}}(\theta) = \int_{\mathcal{X}} (\nabla_\theta \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta))(\nabla_\theta \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta))^\top \mathbb{P}_\theta(\mathrm{d}\mathbf{x})$, where $(\nabla_\theta \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta))_{jk} = \partial^2 \log p(\mathbf{x}|\theta)/\partial x_j \partial \theta_k$. The SM estimators have been shown to be useful for a variety of applications, including imaging models [Koster and Hyvärinen, 2009; Kingma and LeCun, 2010; Swersky et al., 2011], directional statistics [Mardia et al.] and point processes [Sahani et al., 2016]. They have also be shown to be connected to popular inference methods for denoising autoencoders [Vincent, 2011]. They do however have important failure modes, most notably in the case of mixtures [Wenliang et al., 2018].

An interesting fact, first pointed out in an open-access version of Sriperumbudur et al. [2017] and later in Forbes and Lauritzen [2015], is that we can compute the SM estimator for exponential families in closed form. Define the following summary statistics:

- $A(\{\mathbf{y}_j\}_{j=1}^m) = \frac{1}{m} \sum_{j=1}^m \sum_{l=1}^d \frac{\partial T(\mathbf{y}_j)}{\partial (\mathbf{y}_i)_l} \left(\frac{\partial T(\mathbf{y}_j)}{\partial (\mathbf{y}_i)_l} \right)^\top$,
- $B(\{\mathbf{y}_j\}_{j=1}^m) = \frac{1}{m} \sum_{j=1}^m \sum_{l=1}^d \frac{\partial b(\mathbf{y}_j)}{\partial (\mathbf{y}_i)_l} \left(\frac{\partial T(\mathbf{y}_j)}{\partial (\mathbf{y}_i)_l} \right)^\top + \Delta T(\mathbf{y}_j)$,
- $C(\{\mathbf{y}_j\}_{j=1}^m) = \frac{1}{m} \sum_{j=1}^m \frac{1}{2} \|\nabla_{\mathbf{y}} b(\mathbf{y}_j)\|_2^2 + \Delta_{\mathbf{y}} b(\mathbf{y}_j)$.

where b, T are given in Equation 5.16 and $(\mathbf{y}_i)_l$ denotes the l^{th} component of the vector $\mathbf{y}_i$. Then the divergence can be written as a quadratic form, and the estimating equations become linear θ:

$$\begin{aligned} D_{\mathrm{SM}}(\mathbb{Q}^m || \mathbb{P}_\theta) &= \frac{1}{2} \theta^\top A(\{\mathbf{y}_j\}_{j=1}^m) \theta + B(\{\mathbf{y}_j\}_{j=1}^m)^\top \theta + C(\{\mathbf{y}_j\}_{j=1}^m), \\ \hat{\theta}_m^{\mathrm{SM}} &= -B(\{\mathbf{y}_j\}_{j=1}^m) A(\{\mathbf{y}_j\}_{j=1}^m)^{-1}. \end{aligned}$$

The expressions above are particular useful as they circumvent the need for numerical optimisation routines.

Note that Sriperumbudur et al. [2017] generalises the score matching loss to the case where the sufficient statistic and natural parameter are infinite dimensional, and proves consistency with finite sample bounds, both for the infinite and finite dimensional cases. Proposition 1 and 2 in Forbes and Lauritzen [2015] also independently show consistency of these estimators, and provide a central limit theorem.

Minimum Stein Discrepancy Estimators

We propose to generalise the SM methodology originally proposed by Hyvärinen [2006]. Our proposed approach is to consider Stein discrepancies which are based on the score function, or higher derivatives of the log likelihood, within a minimum distance estimator framework. More precisely, we consider estimators of the form $\hat{\theta}_m^{\text{Stein}} = \arg\min_{\theta\in\Theta} D_{\text{Stein}}(\mathbb{Q}^m||\mathbb{P}_\theta)^2$ where

$$D_{\text{Stein}}(\mathbb{Q}^m||\mathbb{P}_\theta) \;\; := \;\; \sup_{g\in\mathcal{G}} \left| \frac{1}{m}\sum_{j=1}^{m} \mathcal{T}_{\mathbb{P}_\theta}[g](\mathbf{y}_j) \right|,$$

$\mathcal{G}$ is a Stein class and $\mathcal{T}_{\mathbb{P}_\theta}$ is a Stein operator adapted to $\mathbb{P}_\theta$. Two potential choices of Stein operators which would not require normalisation of the likelihoods are the Langevin Stein operator and Itô Stein operator. This approach is a generalisation of the SM estimators since the Hyvärinen divergence is a Stein discrepancy:

Proposition 12 (Score-Matching Estimators as Minimum Stein Discrepancy Estimators). *Let $\mathcal{X} \subseteq \mathbb{R}^d$ for $d \in \mathbb{N}$ and consider the Stein operator $\mathcal{S}_{\mathbb{P}}$ in Equation 5.6 for some function $a : \mathcal{X} \to \mathbb{R}^{d\times d}$ taking values in the space of positive semi-definite matrices and function $c : \mathcal{X} \to \mathbb{R}^{d\times d}$ taking values in the space of skew-symmetric matrices. Let $m(\mathbf{x}) = a(\mathbf{x}) + c(\mathbf{x})$ and define the Stein class:*

$$\mathcal{G} := \{g = (g_1,\ldots,g_d) \in (C^1(\mathcal{X}) \cap L^2(\mathcal{X};\mathbb{Q}))^d : \|g_j\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1 \; \forall j = 1,\ldots,d\}.$$

Then, we get a diffusion-based Stein discrepancy of the form:

$$D(\mathbb{Q}||\mathbb{P}_\theta) \;\; = \;\; \int_{\mathcal{X}} \|(\nabla_{\mathbf{x}} \log p_2(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_1(\mathbf{x}))\, m(\mathbf{x})\|_2^2\, p_1(\mathbf{x})\mathrm{d}\mathbf{x} \qquad (5.24)$$

and can obtain a diffusion-based minimum Stein discrepancy estimator:

$$\begin{aligned}\hat{\theta} \;\; = \;\; & \arg\min_{\theta\in\Theta} \int_{\mathcal{X}} \|m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta)\|_2^2 \mathbb{Q}(\mathrm{d}\mathbf{x}) \\ & +2\int_{\mathcal{X}} \left\langle \nabla_{\mathbf{x}}, m(\mathbf{x})m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) \right\rangle \mathbb{Q}(\mathrm{d}\mathbf{x})\end{aligned}$$

In particular, we note that the contrusction above generalises score-matching estimators. Indeed, the score-matching estimator of Hyvärinen [2006] is a diffusion-based minimum Stein discrepancy estimator where $a(\mathbf{x}) = I_{d\times d}$ and $c(\mathbf{x}) = 0$, the non-negative score matching estimator of Hyvärinen [2007] is a diffusion-based minimum Stein discrepancy estimator where $a(\mathbf{x}) = \text{diag}((x_1, \ldots, x_d))$ and $c(\mathbf{x}) = 0$ and finally, the generalised non-negative score matching estimator of Yu et al. [2018] is a diffusion-based minimum Stein discrepancy estimator where $a(\mathbf{x}) = \text{diag}((h_1(x_1)^{\frac{1}{2}}, \ldots, h_d(x_d)^{\frac{1}{2}})$ and $c(\mathbf{x}) = 0$.

Langevin Kernel Stein Discrepancy Estimators

The Hyvärinen divergence is not the only possible choice of Stein discrepancy which can be used for estimation. One drawback of the Hyvärinen divergence is the need for second derivatives of the log-likelihood with respect to the data. It turns out that the KSD with Langevin Stein operator can help us get rid of this requirement. This choice of Stein discrepancy gives us the following estimator:

$$\hat{\theta}^{\text{KSD}} \;=\; \arg\min_{\theta\in\Theta} \text{KSD}(\mathbb{Q}||\mathbb{P}_\theta)^2 \;=\; \arg\min_{\theta\in\Theta} \int_{\mathcal{X}}\int_{\mathcal{X}} k_{\mathbb{P}_\theta}(\mathbf{x},\mathbf{y})\mathbb{Q}(\mathrm{d}\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{y}).$$

where, in this case, the Stein reproducing kernel is based on the Langevin Stein operator adapted to $\mathbb{P}_\theta$ and is of the form:

$$\begin{aligned} k_{\mathbb{P}_\theta}(\mathbf{x},\mathbf{y}) \;&=\; \langle \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), \nabla_{\mathbf{y}} \log p(\mathbf{y}|\theta)\rangle k(\mathbf{x},\mathbf{y}) + \langle \nabla_1, \nabla_2 k(\mathbf{x},\mathbf{y})\rangle \\ &\quad + \langle \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), \nabla_2 k(\mathbf{x},\mathbf{y})\rangle + \langle \nabla_{\mathbf{y}} \log p(\mathbf{y}|\theta), \nabla_1 k(\mathbf{x},\mathbf{y})\rangle. \end{aligned} \tag{5.25}$$

Sufficient conditions for $k_{\mathbb{P}_\theta}$ to be a characteristic Stein kernel, and as a by-product for the KSD with Langevin Stein operator to be a divergence, were summarised in Section 5.1.2. When $k_{\mathbb{P}_\theta}$ is a characteristic Stein kernel, we get the following metric tensor:

Proposition 13 (**Information Metric of the Kernel Stein Discrepancy with Langevin Stein Operator**). *Consider a KSD based on a Langevin Stein operator adapted to some measure $\mathbb{P}_\theta$, with density $p(\mathbf{x}|\theta)$, and base kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Assume $k_{\mathbb{P}_\theta}$ is a characteristic Stein reproducing kernel. The information metric corresponding to this divergence is given by:*

$$g_{jk}(\theta) \;=\; 2\sum_{l=1}^{d} \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y}) \frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial x_l \partial \theta_j} \frac{\partial^2 \log p(\mathbf{y}|\theta)}{\partial y_l \partial \theta_k} \mathbb{P}_\theta(\mathrm{d}\mathbf{x})\mathbb{P}_\theta(\mathrm{d}\mathbf{y}).$$

Given IID realisations $\{\mathbf{y}_j\}_{i=1}^m$ from the model of interest $\mathbb{Q}$, we can obtain an unbiased estimate of the square of the KSD [Liu et al., 2016, Equation 14] using a U-statistic, which leads to a computable estimator:

$$
\begin{aligned}
\mathrm{KSD}_U(\mathbb{Q}^m||\mathbb{P}_\theta)^2 &= \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k_{\mathbb{P}_\theta}(\mathbf{y}_i,\mathbf{y}_j), && (5.26)\\
\hat{\theta}_m^{\mathrm{KSD}} &= \arg\min_{\theta\in\Theta} \mathrm{KSD}_U(\mathbb{Q}^m||\mathbb{P}_\theta)^2. && (5.27)
\end{aligned}
$$

This estimator is particularly useful as it removes the need for second order derivatives of the log-likelihood with respect to the data. The added flexibility of the kernel can also be an advantage.

Similarly to SM, it is possible to obtain a closed form expression for the KSD estimator in the case of statistical models in some exponential family.

Proposition 14 (Kernel Stein Discrepancy for Exponential Family). *Assume $k_{\mathbb{P}_\theta}$ is a characteristic Stein reproducing kernel adapted to some element $\mathbb{P}$ from some exponential family and constructed with the Langevin Stein operator. Define the following summary statistics:*

$$
\begin{aligned}
A(\{\mathbf{y}_j\}_{j=1}^m) &= \frac{1}{m(m-1)}\sum_{i\neq j}^{m}\langle\nabla T(\mathbf{y}_i),\nabla b(\mathbf{y}_j)\rangle k(\mathbf{y}_i,\mathbf{y}_j)+\langle\nabla T(\mathbf{y}_i),\nabla_2 k(\mathbf{y}_i,\mathbf{y}_j)\rangle\\
&\qquad +\langle\nabla T(\mathbf{y}_j),\nabla b(\mathbf{y}_i)\rangle k(\mathbf{y}_i,\mathbf{y}_j)+\langle\nabla T(\mathbf{y}_j),\nabla_1 k(\mathbf{y}_j,\mathbf{y}_i)\rangle.\\
B(\{\mathbf{y}_j\}_{j=1}^m) &= \frac{1}{m(m-1)}\sum_{i\neq j}^{m}\langle\nabla T(\mathbf{y}_i),\nabla T(\mathbf{y}_j)\rangle k(\mathbf{y}_i,\mathbf{y}_j),\\
C(\{\mathbf{y}_j\}_{j=1}^m) &= \frac{1}{m(m-1)}\sum_{i\neq j}^{m}\langle\nabla_1 k(\mathbf{y}_i,\mathbf{y}_j),\nabla b(\mathbf{y}_j)\rangle+\langle\nabla_2 k(\mathbf{y}_i,\mathbf{y}_j),\nabla b(\mathbf{y}_i)\rangle\\
&\qquad +\langle\nabla_1,\nabla_2 k(\mathbf{y}_i,\mathbf{y}_j)\rangle+\langle\nabla b(\mathbf{y}_i),\nabla b(\mathbf{y}_j)\rangle k(\mathbf{y}_i,\mathbf{y}_j).
\end{aligned}
$$

Then, the U-statistic approximation of the KSD based on the Langevin Stein operator and its corresponding estimator are given by:

$$
\begin{aligned}
KSD_U(\mathbb{Q}^m||\mathbb{P}_\theta)^2 &= \theta^\top A(\{\mathbf{y}_j\}_{j=1}^m)\theta+B(\{\mathbf{y}_j\}_{j=1}^m)\theta+C(\{\mathbf{y}_j\}_{j=1}^m),\\
\hat{\theta}_m^{KSD} &= -B(\{\mathbf{y}_j\}_{j=1}^m)A(\{\mathbf{y}_j\}_{j=1}^m)^{-1}.
\end{aligned}
$$

Once again, these expressions will be particularly useful as they allow us to avoid the use of numerical optimisation routines. Note that similar work has recently appeared in Li and Turner [2018], but the aim in that work is to estimate the scores rather than the parameters.

5.2.3 Estimators for Generative Models

We have now completed our introduction to KSD estimators for unnormalised models. A second case of interest is that of generative models. These models have the particular feature that their likelihood cannot be evaluated in closed form, but they can instead be sampled from. Let $(\mathcal{U}, \Sigma_{\mathcal{U}}, \mathbb{U})$ be a probability space. Formally we regard generative models as a family of probability measures $\mathcal{P}_\Theta(\mathcal{X})$ such that for any $\theta \in \Theta$, we can obtain some IID realisations $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ from $\mathbb{P}_\theta$. These realisations are obtained in two steps: first sample IID random variables $\{\mathbf{u}_i\}_{i=1}^n$ from $\mathbb{U}$, then apply some map $G_\theta : \mathcal{U} \to \mathcal{X}$ to each of these realisations to obtain $\mathbb{P}_\theta$ distributed random variables; i.e. $\mathbf{x}_i = G_\theta(\mathbf{u}_i)$ for $i = 1, \ldots, n$. Alternatively, we say $\mathbb{P}_\theta$ is the pushforward of $\mathbb{U}$ under G_θ.

Examples of minimum distance estimators for generative models include estimators based on approximations of the Wasserstein distance [Basu et al., 1998; Bassetti et al., 2006; Genevay et al., 2018]. In this section, we focus instead on a kernel-based estimator related to MMD.

Minimum Maximum Mean Discrepancy Estimators

We propose to use an approximation of the square of the MMD within a minimum distance estimator framework. Once again, this minimum distance estimator originates from a scoring rule, called kernel scoring rule in the literature. The scoring rule which leads to the MMD is well known in the literature [Eaton, 1982; Dawid, 2007; Huszár, 2013; Zawadzki and Lahaie, 2015; Steinwart and Ziegel, 2017; Masnadi-Shirazi, 2017], and takes the form:

$$\begin{aligned} S(\mathbf{x}, \mathbb{P}_\theta) &= k(\mathbf{x}, \mathbf{x}) - 2\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y})\mathbb{P}_\theta(\mathrm{d}\mathbf{y}) + \int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{y}, \mathbf{z})\mathbb{P}_\theta(\mathrm{d}\mathbf{y})\mathbb{P}_\theta(\mathrm{d}\mathbf{z}) \qquad (5.28) \\ &= k(\mathbf{x}, \mathbf{x}) - 2\int_{\mathcal{X}} k(\mathbf{x}, G_\theta(\mathbf{u}))\mathbb{U}(\mathrm{d}\mathbf{u}) + \int_{\mathcal{X}\times\mathcal{X}} k(G_\theta(\mathbf{u}), G_\theta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}). \end{aligned}$$

There is also ample evidence of its applicability to complex generative models, due to the recent line of work on MMD generative adversarial networks [Dziugaite et al., 2015; Li et al., 2015, 2017; Sutherland et al., 2017; Arbel et al., 2018]. Suppose k is characteristic, then the MMD is a divergence, and the scoring rule is a strictly proper scoring rule. However, the scoring rule is not local since it depends on k. The information metric for this divergence is given by:

Proposition 15 (**Information Metric of the Maximum Mean Discrepancy Squared**). *Suppose $\mathbb{P}_\theta$ is a generative model, defined as the pushforward of $\mathbb{U}$ under*

G_θ. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a characteristic kernel. Then the MMD squared is a divergence with associated information metric given by:

$$g(\theta) \;=\; \int_{\mathcal{U}}\int_{\mathcal{U}} \nabla_\theta G_\theta(\mathbf{u})^\top \nabla_1\nabla_2 k(G_\theta(\mathbf{u}), G_\theta(\mathbf{v}))\nabla_\theta G_\theta(\mathbf{v})\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}).$$

where $(\nabla_1\nabla_2 k(\mathbf{x},\mathbf{y}))_{jk} = \partial^2 k(\mathbf{x},\mathbf{y})/\partial x_j \partial y_k$.

Unfortunately, the MMD squared cannot be computed in closed form, but it can be approximated using a U-statistic, and a corresponding estimator can be obtained:

$$\begin{aligned}\hat{\theta}_m^{\mathrm{MMD}} &= \arg\min_{\theta\in\Theta} \mathrm{MMD}_U^2(\mathbb{Q}^m, \mathbb{P}_\theta),\\ \mathrm{MMD}_U^2(\mathbb{Q}^m, \mathbb{P}_\theta) &= \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})\mathbb{P}_\theta(\mathrm{d}\mathbf{x})\mathbb{P}_\theta(\mathrm{d}\mathbf{y}) - \frac{2}{m}\sum_{j=1}^m \int_{\mathcal{X}} k(\mathbf{x},\mathbf{y}_j)\mathbb{P}_\theta(\mathrm{d}\mathbf{x})\\ &\quad + \frac{1}{m(m-1)}\sum_{j\neq j'} k(\mathbf{y}_j,\mathbf{y}_{j'}).\end{aligned}$$

In practice we may not be able to compute expectations with respect to $\mathbb{P}_\theta$ exactly so the loss $\mathrm{MMD}(\mathbb{Q}^m, \mathbb{P}_\theta)$ is intractable. However if the generative map G_θ is sufficiently cheap to evaluate, then approximations via Monte Carlo estimation is feasible by generating $n \gg m$ samples. In this regime, it may be of interest to first consider the estimator $\hat{\theta}_m^{\mathrm{MMD}}$ to understand the behaviour in the limit of large data size. Alternatively, if the generative map is expensive to evaluate, then we would expect the number of realisations of the generative model to be roughly of the same order as the number of data points. In this case, fluctuations arising from the both the approximation of generative distribution $\mathbb{P}_\theta^n$ and data distribution $\mathbb{Q}^m$ will affect the efficiency. We thus study a second U-statistic approximation of the MMD, as well as its corresponding minimum distance estimator:

$$\begin{aligned}\hat{\theta}_{n,m}^{\mathrm{MMD}} &= \arg\min_{\theta\in\Theta} \mathrm{MMD}_{U,U}^2(\mathbb{Q}^m, \mathbb{P}_\theta^n)\\ \mathrm{MMD}_{U,U}^2(\mathbb{Q}^m, \mathbb{P}_\theta^n) &= \frac{1}{n(n-1)}\sum_{i\neq i'} k(\mathbf{x}_i,\mathbf{x}_{i'}) - \frac{2}{mn}\sum_{j=1}^m\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{y}_j)\\ &\quad + \frac{1}{m(m-1)}\sum_{j\neq j'} k(\mathbf{y}_j,\mathbf{y}_{j'}).\end{aligned}$$

An interesting point is that MMD estimators will usually be bias-robust. Take our example of a univariate Gaussian model with unknown standard deviation. In this case, using a Gaussian kernel $k(x,y) = \exp(-(x-y)^2/2\sigma^2)$, the scoring rule can be derived in closed form using Gaussian identities (see for example Appendix C of Briol et al. [2015a] or Example 3 in Sriperumbudur et al. [2012]). This allows us to notice that $\text{IF}_{\text{MMD}}(z, \mathbb{P}_\theta) = O\left((z^2/\theta + \sigma^2)\exp(-z^2(2\theta^2 + 2\sigma^2)^{-1})\right)$, which is clearly bounded in z and is therefore bias-robust. Furthermore, the choice of lengthscale σ will impact the gross-error sensitivity. More generally, MMD estimators will be bias-robust under the following assumptions:

Proposition 16 (Bias-Robustness of Maximum Mean Discrepancy Estimation). *Consider an MMD estimator for a model $\mathbb{P}_\theta$, seen as the pushforward of some measure $\mathbb{U}$ through the parametric map G_θ, based on the reproducing kernel k. Assume that (i) k is characteristic, (ii) $\| - 2\int_{\mathcal{U}} \nabla_\theta k(\mathbf{x}, G_\theta(\mathbf{u}))\mathbb{U}(\mathrm{d}\mathbf{u}) + \int_{\mathcal{U}}\int_{\mathcal{U}} \nabla_\theta k(G_\theta(\mathbf{u}), G_\theta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v})\|_\infty < \infty \; \forall \mathbf{x} \in \mathcal{X}$, and (iii) the matrix given by $\int_{\mathcal{U}}\int_{\mathcal{U}} \nabla_\theta \nabla_\theta k(G_\theta(\mathbf{u}), G_\theta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v})$ is invertible. Then, minimum MMD estimators are bias-robust.*

Although these conditions might be challenging to check on a case by case basis, they only usually required assumptions on the tails of the kernel and generative map, as well as that of their derivatives with respect to the parameter. The conditions can hence be useful for selecting kernels.

Unfortunately, since the MMD estimators are not based on the score function, the estimating equations will not be linear and it will not usually be possible to solve them explicitly in the case of exponential families.

5.2.4 Practical Considerations

As will now be obvious, the choice of loss function (or equivalently of kernel and kernel hyperparameters), and of numerical optimisation routine will be of great importance for practical implementation. These tuning choices will influence the performance of our estimators in three main ways: the asymptotic efficiency of the estimator, the robustness of the estimator and the difficulty of optimising the loss function.

Numerical Optimisation

Recall our goal of inferring the parameter $\theta \in \Theta$ by minimising the loss function $L(\theta) = \text{KSD}(\mathbb{Q}||\mathbb{P}_\theta)^2$ or $L(\theta) = \text{MMD}(\mathbb{Q}, \mathbb{P}_\theta)^2$. These loss functions are usually non-convex, and potentially high-dimensional when the parameter space is large, which might lead to computational challenges for practical implementation.

Several optimisation algorithms could be used, but we propose to focus on gradient-based methods. A common approach is gradient descent, which consists of initialising at $\theta_0 \in \Theta$, then iterating over descent steps. For the t^{th} iteration, we have the following update:

$$\theta^{(t)} = \theta^{(t-1)} - \eta_t \nabla_\theta L(\theta^{(t-1)}),$$

where $\{\eta_t\}_{t\in\mathbb{N}}$ is a step size sequence chosen to guarantee convergence to a local minimum. To use gradient descent for our minimum distance estimators based on KSDs and MMDs, we will need the gradient of these statistical divergences with respect to the parameters of the model:

Proposition 17 (Gradients of the Kernel Stein Discrepancy and Maximum Mean Discrepancy Loss Functions). *The gradient of the KSD loss function $L^{KSD}(\theta) := KSD(\mathbb{Q}||\mathbb{P}_\theta)^2$ is given by:*

$$\begin{aligned}\nabla_\theta L^{KSD}(\theta) \;=\; & \int_{\mathcal{X}}\int_{\mathcal{X}} \Big[k(\mathbf{x},\mathbf{y})\nabla_\theta\nabla_{\mathbf{x}}\log p(\mathbf{x}|\theta)\nabla_{\mathbf{y}}\log p(\mathbf{y}|\theta) \\ & + k(\mathbf{x},\mathbf{y})\nabla_\theta\nabla_{\mathbf{y}}\log p(\mathbf{y}|\theta)\nabla_{\mathbf{x}}\log p(\mathbf{x}|\theta) + \big(\nabla_\theta\nabla_{\mathbf{x}}\log p(\mathbf{x}|\theta)\big)\nabla_2 k(\mathbf{x},\mathbf{y}) \\ & + \nabla_\theta\nabla_{\mathbf{y}}\log p(\mathbf{y}|\theta)\nabla_1 k(\mathbf{x},\mathbf{y})\Big]\mathbb{Q}(\mathrm{d}\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{y}).\end{aligned}$$

The gradient of the MMD loss function $L^{MMD}(\theta) := MMD(\mathbb{Q},\mathbb{P}_\theta)^2$, when $\mathbb{P}_\theta$ is the pushforward of some base measure $\mathbb{U}$ through the map G_θ, is given by:

$$\begin{aligned}\nabla_\theta L^{MMD}(\theta) \;=\; & \int_{\mathcal{U}}\int_{\mathcal{U}} \nabla_\theta G_\theta(\mathbf{u})\left(\nabla_1 k(G_\theta(\mathbf{u}),G_\theta(\mathbf{v})) + \nabla_2 k(G_\theta(\mathbf{v}),G_\theta(\mathbf{u}))\right)^\top \mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) \\ & -2\int_{\mathcal{U}}\int_{\mathcal{X}} \nabla_\theta G_\theta(\mathbf{u})\nabla_1 k(G_\theta(\mathbf{u}),\mathbf{y})\mathbb{U}(\mathrm{d}\mathbf{u})^\top\mathbb{Q}(\mathrm{d}\mathbf{y}).\end{aligned}$$

The usual way to motivate gradient descent methods is to say that they sequentially decrease the objective function in the optimal direction. However, since we are optimising within a specific statistical manifold, notions of distance are different from Euclidean spaces and the classic gradient descent algorithm does not decrease the objective in an optimal direction anymore. The algorithm is still a valid optimisation algorithm, but it will require a large number of iterations to attain the minimum of the loss function. On a manifold, the gradient vector of the function L (i.e. the optimal descent direction) is given by the vector field $\nabla^g L$, which in a local coordinate system is: $\nabla^g L(\theta) = g^{-1}(\theta)\nabla_\theta L(\theta)$, where $g^{-1}(\theta)$ is the inverse of the matrix $g(\theta)$. The natural generalisation of the gradient descent is then to follow the geodesics of the manifold: $\mathbf{p}_t = \exp_{\mathbf{p}_{t-1}}(-\eta_t \nabla^g L(\mathbf{p}_{t-1}))$, where exp maps the tangent vector v_p at some $p \in \mathcal{M}$ to the point $\exp_p(v_p) := \gamma(1) \in \mathcal{M}$, where γ is the unique geodesic s.t. $\gamma(0) = p$, $\dot{\gamma} = v_p$. However it is often hard to follow geodesics exactly in practice. Instead, the Euclidean space formula for the exponential, $\exp_\theta \mathbf{v} = \theta + \mathbf{v}$, suggests the following iterations: $\theta^{(t)} = \theta^{(t-1)} - \eta_t \nabla^g L(\theta^{(t-1)})$.

This corresponds to natural gradient algorithms [Amari, 1998, 2016]. These algorithms were previously introduced in the context of the KL divergence and SM divergence, but we can straightforwardly derive similar algorithms for the KSD and MMD using some of our previously-derived formulae for information metrics in Propositions 13 and 15. Raskutti and Mukherjee [2015] also reinterpreted this algorithm as a mirror descent algorithm, whilst Pascanu and Bengio [2014]; Martens [2014] demonstrated its connections to many popular optimisation algorithms for training large machine learning models.

For minimum distance estimation based on KSD, we can use the information metric derived in Proposition 13 and the gradient derived in Proposition 17. Since these will not be available in closed form, we can approximate the trajectories of the natural gradient algorithm by using U-statistic approximations of these quantities. In this case, we target the minimiser of $\hat{L}_U^{\mathrm{KSD}}(\theta) = \mathrm{KSD}_U^2(\mathbb{Q}^m || \mathbb{P}_\theta)$. The gradient descent algorithm follows the iterations $\theta^{(t)} = \theta^{(t-1)} - \eta_t \nabla_\theta \hat{L}_U^{\mathrm{KSD}}(\theta^{(t-1)})$ whilst the natural gradient descent algorithm follows the iterations

$$\theta^{(t)} \quad = \quad \theta^{(t-1)} - \eta_t (g_U^{\mathrm{KSD}}(\theta^{(t-1)}))^{-1} \nabla_\theta \hat{L}_U^{\mathrm{KSD}}(\theta^{(t-1)}).$$

On the other hand, for minimum distance estimation based on MMD, we propose to target the minimiser of $\hat{L}_{U,U}^{\mathrm{MMD}}(\theta) = \mathrm{MMD}_{U,U}^2(\mathbb{Q}^m, \mathbb{P}_\theta^n)$. To do so, we can use U-statistic approximations of the information metric in Proposition 15 and gradient in Proposition 17. The gradient descent algorithm follows the iterations $\theta^{(t)} = \theta^{(t-1)} - \eta_t \nabla_\theta \hat{L}_{U,U}^{\mathrm{MMD}}(\theta^{(t-1)})$ whilst the natural gradient descent algorithm follows:

$$\theta^{(t)} \quad = \quad \theta^{(t-1)} - \eta_t (g_{U,U}^{\mathrm{MMD}}(\theta^{(t-1)}))^{-1} \nabla_\theta \hat{L}_{U,U}^{\mathrm{MMD}}(\theta^{(t-1)}).$$

Note that we cannot expect to find a unique minimum to this problem since realisations from the generative model are obtained at each iteration. This is necessary in order to compute $\hat{L}_{U,U}^{\mathrm{MMD}}$ at a new parameter value, but implies that a different objective function is optimised at each iteration. Consistency properties of MMD estimators and of U-statistics approximations do however guarantee that the algorithm above will move towards a minimum of the idealised loss function $\mathrm{MMD}(\mathbb{Q}, \mathbb{P}_\theta)^2$ as n and m grow.

The U-statistic approximations might lead to parameter values outside the domain Θ, so we also introduce a projection operator $\text{Proj} : \mathbb{R}^p \to \Theta$, applied after each step, and which maps parameters to their closest value in Θ in terms of some norm to be defined. The approximation may also mean that the metric tensor is not invertible, and regularisation might be required to resolve this issue. Finally, using the same samples for the U-statistic of the gradient and of the information metric may lead to strong biases.

In any case, finding the minimiser may still be challenging for several reasons:

1. The gradient descent procedures will be expensive in certain settings. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^p$. For KSD estimation, the cost of each iteration of gradient descent is $O(mpd)$, whilst the cost of each iteration of natural gradient descent is $O(m^2p^2d + p^3)$. On the other hand, for MMD estimation the cost of each iteration of gradient descent is $O\left((n^2 + nm)pd^2\right)$ whilst the cost of each natural gradient descent iteration is $O\left((n^2 + nm)p^2d^2 + p^3\right)$. The cost for MMD estimation is linear in the number of data points m, but quadratic in the number of simulated samples n. We note that taking $n = m$ is optimal in terms of computational cost.

   This computational cost could be made linear in n by considering approximations of the MMD or KSD as found in Chwialkowski et al. [2015]; Jitkrittum et al. [2017]. In large data settings, subsampling b elements uniformly at random from $\{\mathbf{y}_j\}_{j=1}^m$ may lead to significant speed-ups. The additional term in the natural gradient descent algorithm incurs a $O(p^3)$ cost due to the need to invert a matrix, which for large-p settings may be prohibitive. In these cases, approximate linear solvers could also be used to reduce this cost.

2. The gradient of the generator $\nabla_\theta G_\theta$ may not be available, precluding exact gradient descent inference. In this case, the method of finite difference stochastic approximation [Kushner and Yin, 2003] can be used to approximate the descent direction. Finally, it is important to point out that the loss function may be non-convex and that we might converge to a local minimum.

Kernel Selection

Kernel selection for KSD and MMD estimators is delicate, since it will significantly influence the geometry of the statistical manifolds, and will hence have a significant impact on a range of issues including the efficiency and robustness of the estimators.

This was clearly demonstrated by the Gaussian distribution example used throughout the section. This example highlighted the impact of the choice of kernel hyperparameters on the robustness of the method. First, the KSD estimator was shown to be non-bias-robust, but the choice of lengthscale could help reduce the growth of the influence function. Later on, the MMD estimator was shown to be bias-robust, but the choice of lengthscale could impact the gross-error sensitivity. The kernel can therefore be seen as an additional free parameters to adapt these estimators to the problem at hand; something which is not possible with estimators based on the KL or SM divergence.

Since the choice of kernel will also impact the loss function itself, practical considerations should also prevail and it might be of interest to select a kernel which makes the loss function easy to minimise with gradient descent. The natural gradient algorithm will however be able to alleviate some of these issues by adapting directly to the geometry induced by the choice of kernel.

Previous work on the use of maximum mean discrepancy for hypothesis testing could also guide our choice of kernel. A common approach in that case was to study the asymptotic distribution of the tests, and choose kernel parameters so as to minimise the power of the test. Extensions where a linear combination of kernels whose weights are optimised was proposed in Gretton et al. [2012b], and Sutherland et al. [2017] used this approach for inference with generative adversarial networks. Note that this will require using held out data, which might therefore decrease the accuracy of our estimators.

Simulation Study

To highlight in detail all of the important theoretical and practical details discussed earlier in the section, our simulation study will focus solely on Gaussian models. Although this is of course not the intended class of models for KSD and MMD estimators, working with Gaussian models will be convenient because they are simple enough to be analysed in detail.

Gaussian Models as Unnormalised Models

We begin by discussing the estimators which make use of the log-likelihood of the model in a normalised or unnormalised form: KL, SM and KSD estimators. We will focus on a problem where our target is a multi-dimensional isotropic Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with mean $\mu = (\mu_1, \dots, \mu_d)$ and $d \times d$ scaled diagonal covariance matrix Σ with diagonal entries $(\sigma_1^2, \dots, \sigma_d^2)$. To begin with, we will assume that both the mean vector and the diagonal values of the covariance function are unknown such that $\theta = (\mu_1, \dots, \mu_d, \sigma_1, \dots, \sigma_d)$ and $\Theta = \mathbb{R}^d \times \mathbb{R}_+^d$.

The loss function for the one-dimensional case is plotted as a function of

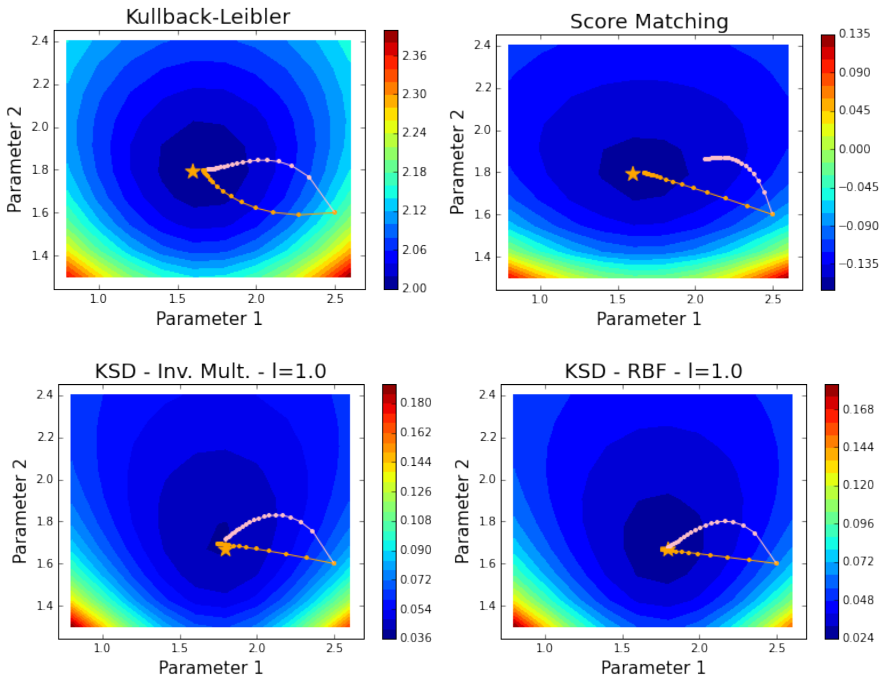

Figure 5.2: *Descent trajectories of gradient descent and natural gradient descent algorithms on a one-dimensional Gaussian model for estimators based on the KL, SM and KSD divergences.* We compare 20 iterations of gradient descent (pink) and natural gradient descent (orange) with constant step sizes. The loss function optimised are the KL, SM and KSD divergence with inverse-multiquadric kernel and Gaussian RBF kernel, each computed using $m = 100$ realisations. The minimum of the empirical loss functions are represented with an orange star. All of the optimisation algorithms were initialised at the $\theta_0 = (2.5, 1.6)$ and data was obtained from a model with $\theta^* = (1.5, 2)$.

the mean and standard deviation in Figure 5.2. Clearly, the loss function varies in geometry according to the choice of divergence, but is always convex since the Gaussian distribution is an example of the exponential family of distributions. For illustration, we still used numerical optimisation routines to tackle this problem (although a closed form solution was provided in Proposition 14). As seen in Figure 5.2, natural gradient algorithms are able to leverage knowledge of the geometry of the statistical manifold to provide more efficient updates towards the minimum. Indeed, the algorithm approximates a geodesic between the starting position and the minimum, resulting in a direct line when seen as a function of the parameters, which are the coordinates of the manifold. On the other hand, the gradient descent follows descent directions which are perpendicular to the contour lines of the heat

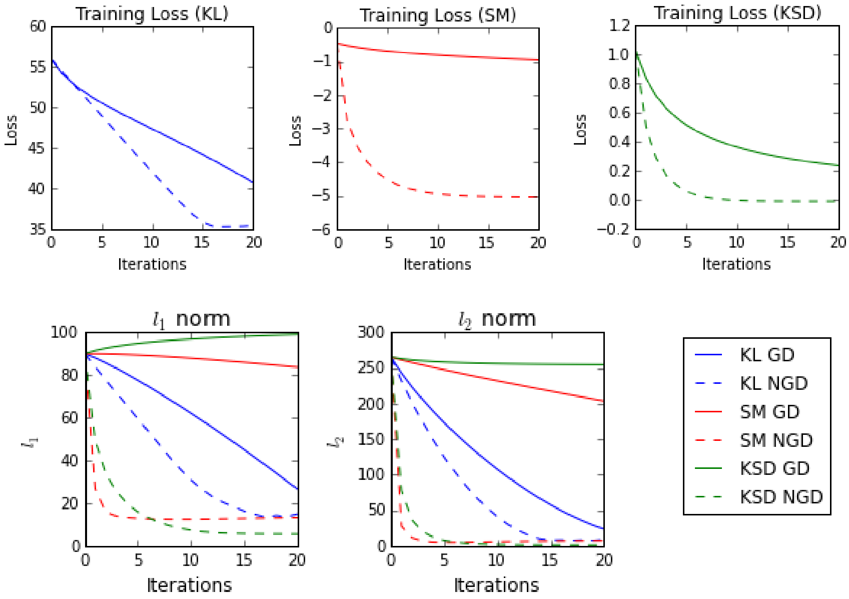

Figure 5.3: *Performance of natural gradient descent algorithms on a* 20*-dimensional Gaussian model for estimators based on the KL, SM and KSD divergences.* Consider an inference problem in an M-closed setting with $m = 300$ samples of a 20 dimensional Gaussian model with 40 parameters (mean vector and diagonal entries of the covariance function). The top plots compares the speed at which the gradient descent (GD) algorithms (full line) and the natural gradient descent (NGD) algorithms (dashed line) minimise each loss function. The KSD was computed with an inverse-multiquadric kernel with lengthscale parameter $l = 1$. The bottom plots compute l_1 and l_2 errors between the estimated parameter at iteration t and the true parameter θ^*.

map, which results in slower convergence towards the minimum.

We now extend this experiment to a Gaussian distribution on $\mathcal{X} = \mathbb{R}^{20}$, in which case the parameter space $\Theta \subset \mathbb{R}^{40}$. This is significantly more challenging since in a high-dimensional data space $\mathcal{X}$, empirical approximations of the divergences will not be as accurate. Furthermore, the optimisation problem also becomes more challenging in higher dimensions. Figure 5.3 compares the use of the KL, SM and KSD for estimation of this problem. Once again, the natural gradient algorithms are able to minimise the loss functions in fewer iterations. This time, the speed up obtained by the natural gradient algorithm is much more significant than in lower dimensions. This clearly highlights the advantage of making use of the geometry of the statistical inference problem.

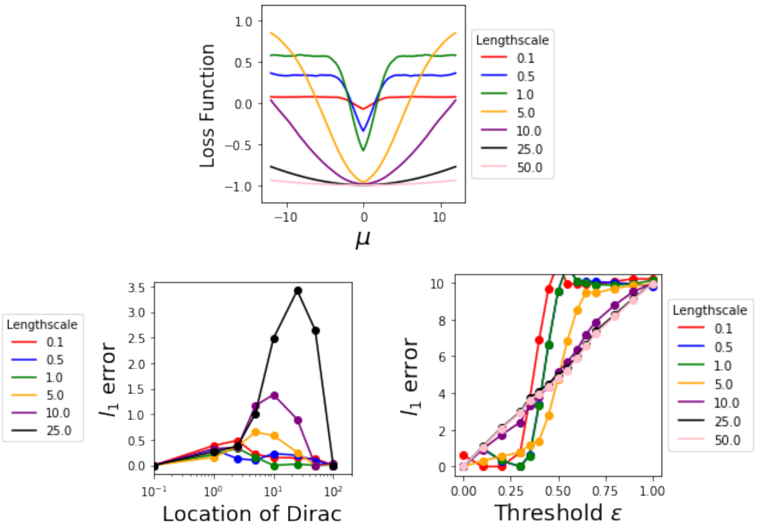

Figure 5.4: *Maximum mean discrepancy estimator based on a Gaussian RBF kernel for a Gaussian location model. Top:* Comparison of the loss landscape for various lengthscale values. *Bottom Left:* Robustness problem with varying location z for the Dirac but threshold fixed to $\epsilon = 0.2$. *Bottom Right:* Robustness problem with varying threshold but fixed location for the Dirac at $z = 10$.

Gaussian Models as Generative Models

We now consider inference for generative models using the MMD. We use a synthetic generative model $\mathbb{P}_\theta$ which is a multi-dimensional isotropic Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with mean vector μ and covariance σ^2 where $\sigma^2 > 0$ so that $\theta = (\mu, \sigma)$. In this case $\mathbb{U}$ is a standard Gaussian distribution $\mathcal{N}(0, 1)$ distribution on $\mathcal{U} = \mathbb{R}$ and $G_{\mu,\sigma}(u) = \mu + \sigma u$. We have that $\nabla_\mu G_{\mu,\sigma}(u) = 1$ and $\nabla_\sigma G_{\mu,\sigma}(u) = u$. For simplicity, and to understand the performance of the model for location and scale parameters separately, we first study the case where μ^* is unknown but σ^* known, which we call location model, then later move on to the case where μ^* is known but σ^* unknown, which we call scale model.

Starting with the location model, we first generate realisations from the Gaussian model with known scale parameter $\sigma = 1$ and unknown location parameter $\mu^* = \theta^* = 0$. The landscape of the loss function of an MMD estimator with Gaussian RBF kernel is presented in Figure 5.4 (top). We notice that the choice of lengthscale has a significant influence on this landscape. When the lengthscale is smaller than 5 or larger than 25, large parts of the loss function are flat. In those case, optimising this loss function will be challenging with gradient-based methods. We will hence need to repeatedly reinitialise the algorithm to be able to minimise the function.

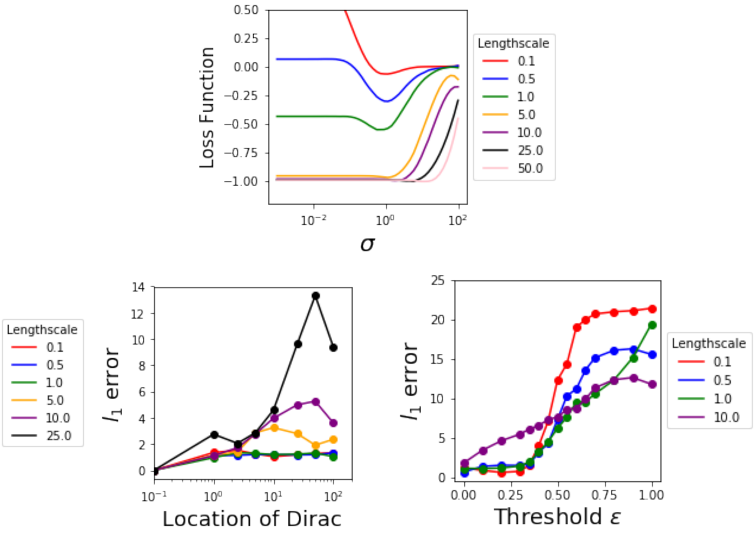

Figure 5.5: *Maximum mean discrepancy estimator based on a Gaussian RBF kernel for a Gaussian scale model. Top:* Comparison of the loss landscape for various lengthscale values. *Bottom Left:* Robustness problem with varying location z for the Dirac but threshold fixed to $\epsilon = 0.2$. *Bottom Right:* Robustness problem with varying threshold but fixed location for the Dirac at $z = 10$.

When the lengthscale is in the interval $[5, 25]$, the loss function will be amenable to gradient-based methods.

To highlight some of the issues surrounding robustness of kernel estimators, we also propose two additional experiments in an M-open setting. For these experiments, the data is generated from a measure $\mathbb{Q}$ corresponding to a mixture of the model $\mathbb{P}_*$ with unknown parameter $\mu^* = 0$, attributed a weight $1 - \epsilon$ for some $\epsilon \in [0, 1]$, and a Dirac measure at some point z, considered to be an outlier and attributed a weight of ϵ. Figure 5.4 (bottom left) shows the l_1 error between the true value θ^* and the MMD estimator $\theta^{\text{MMD}}_{n,m}$ for $n = m = 500$ as a function of the location $\mathbf{z}$ of the Dirac when $\epsilon = 0.2$. The MMD estimator has the desirable property that as $z \to \infty$, the estimator ignores the corruption from the outlier. This clearly illustrates one of the advantages of bias-robust estimators. In Figure 5.4 (bottom right), $z = 10$ but we vary ϵ. In this case, the corruption only affects the estimator in a significant manner when ϵ approaches a value around 0.5, in which case we realistically cannot consider $\mathbf{z}$ as an outlier anymore. In both of these plots, we see that the lengthscale has a significant impact on the robustness of the estimator.

Moving on to the scale model, we repeat the three experiments previously performed on the location model in Figure 5.5. The same conclusions can be obtained in this case: once again a value of the lengthscale which is either too large or too small will make the loss function impossible to minimise using gradient-based method, and the lengthscale has a significant impact on the robustness.

Summary

Kernel-based estimator can be useful for a variety of challenging statistical inference problems involving complex intractable models such as unnormalised or generative models. In this section, we framed the study of these estimators in the context of minimum distance estimators and strictly proper scoring rules, and discussed their bias-robustness.

Clearly, much more work remains from a theoretical viewpoint. First, ongoing work is focusing on proving consistency of these estimators, as well central limit theorems for the M-closed setting. These results do not follow directly from classical consistency and central limit theorem results from the scoring rule literature [Dawid and Musio, 2014] due to the necessity of using (multiple) U-statistic approximations of quantities of interest. It will also be interesting to study other types of robustness [Huber and Ronchetti, 2009].

From the point of view of applications, it will also be interesting to test some of the methodology described in this section on a wide range of models, including unnormalised graphical models from the imaging literature, or complex generative models from the ABC literature.

Chapter 6

Discussion

6.1 Contributions of the Thesis

Kernel methods have been used extensively across the computational sciences, including in statistics, machine learning, applied mathematics and engineering. The reason for their popularity lies in their ease of use, with the reproducing property providing a useful tool which renders many quantities of interest computable.

The goal of this thesis was to demonstrate that these advantages can also be useful in building algorithms in computational statistics. We highlighted how reproducing kernels can be used to tackle two of the most pressing problems in this area (introduced at length in Chapter 1): the numerical approximation of integrals of expensive and highly complex functions, and the construction of statistical estimators for inference within models where the likelihood cannot be evaluated.

To do so, the thesis began by reviewing known results on reproducing kernel Hilbert spaces, stochastic processes, and Bayesian nonparametrics (in Chapter 2). All of these notions were used throughout the following chapters, which contain the novel contributions of the thesis. The first part of the thesis began with the use of kernel methods in Bayesian probabilistic numerical methods, and highlighted their use in the well-known Bayesian quadrature (BQ) algorithm:

- In Chapter 3, we first showed how BQ can be formally analysed using the theory of reproducing kernel Hilbert spaces (RKHS). This allowed us to provide some theoretical guarantees on its asymptotic performance in the form of consistency and contraction rates. This contribution helped fill a major gap in the numerical analysis and probabilistic numerics literatures, which was preventing the large-scale use of BQ in statistical computation.

  We then provided an extensive simulation study which was devoted to the uncertainty quantification properties of these algorithms, and then applied BQ on a wide range of problems in computational statistics. This assessment should be helpful for readers interested in understanding the advantages, but also limitations, of the methodology.

The conclusions of this chapter are clear, and can be interpolated to most Bayesian probabilistic numerical methods. Providing exact Bayesian uncertainty quantification for the output of numerical methods is computationally expensive, and the associated model selection is a delicate task. It should therefore only be attempted in situations where the function underlying the numerical method is expensive and understanding the associated epistemic uncertainty is of importance for the application.

- With this last point in mind, Chapter 4 used insights from the theory of kernel methods to develop novel extensions to BQ. These aimed at pushing the performance capabilities of the algorithm to the fullest, in the sense of requiring a number of integrand evaluation n as small as possible.

  Section 4.1 began with a novel extension of BQ to vector-valued RKHS, which is helpful when multiple related integrals need to be computed simultaneously or sequentially. This extension allowed us to build estimators which re-use information to estimate multiple integrals. As such, these estimators are significantly less data intensive, but come with an increase in computational cost. Once again, theoretical work from the RKHS literature was essential in proving consistency and contraction results.

  We then proposed two new algorithms for efficient point selection.

  i. The first algorithm, called Frank-Wolfe Bayesian Quadrature (FWBQ) and presented in Section 4.2, used the fact that function-space conditional gradient algorithms can be made tractable in RKHS. In this setting, new points can be obtained analytically in terms of kernel evaluations. This property is convenient as it allowed us to build a practical algorithm based on experimental-design principles, with theoretical properties that could be formally analysed.

  ii. The second algorithm, called sequential Monte Carlo Bayesian quadrature (SMC-BQ) and presented in Section 4.3, attempts to approximate an optimal importance sampling distribution for BQ algorithms. Here, we make use of the fact that an upper bound on the integration error for functions in a RKHS can be straightforwardly approximated to create an efficient criterion for selection of the sampling distribution. Once again, this should provide useful methodology for applications of BQ.

The second part of the thesis (in Chapter 5) then discussed how to make use of kernel methods when models have a likelihood which cannot be evaluated in closed form, but only in some unnormalised or generative form.

- In the case of unnormalised models, we looked at the recent combination of reproducing kernels with Stein's method to construct a statistical divergence called kernel Stein discrepancy (KSD). We then highlighted two novel algorithms which extend BQ and FW to the design of numerical integration methodologies for integration against unnormalised densities. The extension is significant since these algorithms were previously restricted to cases where kernel means can be obtained in closed form and can now be applied to a wide range of problems. In particular, they can now be used in Bayesian inference where integrals often need to be computed against unnormalised posterior densities.
- We then studied some existing and novel minimum distance statistical estimators based on kernel-based discrepancies, such as the maximum mean discrepancy and the KSD. We discussed the flexibility of these methods, and highlighted how the choice of kernel can be used to adapt the geometry induced by the divergence to the need of the application at hand. We then provided novel numerical optimisation algorithms which exploit the geometry induced by these discrepancies to provide efficient implementation of our estimators.

6.2 Remaining Challenges

Future work related to each specific algorithm was already highlighted in the relevant chapters, but we point out common themes below.

- **Kernel selection.** A key gap in the literature on reproducing kernels is a satisfactory answer to the question of kernel choice. For Bayesian probabilistic numerical methods or kernel-based statistical estimators, we have highlighted that this choice will have a significant impact on performance, and proposed some heuristics for making this choice. However, further work will be required before we can make full use of the capabilities of these methods.

  For Bayesian probabilistic numerical methods, kernel selection is part of the problem of eliciting infinite-dimensional priors. Eliciting such information from domain experts is a challenging task which will require extensive further work. Some work has discussed cases in which maximum a-posteriori estimates of Bayesian algorithms correspond to existing methods in numerical analysis for certain choices of kernels (See Särkkä et al. [2016]; Karvonen and Särkkä [2017] for integration, or Owhadi [2015]; Cockayne et al. [2016] for differential equations). This work is a useful first step in this direction. However, further work should focus on the construction of kernels which include all of the information available to the user, including boundary conditions or knowledge of the smoothness satisfied by solutions of the differential equation.

For kernel-based statistical estimators, the choice of kernel relates to the question "which reproducing kernel can distinguish two probability measures in the most efficient manner?". The answer to this question obviously depends on these two measures, and on the form in which these are available. For two empirical measures and a fixed functional form of kernel, it is possible to choose kernel parameters based on the asymptotic distribution of a kernel two sample test [Gretton et al., 2012a]. It is, however, not clear that this choice will work well for small sample sizes. Another issue is how to choose the functional form itself. Gorham and Mackey [2017] highlighted how KSDs are highly sensitive to the choice of base kernel, but it is still unclear how to make this choice in general.

- **Approximate computation.** Kernels provide significant advantages over alternative methods due to the tractability provided by the kernel trick, but this usually comes with increased computational cost. Kernel-based interpolants usually have an $O(n^3)$ cost, and kernel-based discrepancies usually require $O(n^2)$ computations. Many approximation schemes exist for both cases, but it is still unclear whether these schemes can be combined with the algorithms in this thesis whilst simultaneously retaining the theoretical results.

  For Bayesian probabilistic numerical methods, the use of approximate kernel interpolants could help build algorithms which are (computationally) competitive with non-Bayesian algorithms. This will however require careful assessment of the impact of the approximations on the resulting posterior, and tight bounds on the distance between the exact and approximate posterior to assess whether the uncertainty quantification provided is still useful.

  For kernel-based estimators, many of the approximate methods from kernel hypothesis testing Gretton et al. [2009]; Jitkrittum et al. [2017] could be adapted to the statistical estimators. This will however require novel theoretical results guaranteeing consistency of these estimators, as well as novel numerical optimisation algorithms.

- **Kernels on non-Euclidean spaces.** Many of the algorithms proposed in this thesis could be adapted to more general domains. Indeed, the thesis usually focused on applications where the domain $\mathcal{X}$ was a Euclidean space or a sphere, but kernels existing for other spaces such as spaces of integers, graphs, time series and strings also exist [Schölkopf and Smola, 2002; Rasmussen and Williams, 2006]. Adapting the algorithms in this thesis to these spaces could provide significant performance enhancements for specific applications. For example, Oates et al. [2017a]; Ehler et al. [2019] demonstrated how BQ and its variant with Stein reproducing kernels can be formally constructed and analysed on manifolds. None of the other spaces mentioned have yet been considered in statistical computation, but have however been shown to be useful for applications such as natural language processing [Lodhi et al., 2017] or chemistry [Vert and Mahé, 2009]. We can therefore hope that they could be helpful to extend our algorithms.

Overall, this thesis has highlighted several areas where the theory of kernel methods (and associated fields) can provide insight and novel tools for statistical computation. The hope is that these contributions will help statistical methodology cope with the ever increasing computational needs of large-scale applications.

Bibliography

A. Abdulle and G. Garegnani. Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration. *arXiv:1801.01340*, 2017.

R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Academic Press Inc., 2003.

R. J. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. Institute of Mathematical Statistics, 1990.

Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 139–153, 2006.

M. A. Álvarez and N. D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12:1459–1500, 2011.

S.-I. Amari. *Differential Geometrical Methods in Statistics*. Springer-Verlag, 1987.

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

S.-I. Amari. *Information Geometry and Its Applications*. Springer, 2016.

A. Anastasiou. Bounds for the normal approximation of the maximum likelihood estimator from m-dependent random variables. *Statistics and Probability Letters*, 129(1):171–181, 2017.

A. Anastasiou and G. Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218, 2017.

A. Anastasiou and G. Reinert. Bounds for the asymptotic distribution of the likelihood ratio. *arXiv:1806.03666*, 2018.

S. Andradóttir, D. P. Heyman, and T. J. Ott. Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Transactions on Modeling and Computer Simulation*, 3(3):167–189, 1993.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18 (4):343–373, 2008.

E. Angelino, M. J. Johnson, and R. P. Adams. Patterns of scalable Bayesian inference. *Foundations and Trends in Machine Learning*, 9(2-3):119–247, 2016.

M. Arbel, D. J. Sutherland, M. Binkowski, and A. Gretton. On gradient regularizers for MMD GANs. *arXiv:1805.11565*, 2018.

A. V. Arkhangel'skii and L. S. Pontryagin. General Topology I: The Basic Concepts and Constructions of General Topology. Springer, 1991.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

R. Assaraf and M. Caffarel. Zero-variance principle for Monte Carlo algorithms. *Physical Review Letters*, 83(23):4682, 1999.

I. Babuska and J. M. Melenk. The partition of unity method. *International Journal for Numerical Methods in Engineering*, 40(4):727–758, 1997.

I. Babuska, U. Banerjee, and J. E. Osborn. Survey of meshless and generalized finite element methods: A unified approach. *Acta Numerica*, 12(2003), 2003.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *International Conference on Learning Theory*, pages 185–209, 2013.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(19):1–38, 2017.

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, pages 1355–1362, 2012.

O. Bachem, M. Lucic, and A. Krause. Practical coreset constructions for Machine Learning. *arXiv:1703.06476*, 2017.

N. S. Bakhvalov. On the optimality of linear methods for operator approximation in convex classes of functions. *USSR Computational Mathematics and Mathematical Physics*, 11 (4):244–249, 1971.

N. S. Bakhvalov. On the approximate calculation of multiple integrals. *Journal of Complexity*, 31(4):502–516, 2015.

A. D. Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, 25:175–184, 1988.

A. D. Barbour and L. H. Y. Chen. *An introduction to Stein's method.* World Scientific, 2005.

A. D. Barbour and L. H. Y. Chen. Stein's (magic) method. *arXiv:1411.1179*, pages 1–26, 2014.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18:1–43, 2017.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, 1978.

A. Barp, F.-X. Briol, A. D. Kennedy, and M. Girolami. Geometry and dynamics for Markov chain Monte Carlo. *Annual Reviews in Statistics and Its Applications*, 5, 2018.

S. Bartels and P. Hennig. Probabilistic approximate least-squares. *International Conference on Artificial Intelligence and Statistics*, pages 676–684, 2016.

F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76:1298–1302, 2006.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach.* CRC Press, 2011.

B. Bauer, L. Devroye, M. Kohler, A. Krzyżak, and H. Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017.

R. K. Beatson, W. A. Light, and S. Billings. Fast solution of the radial basis function interpolation equations: Domain decomposition methods. *SIAM Journal on Scientific Computing*, 22(5):1717–1740, 2001.

M. A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

J. Beck and S. Guillas. Sequential design with mutual information for computer experiments (MICE): emulation of a tsunami model. *SIAM Journal on Uncertainty Quantification*, 4: 739–766, 2016.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.* Springer-Verlag, 1984.

S. Bergman. Uber die Entwicklung der harmonischen Funktionen der Ebene und des Raumes nach Orthogonalfunktionen. *Mathematische Annalen*, 1922.

S. Bergman and M. Schiffer. *Kernel Functions and Elliptic Differential Equations in Mathematical Physics.* Academic Press Inc., New York, 1953.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science+Business Media, New York, 2004.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Inference in generative models using the Wasserstein distance. *arXiv:1701.05146*, 2017.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36(2), 1974.

A. Beskos, F. J. Pinski, J. M. Sanz-Serna, and A. M. Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201–2230, 2011.

A. Beskos, M. Girolami, S. Lan, P. E. Farrell, and A. M. Stuart. Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335:327–351, 2017.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*, 2017.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2012.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

A. Bondarenko, D. Radchenko, and M. Viazovska. Optimal asymptotic bounds for spherical designs. *Annals of Mathematics*, 178(2):443–452, 2013.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

J. Brauchart, E. Saff, I. H. Sloan, and R. S. Womersley. QMC designs: optimal order quasi Monte Carlo integration schemes on the sphere. *Mathematics of Computation*, 83: 2821–2851, 2014.

F.-X. Briol and M. Girolami. Bayesian numerical methods as a case study for statistical data science. In *Statistical Data Science*, pages 99–110. World Scientific, 2018.

F-X. Briol, C. J. Oates, M. Girolami, and M. A. Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances In Neural Information Processing Systems 28*, pages 1162–1170, 2015a.

F-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *To appear in "Statistical Science" with discussion and rejoinder, arXiv:1512.00933*, 2015b.

F-X. Briol, J. Cockayne, and O. Teymur. Comments on "Bayesian solution uncertainty quantification for differential equations" by Chkrebtii, Campbell, Calderhead & Girolami. *Bayesian Analysis*, 11(4):1285–1293, 2016.

F-X. Briol, C. J. Oates, J. Cockayne, W. Y. Chen, and M. Girolami. On the sampling problem for kernel quadrature. In *Proceedings of the International Conference on Machine Learning*, pages 586–595, 2017.

F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Rejoinder for "Probabilistic integration: a role in statistical computation?". *Statistical Science (to appear), arXiv:1811.10275*, 2018.

J. Brouillat, C. Bouville, B. Loos, C. Hansen, and K. Bouatouch. A Bayesian Monte Carlo approach to global illumination. *Computer Graphics Forum*, 28(8):2315–2329, 2009.

M. D. Buhmann. *Radial Basis Functions*. Cambridge University Press, 2003.

T. Bui-Thanh and M. Girolami. Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo. *Inverse Problems*, 30(11):114014, 2014.

S. Byrne and M. Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40:825–845, 2013.

A. Caimo and A. Mira. Efficient computational strategies for doubly intractable problems with applications to Bayesian social networks. *Statistics and Computing*, 25:113–125, 2015.

B. Calderhead. A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413, 2014.

B. Calderhead and M. Girolami. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis*, 53(12):4028–4045, 2009.

B. Calderhead and M. Girolami. Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Journal of Royal Society: Interface Focus*, 1: 821–835, 2011.

E. Cameron and A. N. Pettitt. Approximate Bayesian computation for astronomical model analysis: A case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 2012.

T. Campbell and T. Broderick. Automated scalable Bayesian inference via Hilbert coresets. *arXiv:1710.05053*, 2017.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanita. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010.

L. H. Y. Chen, L. Goldstein, and Q-M. Shao. *Normal Approximation by Stein's Method.* Springer, 2011.

W. Y. Chen, L. Mackey, J. Gorham, F-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the International Conference on Machine Learning, PMLR 80:843-852*, 2018.

W. Y. Chen, A. Barp, F.-X. Briol, L. Mackey, J. Gorham, M. Girolami, and C. J. Oates. Stein point Markov Chain Monte Carlo. 2019.

Y. Chen, M. Welling, and A. J. Smola. Super-samples from kernel herding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. Girolami. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267, 2016.

T. Choi and M. J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98:1969–1987, 2007.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016.

J. Cockayne, C. J. Oates, T. Sullivan, and M. Girolami. Probabilistic meshless methods for partial differential equations and Bayesian inverse problems. *arXiv:1605.07811*, 2016.

J. Cockayne, C. J. Oates, T. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *arXiv:1701.04006*, 2017.

J. Cockayne, C. J. Oates, and M. Girolami. A Bayesian conjugate gradient method. *arXiv:1801.05242*, 2018.

P. R. Conrad, M. Girolami, S. Särkkä, A. M. Stuart, and K. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Statistics and Computing*, 27(4):1065–1082, 2017.

S. Conti and A. O'Hagan. Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140:640–651, 2010.

S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

N. Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239–252, 1990.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

A. C. Damianou and N. D. Lawrence. Deep Gaussian processes. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, 31:207–215, 2013.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification*. 2016.

T. M. Davies and D. J. Bryant. On circulant embedding for Gaussian random fields in R. *Journal of Statistical Software*, 55(9), 2013.

P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Courier Corporation, 2007.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

A. P. Dawid and M. Musio. Theory and applications of proper scoring rules. *Metron*, 72 (2):169–183, 2014.

A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *Annals of Statistics*, 40(1):593–608, 2012.

A. P. Dawid, M. Musio, and L. Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016.

Y. A. de Montjoye, L. Radaelli, V. K. Singh, and A. S Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

E. De Vito, V. Umanità, and S. Villa. An extension of Mercer theorem to matrix-valued measurable kernels. *Applied and Computational Harmonic Analysis*, 34(3):339–351, 2013.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statisitical Methodology*, 68:411–436, 2006.

P. Dellaportas and I. Kontoyiannis. Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(1):133–161, 2012.

D. Dey, P. Muller, and D. Sinha. *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, Lecture Notes in Statistics, 1998.

P. Diaconis. Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, pages 163–175, 1988.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Annals of Statistics*, 14(1):1–26, 1986.

J. Dick and F. Pillichshammer. *Digital Nets and Sequences - Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.

J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(2): 193–227, 1984.

P. J. Diggle, P. Moraga, B. Rowlingson, and B. M. Taylor. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

J. L. Doob. Application of the theory of martingales. *Le Calcul des Probabilites et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique*, 13:23–27, 1949.

J. L. Doob. *Stochastic Processes.* Wiley, 1953.

R Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69, 2005.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of Nonlinear Filtering.* Oxford University Press, 2011.

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13: 3475–3506, 2012.

M. F. Driscoll. The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. 26(4):309–316, 1973.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2002.

M. M. Dunlop, M. Girolami, A. M. Stuart, and A. L. Teckentrup. How deep are deep Gaussian processes? *arXiv:1711.11280*, 2017.

D. K. Duvenaud. *Automatic model construction with Gaussian processes.* PhD thesis, University of Cambridge, 2014.

C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19, Berlin, 2008. Springer.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.

M. L. Eaton. A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics III*, pages 329–352, 1982.

J. L. Eftang and B. Stamm. Parameter multi-domain 'hp' empirical interpolation. *International Journal for Numerical Methods in Engineering*, 90:412–428, 2012.

M. Ehler, M. Gräf, and C. J. Oates. Optimal Monte Carlo integration on closed manifolds. *Statistics and Computing, to appear*, 2019.

W. Ehm and T. Gneiting. Local proper scoring rules of order two. *Annals of Statistics*, 40 (1):609–637, 2012.

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

G. Fasshauer, F. J. Hickernell, and H. Woźniakowski. On dimension-independent rates of convergence for function approximation with Gaussian kernels. *SIAM Journal on Numerical Analysis*, 50(1):247–271, 2012.

G. E. Fasshauer. Positive definite kernels: past, present and future. *Dolomite Research Notes on Approximation*, 4:21–63, 2011.

J. Feldman. Equivalence and perpendicularity of Gaussian processes. *Pacific Journal of Mathematics*, 8(4):699–708, 1958.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

M. Filippone and M. Girolami. Exact-approximate Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 222:309–368, 1922.

J. Fitzsimons, K. Cutajar, M. A. Osborne, S. Roberts, and M. Filiponne. Bayesian inference of log determinants. *Uncertainty in Artificial Intelligence*, 2017.

S. Flaxman, D. Sejdinovic, J. P. Cunningham, and S. Filippi. Bayesian learning of kernel embeddings. In *Uncertainty in Artificial Intelligence*, pages 182–191, 2016.

M. S. Floater and A. Iske. Multistep scattered data interpolation using compactly supported radial basis functions. *Journal of Computational and Applied Mathematics*, 73:65–78, 1996.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications.* Wiley, 1984.

P. G. M. Forbes and S. Lauritzen. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and Its Applications*, 473:261–283, 2015.

V. Fortuin and G. Ratsch. Deep mean functions for meta-learning in Gaussian processes. *arXiv:1901.08098*, 2019.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, 27(4):1119–1140, 1999.

N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(3):589–607, 2008.

N. Friel, M. Hurn, and J. Wyse. Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723, 2014.

E. Fuselier, T. Hangelbroek, F. J. Narcowich, J. D. Ward, and G. B. Wright. Kernel based quadratures on spheres and other homogeneous spaces. *Numerische Mathematik*, 127(1): 57–92, 2014.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *Proceedings of the International Conference on Machine Learning*, pages 541–549, 2015.

R. E. Gaunt, A. M. Pickett, and G. Reinert. Chi-square approximation by Stein's method with application to Pearson's statistic. *Annals of Applied Probability*, 27(2):720–756, 2017.

A. Gelman and X-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR 84*, pages 1608–1617, 2018.

M. Gerber and N. Chopin. Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(3):509–579, 2015.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, (1):156–163, 1991.

F. Ghaderinezhad and C. Ley. A general measure of the impact of priors in Bayesian statistics via Stein's Method. *arXiv:1803.00098*, 2018.

S. Ghosal and A. van Der Vaart. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

S. Ghosal, J. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

I. I. Gikhman and A. V. Skorokhod. *Introduction to the Theory of Random Processes*. Dover Publications Inc., 1969.

M. B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

W. R. Gilks, G. O. Roberts, and E. I. George. Adaptive direction sampling. *Journal of the Royal Statistical Society Series D: The Statistician*, 43(1):179–189, 1994.

E. Gine and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

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.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2004.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

G. H. Golub and J. H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1292–1301, 2017.

J. Gorham, A. Duncan, L. Mackey, and S. J. Vollmer. Measuring sample quality with diffusions. *arXiv:1506.03039*, 2016.

T. Graepel, K. Lauter, and M. Naehrig. ML confidential: Machine Learning on encrypted data. In *International Conference on Information Security and Cryptology*, pages 1–21, 2012.

R. B. Gramacy and D. W. Apley. Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578, 2015.

J. Grazzini, M. G. Richiardi, and M. Tsionas. Bayesian estimation of agent-based models. *Journal of Economic Dynamics & Control*, 77:26–47, 2017.

L. Greegard and J. Strain. The fast Gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

P. J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82:711–732, 1995.

P. J. Green, K. Latuszyski, M. Pereyra, and C. P. Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25: 835–862, 2015.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2006.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. *Neural Information Processing Systems*, pages 585–592, 2008.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, pages 673–681, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, and B. Schölkopf. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012a.

A. Gretton, B. K. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems 25*, pages 1214–1222, 2012b.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, pages 5228–5235, 2004.

G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, 2001.

S. Grunewalder. Compact convex projections. *Journal of Machine Learning Research*, 18 (219):1–43, 2018.

T. Gunter, R. Garnett, M. A. Osborne, P. Hennig, and S. Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2014.

M. U. Gutmann and A. Hyvärinen. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7 (2):223–242, 2001.

E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer Series in Computational Mathematics, second edition, 1996.

E. Hairer, S. P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer Series in Computational Mathematics, second edition, 1993.

Y. Hajizadeh, M. Christie, and V. Demyanov. Ant colony optimization for history matching and uncertainty quantification of reservoir models. *Journal of Petroleum Science and Engineering*, 77(1):78–92, 2011.

A. R. Hall. *Generalized Method of Moments*. Oxford University Press, 2005.

H. Hammer and H. Tjelmeland. Control variates for the Metropolis-Hastings algorithm. *Scandinavian Journal of Statistics*, 35(3):400–414, 2008.

F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

F. Hartig, J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. Statistical inference for stochastic simulation models – theory and application. *Ecology Letters*, 14:816–827, 2011.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

S. G. Henderson and P. W. Glynn. Approximating martingales for variance reduction in Markov process simulation. *Mathematics of Operations Research*, 27(2):253–271, 2002.

P. Hennig. Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25 (1):234–260, 2015.

P. Hennig and S. Hauberg. Probabilistic solutions to differential equations and their application to Riemannian statistics. In *Artificial Intelligence and Statistics*, pages 347–355, 2014.

P. Hennig and M. Kiefel. Quasi-Newton methods: a new direction. *Journal of Machine Learning Research*, 14:843–865, 2013.

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Journal of the Royal Society A*, 471(2179), 2015.

K. Hesse and I. H. Sloan. Worst-case errors in a Sobolev space setting for cubature over the sphere S2. *Bulletin of the Australian Mathematical Society*, 71(1):81–105, 2005.

F. J. Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322, 1998.

F. J. Hickernell, C. Lemieux, and A. B. Owen. Control variates for quasi-Monte Carlo. *Statistical Science*, 20(1):1–31, 2005.

D. M. Higdon. Space and space-time modeling using process convolutions. *Quantitative methods for current environmental issues*, pages 37–56, 2002.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

S. P. Huang, S. T. Quek, and K. K. Phoon. Convergence study of the truncated Karhunen-Loeve expansion for simulation of stochastic processes. *International Journal for Numerical Methods in Engineering*, 52(9):1029–1043, 2001.

P. J. Huber and E. M. Ronchetti. *Robust Statistics.* Wiley, 2009.

S. Hug, M. Schwarzfischer, J. Hasenauer, C. Marr, and F. J. Theis. An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule. *Statistics and Computing*, 26:663–677, 2016.

J. H. Huggins and L. Mackey. Random feature Stein discrepancies. 2018.

J. H. Huggins, T. Campbell, and T. Broderick. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

F. Huszár. *Scoring rules, Divergences and Information in Bayesian Machine Learning.* Phd thesis, University of Cambridge, 2013.

F. Huszár and D. K. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 377–385, 2012.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2006.

A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51(5):2499–2512, 2007.

A. Iske. *Multiresolution Methods in Scattered Data Modelling.* Springer, 2004.

M. Jaggi. Revisiting Frank-Wolfe: projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, volume 28, pages 427–435, 2013.

J. Jewson, J. Q. Smith, and C. Holmes. Principled Bayesian minimum divergence inference. 2018.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 261–270, 2017.

V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

V. R. Joseph, T. Dasgupta, R. Tuo, and C. F. J. Wu. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1):64–74, 2015.

V. R. Joseph, D. Wang, L. Gu, S. Lv, and R. Tuo. Deterministic sampling of expensive posteriors using minimum energy designs. *arXiv:1712.08929*, 2017.

J. B. Kadane and G. W. Wasilkowski. Average case epsilon-complexity in computer science: A Bayesian view. In *Bayesian Statistics 2, Proceedings of the Second Valencia International Meeting*, pages 361–374, 1985.

M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, pages 3288–3296, 2016.

M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *arXiv:1709.00147*, 2017.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

R. Karakida, M. Okada, and S.-I. Amari. Adaptive natural gradient learning algorithms for unnormalized statistical models. *Artificial Neural Networks and Machine Learning (ICANN)*, pages 427–434, 2016.

S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press Inc., 1975.

T. Karvonen and S. Särkkä. Classical quadrature rules via Gaussian processes. *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017.

T. Karvonen and S. Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):697–720, 2018.

T. Karvonen, C. J. Oates, and S. Särkkä. A Bayes-Sard cubature method. *arXiv:1804.03016*, 2018.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

R. E. Kass and P. W. Vos. *Geometric Foundations of Asymptotic Inference*. Wiley, 1997.

M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214, 2017.

D. J. Kaufman, J. Murphy-Bollinger, J. Scott, and K. L. Hudson. Public opinion about the importance of privacy in biobank research. *The American Journal of Human Genetics*, 85(5):643–654, 2009.

M. C. Kennedy. Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8(4):365–375, 1998.

M. C. Kennedy and A. O. Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(3):425–464, 2001.

J. T. Kent. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(1):71–80, 1982.

H. Kersting and P. Hennig. Active uncertainty calibration in Bayesian ODE solvers. In *Uncertainty in Artificial Intelligence*, pages 309–318, 2016.

D. P. Kingma and Y. LeCun. Regularized estimation of image statistics by score matching. In *Advances in Neural Information Processing Systems*, pages 1126–1134, 2010.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

J. Knoblauch, J. Jewson, and T. Damoulas. Doubly robust Bayesian inference for non-stationary streaming data with β-divergences. In *Advances in Neural Information Processing Systems*, 2018.

L. B. Koralov and Y. G. Sinai. *Theory of Probability and Random Processes.* Springer, 2007.

U. Koster and A. Hyvärinen. Natural image statistics: energy-based models estimated by score matching. In *2009 International Workshop on Local and Non-Local Approximation in Image Processing*, pages 16–25, 2009.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian Processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

E. Kreyszig. *Introductory Functional Analysis with Applications.* Wiley, 1989.

D. G. Krige. *A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand.* PhD thesis, University of Witwatersrand, 1951.

S. Kristoffersen. *The Empirical Interpolation Method.* PhD thesis, Norwegian University of Science and Technology, 2013.

F. Y. Kuo. Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. *Journal of Complexity*, 19(3):301–320, 2003.

F. Y. Kuo, W. T. M. Dunsmuir, I. H. Sloan, M. P. Wand, and R. S. Womersley. Quasi-Monte Carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability*, 10(2):239–275, 2008.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications.* Springer, 2003.

S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 544–552, 2015.

S. Lan, T. Bui-Thanh, M. Christie, and M. Girolami. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. *Journal of Computational Physics*, 308:81–101, 2016.

B. Larget and D. L. Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–759, 1999.

F. M. Larkin. Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2(3):379–422, 1972.

M. Lazaro-Gredilla, J. Quinonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11: 1865–1881, 2010.

Q. T. Le Gia, I. H. Sloan, and H. Wendland. Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. *Applied and Computational Harmonic Analysis*, 32:401–412, 2012.

A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.

C. Ley, G. Reinert, and Y. Swan. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *Annals of Applied Probability*, 27(1):216–241, 2017.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

F. Li and N. R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214, 2010.

Y. Li and R. E. Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *Proceedings of the International Conference on Machine Learning*, volume 37, pages 1718–1727, 2015.

H. C. Lie, A. M. Stuart, and T. J. Sullivan. Strong convergence rates of probabilistic integrators for ordinary differential equations. *arXiv:1703.03680*, 2017.

L. Lin, M. Drton, and A. Shojaie. Estimation of High-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854, 2016.

J. Lintusaari, M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1): 66–82, 2017.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2001.

Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, pages 3118–3126, 2017.

Q. Liu and J. D. Lee. Black-Box Importance Sampling. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 952–961, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2370–2378, 2016.

Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation. In *International Conference on Machine Learning*, 2016.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2017.

M. Loève. *Probability Theory*, volume 2. Springer, 1978.

M. Lukić and J. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467, 2015.

S. Lyu. Interpretation and generalization of score matching. In *Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2009.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

E. Magnani, H. Kersting, M. Schober, and P. Hennig. Bayesian filtering for ODEs with bounded derivatives. *arXiv:1709.08471*, 2017.

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

M. Mahsereci and P. Hennig. Probabilistic line searches for stochastic optimization. In *Advances In Neural Information Processing Systems*, pages 181–189, 2015.

V. Mameli and L. Ventura. Higher-order asymptotics for scoring rules. *Journal of Statistical Planning and Inference*, 165:13–26, 2015.

K. V. Mardia, J. T. Kent, and A. K. Laha. Score matching estimators for directional distributions. *arXiv:1604.08470*, page 2016.

J-M. Marin, K. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics*, pages 459–507. 2005.

J-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

R. Marques. *Bayesian and quasi-Monte Carlo spherical integration for global illumination.* PhD thesis, Université de Rennes 1, 2013.

R. Marques, C. Bouville, M. Ribardiere, P. Santos, and K. Bouatouch. A spherical Gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1619–1632, 2013.

J. Martens. New insights and perspectives on the natural gradient method. *arXiv:1412.1193*, 2014.

J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal of Scientific Computing*, 34(3):1460–1487, 2012.

H. Masnadi-Shirazi. Strictly proper kernel scoring rules and divergences with an application to kernel two-sample hypothesis testing. *arXiv:1704.02578*, 2017.

B. Matérn. *Spatial Variation.* Springer, 1960.

B. Mau, M. A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55(1):1–12, 1999.

J. Mercer. Functions of positive and negative type and their conection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209:415–446, 1909.

E. C. Merkle and M. Steyvers. Choosing a strictly proper scoring rule. *Decision Analysis*, 10(4):292–304, 2013.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6): 1087–1092, 1953.

P.-A. Meyer. Stochastic processes from 1950 to the present. *Electronic Journal for History of Probability and Statistics*, 5(1), 2009.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* 1993.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

A. Mira, R. Solgi, and D. Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.

J. Mockus. *Bayesian Approach to Global Optimization: Theory and Applications*. Kluwer Academic Publishers, 1989.

L. Mohamed, M. Christie, and V. Demyanov. Comparison of stochastic sampling algorithms for uncertainty quantification. *SPE Journal*, 15:31–38, 2010.

S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. *arXiv:1610.03483*, 2016.

J. Møller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman & Hall, 2004.

J. Møller, A. N. Pettitt, and R. Reeves. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

K. Monterrubio-Gómez, L. Roininen, S. Wade, T. Damoulas, and M. Girolami. Posterior inference for sparse hierarchical non-stationary models. *arXiv:1804.01431*, 2018.

M. T. Moores, A. N. Pettitt, and K. Mengersen. Scalable Bayesian inference for the inverse temperature of a hidden Potts model. *arXiv:1503.08066*, 2015.

S. Mosbach and A. G. Turner. A quantitative probabilistic investigation into the accumulation of rounding errors in numerical ODE solution. *Computers & Mathematics with Applications*, 57(7):1157–1167, 2009.

C. T. Mouat. *Fast algorithms and preconditioning techniques for fitting radial basis functions*. PhD thesis, University of Canterbury, 2001.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2016.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

P. Müller, F. A. Quintana, A. Jara, and T. Hanson. *Bayesian Nonparametric Data Analysis*. Springer, 2015.

S. A. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

I. Murray, Z. Ghahramani, and D. J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006.

F. J. Narcowich, J. D. Ward, and H. Wendland. Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24:175–186, 2006. ISSN 01764276.

J. C. Naylor and A. F. M. Smith. Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 31(3):214–225, 1982.

R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366, 1996.

R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. 2011.

N. J. Newton. Variance reduction for simulated diffusions. *SIAM Journal on Applied Mathematics*, 54(6):1780–1805, 1994.

H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* Society for Industrial and Applied Mathematics, 1992.

Y. Nishiyama and K. Fukumizu. Characteristic kernels and infinitely divisible distributions. *Journal of Machine Learning Research*, 17(180):1–28, 2016.

J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, 2006.

E. Novak. On the power of adaption. *Journal of Complexity*, 12:199–237, 1996.

E. Novak. Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 161–183, 2016.

E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems Volume I: Linear Information.* European Mathematical Society Publishing House, EMS Tracts in Mathematics 6, 2008.

E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems, Volume II : Standard Information for Functionals.* European Mathematical Society Publishing House, EMS Tracts in Mathematics 12, 2010.

J. Oakley. Eliciting Gaussian process priors for complex computer codes. *Journal of the Royal Statistical Society Series D: The Statistician*, 51(1):81–97, 2002.

C. J. Oates and M. Girolami. Control functionals for quasi-Monte Carlo integration. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 51, pages 56–65, 2016.

C. J. Oates and T. J. Sullivan. A modern retrospective on probabilistic numerics. *arXiv:1901.04457*, 2019.

C. J. Oates, F. Dondelinger, N. Bayani, J. Korkola, J. W. Gray, and S. Mukherjee. Causal network inference using biochemical kinetics. *Bioinformatics*, 30(17):i468–i474, 2014.

C. J. Oates, T. Papamarkou, and M. Girolami. The controlled thermodynamic integral for Bayesian model comparison. *Journal of the American Statistical Association*, 2016.

C. J. Oates, A. Barp, and M. Girolami. Posterior integration on an embedded Riemannian manifold. *arXiv:1712.01793*, 2017a.

C. J. Oates, J. Cockayne, and R. G. Aykroyd. Bayesian probabilistic numerical methods for industrial process monitoring. *arXiv:1707.06107*, 2017b.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society B: Statistical Methodology*, 2017c.

C. J. Oates, S. Niederer, A. Lee, F-X. Briol, and M. Girolami. Probabilistic models for integration error in the assessment of functional cardiac models. *Advances in Neural Information Processing*, 2017d.

C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on Stein's identity. *Bernoulli*, 2018.

J. Oettershagen. *Construction of optimal cubature algorithms with applications to econometrics and uncertainty quantification.* PhD thesis, Rheinischen Friedrich-Wilhelms-Universität Bonn, 2017.

A. O'Hagan. Monte Carlo is fundamentally unsound. *Journal of the Royal Statistical Society Series D: The Statistician*, 36(2):247–249, 1984.

A. O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29: 245–260, 1991.

A. O'Hagan. Some Bayesian numerical analysis. *Bayesian Statistics*, 4:345–363, 1992.

M. A. Osborne, D. K. Duvenaud, R. Garnett, C. E. Rasmussen, S. Roberts, and Z. Ghahramani. Active learning of model evidence using Bayesian quadrature. In *Advances In Neural Information Processing Systems*, pages 46–54, 2012.

A. B. Owen. Monte Carlo variance of scrambled net quadrature. *SIAM Journal on Numerical Analysis*, 34(5):1884–1910, 1997.

H. Owhadi. Bayesian numerical homogenization. *SIAM Multiscale Modeling & Simulation*, 13(3):818–828, 2015.

H. Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Review*, 59(1):99–149, 2017.

H. Owhadi and C. Scovel. Universal scalable robust solvers from computational information games and fast eigenspace adapted multiresolution analysis. *arXiv:1703.10761*, 2017.

H. Owhadi and L. Zhang. Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients. *Journal of Computational Physics*, 347:99–128, 2017.

H. Owhadi, C. Scovel, and T. Sullivan. On the brittleness of Bayesian inference. *SIAM Review*, 57(4):566–582, 2015.

J. Paisley, D. Blei, and M. I. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.

T. Papamarkou, A. Mira, and M. Girolami. Zero variance differential geometric Markov chain Monte Carlo algorithms. *Bayesian Analysis*, 9(1):97–128, 2014.

L. Pardo. *Statistical Inference Based on Divergence Measures*, volume 170. Chapman and Hall/CRC, 2005.

H. Park, C. Scheidt, D. Fenwick, A. Boucher, and J. Caers. History matching and uncertainty quantification of facies models with multiple geological interpretations. *Computational Geosciences*, 17(4):609–621, 2013.

M. Parry. Extensive scoring rules. *Electronic Journal of Statistics*, 10(1):1098–1108, 2016.

M. Parry, A. P. Dawid, and S. Lauritzen. Proper local scoring rules. *Annals of Statistics*, 40(1):561–592, 2012.

L. Parussini, D. Venturi, P. Perdikaris, and G. E. Karniadakis. Multi-fidelity Gaussian process regression for prediction of random fields. *Journal of Computational Physics*, 336:36–50, 2017.

R. Pascanu and Y. Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representation*, 2014.

G. A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations.* Springer, 2014.

G. Pedrick. *Theory of reproducing kernels for Hilbert spaces of vector valued functions.* PhD thesis, University of Kansas, 1957.

B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *ACDL Technical Report TR16-1*, 2016a.

B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM Journal of Scientific Computing*, 38(5), 2016b.

P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. Nonlinear information fusion algorithms for robust multi-fidelity modeling. *Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, 473(2198), 2016.

N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems part II: stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal of Scientific Computing*, 36(4): 1525–1555, 2014.

M. Pharr and G. Humphreys. *Physically based rendering: From theory to implementation.* Morgan Kaufmann Publishers Inc., 2004.

N. S. Pillai, Q. Wu, F. Liang, S. Mukherjee, and R. L. Wolpert. Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, 8:1769–1797, 2007.

T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

H. Poincaré. *Calcul des probabilites.* Gauthier-Villars, Paris, 1896.

J. Quinonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances In Neural Information Processing Systems*, pages 1177–1184, 2007.

M. Raissi and G. E. Karniadakis. Deep multi-fidelity Gaussian processes. *arXiv:1604.07484*, 2016.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Inferring solutions of differential equations using noisy multi-fidelity data. *Journal of Computational Physics*, 335:736–746, 2017.

G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 489–496, 2002.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

F. Riesz and B. S. Nagy. *Functional Analysis.* Dover Publications, 1990.

K. Ritter. *Average-Case Analysis of Numerical Problems.* Springer, 2000.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* Springer-Verlag, 1994.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

C. P. Robert and G. Casella. A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115, 2011.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60 (1):255–268, 1998.

J. C. Robinson. *An Introduction to Ordinary Differential Equations*. Cambridge University Press, 2004.

N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

P. J. Rossky, J. D. Doll, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *Journal of Chemical Physics*, 69(10):4628, 1978.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximatins. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian computing with INLA: A review. *Annual Reviews of Statistics and Its Applications*, 4: 395–421, 2016.

M. Sahani, G. Bohner, and A. Meyer. Score-matching estimators for continuous-time point process regression models. In *2016 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–5, 2016.

S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

S. Särkkä, J. Hartikainen, L. Svensson, and F. Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. *Journal of Advances in Information Fusion*, 11(1):31–46, 2016.

R. Scalettar, D. J. Scalapino, and R. L. Sugar. New algorithm for the numerical simulation of fermions. *Physical Review B*, 34(11):7911–7917, 1986.

R. Schaback and H. Wendland. Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15, 2006.

M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge-Kutta means. In *Advances in Neural Information Processing Systems*, pages 739–747, 2014.

M. Schober, S. Särkkä, and P. Hennig. A probabilistic model for the numerical solution of initial value problems. *Statistics and Computing, to appear*, 2018.

I. J. Schoenberg. On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space. *Annals of Mathematics*, 38(4):787–793, 1937.

B. Schölkopf and A. J. Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

L. Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associes (noyaux reproduisants). *Journal d'Analyse Mathematique de Jerusalem*, 13:115–256, 1964.

L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie*, 4:10–26, 1965.

D. Sejdinovic, H. Strathmann, M. Lomeli Garcia, C. Andrieu, and A. Gretton. Kernel adaptive Metropolis-Hastings. In *Proceedings of the International Conference on Machine Learning*, pages 1665–1673, 2014.

S. Sharma. Markov chain Monte Carlo methods for Bayesian data analysis in astronomy. *Annual Review of Astronomy and Astrophysics*, 55:213–259, 2017.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *Annals of Statistics*, 37(6):3960–3984, 2009.

W. Sickel and T. Ullrich. Tensor products of Sobolev-Besov spaces and applications to approximation from the hyperbolic cross. *Journal of Approximation Theory*, 161(2):748–786, 2009.

V. Sinescu, F. Y. Kuo, and I. H. Sloan. On the choice of weights in a function space for quasi-Monte Carlo methods for a class of generalised response models in statistics. In *Monte Carlo and Quasi-Monte Carlo Methods 2012*. 2012.

J. Skilling. Bayesian solution of ordinary differential equations. In *Maximum Entropy and Bayesian Methods*, volume 50, pages 23–37, 1991.

I. H. Sloan and H. Woźniakowski. When are Quasi-Monte Carlo algorithms efficient for high dimensional integrals? *Journal of Complexity*, 14(1):1–33, 1998.

A. F. M. Smith, A. M. Skene, J. E. H. Shaw, J. C. Naylor, and M. Dransfield. The implementation of the Bayesian paradigm. *Communications in Statistics - Theory and Methods*, 14(5):1079–1102, 1985.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances In Neural Information Processing Systems*, pages 2951–2959, 2012.

A. Sommariva and M. Vianello. Numerical cubature on scattered data by radial basis functions. *Computing*, 76(3-4):295–310, 2006.

L. Song. *Learning via Hilbert space embedding of distributions.* PhD thesis, The University of Sydney, 2008.

B. K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:30, 2010a.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11, 2010b.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and . R G Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, and A. Hyvärinen. Density Estimation in Infinite Dimensional Exponential Families. *Journal of Machine Learning Research*, 18, 2017.

A. Srivastava and E. Klassen. Bayesian and geometric subspace tracking. *Advances in Applied Probability*, 56:43–56, 2004.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602. University of California Press, 1972.

M. Stein. *Interpolation of Spatial Data - Some Theory for Kriging.* Springer Science+Business Media, 1999.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. 2017.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52 (10):4635–4643, 2006.

J. Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain Journal Mathematics*, 6(3):409–434, 1976.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabó, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems*, pages 955–963, 2015.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

M. A. Suchard, Q. Wang, C. Chan, J. Frelinger, A. Cron, and M. West. Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19(2):419–438, 2010.

A. B. Suldin. Wiener measure and its applications to approximation methods. I. *I. Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika*, (6):145–158, 1959.

T. J. Sullivan. *Introduction to Uncertainty Quantification.* Springer, 2016.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. J. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representation*, 2017.

R. H. Swendsen and J-S. Wang. Replica Monte Carlo simulation of spin glasses. *Physical Review Letters*, 57(21):2607–2609, 1986.

K. Swersky, M. A. Ranzato, D. Buchman, B. M. Marlin, and N. de Freitas. On autoencoders and score matching for energy based models. In *International Conference on Machine Learning*, pages 1201–1208, 2011.

B. Szabó, A. van der Vaart, and J. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

Z. Szabó, B. K. Sriperumbudur, B. Poczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

O. Teymur, K. Zygalakis, and B. Calderhead. Probabilistic linear multistep methods. In *Neural Information Processing Systems*, pages 4314–4321, 2016.

O. Teymur, B. Calderhead, H. Cheng Lie, and T. Sullivan. Implicit Probabilistic Integrators for ODEs. *Advances in Neural Information Processing Systems, to appear*, 2018.

J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski. *Information, Uncertainty, Complexity.* Addison-Wesley, 1983.

L. N. Trefethen. The definition of numerical analysis. *SIAM News*, 25(6), 1992.

L. N. Trefethen and D. Bau. *Numerical Linear Algebra.* Society for Industrial and Applied Mathematics, 1997.

P. Turaga, A. Veeraraghavan, and R. Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

J. M. Ver Hoef and R. P. Barry. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69(907):275–294, 1998.

J.-P. Vert and P. Mahé. Graph kernels based on tree patterns for molecules. *Machine Learning*, pages 3–35, 2009.

P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 1674:1661–1674, 2011.

G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, 1991.

Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. de Freitas. Bayesian optimization in high dimensions via random embeddings. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.*, pages 1778–1784, 2013.

H. Wendland. *Scattered data approximation.* Cambridge University Press, 2005.

L. Wenliang, D. J. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. *arXiv:1811.08357*, 2018.

D. Williams. *Probability with martingales.* Cambridge University Press, 1991.

A. G. Wills and T. B. Schön. On the construction of probabilistic Newton-type algorithms. *arXiv:1704.01382*, 2017.

S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

Z. Wu and R. Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis*, 13(1):13–27, 1993.

X. Xi, F-X. Briol, and M. Girolami. Bayesian quadrature for multiple related integrals. *International Conference on Machine Learning, PMLR 80:5369-5378*, 2018.

W. Xu and M. L. Stein. Maximum likelihood estimation for a smooth Gaussian random field model. *SIAM Journal on Uncertainty Quantification*, 5(1):138–175, 2017.

Y. Yang and D. B. Dunson. Bayesian manifold regression. *Annals of Statistics*, 44(2): 876–905, 2016.

G. A. Young and R. L. Smith. *Essentials of Statistical Inference.* Cambridge University Press, 2005.

S. Yu, M. Drton, and A. Shojaie. Graphical models for non-negative data using generalized score matching. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR 84*, pages 1781–1790, 2018.

Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH '99)*, pages 215–224, 1999.

E. Zawadzki and S. Lahaie. Nonparametric scoring rules. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI 2015)*, pages 3635–3641, 2015.

Y. Zhou, A. M. Johansen, and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.

Z. Zhu, R. Wan, and M. Zhong. Neural control variates for variance reduction. *arXiv:1806.00159*, 2018.

Appendix A

Background Material

This thesis requires some basic understanding of functional analysis and probability theory. For completeness, we now provide a brief introduction to the important concepts from these fields that are used in the main text.

A.1 Topology and Functional Analysis

Denote by $\mathcal{X}$ some abstract set. In this section, we will start by discussing some useful examples of such sets for computational statistics, which will be required to formalise the methodology throughout this thesis. Most of these notions will be used to formalise certain properties of sets which readers will find intuitive from the Euclidean space setting. Specifically, we will discuss metric space, vector spaces and inner product spaces of functions and measures. Most of the material in this section is based on Kreyszig [1989]; Folland [1984]. The basic space which we will work with is a topological space. Denote by $\emptyset$ the empty set, $A \cup B$ the union of the sets A and B, and $A \cap B$ the intersection of the sets A and B.

Definition 3 (**Topological Space**, Kreyszig [1989] p19)**.** *We call topological space any pair $(\mathcal{X}, C)$ consisting of a space $\mathcal{X}$ and collection of open subsets of $\mathcal{X}$ denoted C such that: (i) $\emptyset \in C$ and $\mathcal{X} \in C$, (ii) any arbitrary union (countable or uncountable) of elements of C is in C, and (iii) the intersection of finitely many elements of C is in C.*

An class of topological spaces often used in probability theory are the Hausdorff spaces. A space $\mathcal{X}$ is a Hausdorff space (See p27 in Arkhangel'skii and Pontryagin [1991]) if any two distinct points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ can be separated by disjoint neighbourhoods; i.e., there exist open subsets $\mathcal{Y}$ and $\mathcal{Z}$ of $\mathcal{X}$ such that $\mathbf{x} \in \mathcal{Y}, \mathbf{y} \in \mathcal{Z}$ and $\mathcal{Y} \cap \mathcal{Z} = \emptyset$.

A useful property of topological spaces is compactness. A topological space $\mathcal{X}$ is compact if for every collection C of open subsets of $\mathcal{X}$ such that $\mathcal{X} = \bigcup_{\mathbf{x} \in C} \mathbf{x}$, there is a finite subset F of C such that $\mathcal{X} = \bigcup_{\mathbf{x} \in F} \mathbf{x}$.

The simplest example of topological space on which a notion of distance can be defined is called a metric space:

Definition 4 (**Metric space**. Kreyszig [1989], Definition 1.1-1)**.** *A metric space is a pair $(\mathcal{X}, d)$, where $\mathcal{X}$ is a set and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a metric on $\mathcal{X}$; i.e. $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ we have: (i) $d(\mathbf{x}, \mathbf{y})$ is real-valued, finite and non-negative, (ii) $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$, (iii) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry), and (iv) $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ (triangle inequality).*

Another useful property is that a subspace $\mathcal{Y}$ of a metric space $\mathcal{X}$ is a dense subspace if and only if for every point in $\mathcal{X}$ exists as a limit of a sequence in the subspace $\mathcal{Y}$.

All metric spaces are Hausdorff spaces. A simple example of metric space which will be familiar to most readers is the Euclidean space $\mathcal{X} = \mathbb{R}^d$ ($d \in \mathbb{N}$) combined with the 2-norm metric $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$ for all $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = (y_1, \ldots, y_d)$ in $\mathcal{X}$.

It is also possible to consider metric spaces of functions. For example, the space $\mathcal{X} = C[a, b]$ of all real-valued continuous functions on some interval $[a, b] \subset \mathbb{R}$, together with the metric $d_1(f, g) = \max_{x \in [a,b]} |f(x) - g(x)|$ for $f, g \in \mathcal{X}$ forms a metric space. Similarly, so does the space $\mathcal{X} = L^2[a, b]$ of square-integrable functions with the metric $d_2(f, g) = \sqrt{\int_a^b (f(x) - g(x))^2 \, \mathrm{d}x}$. In this case the integral is defined as a Lebesgue integral and each element of this space are equivalent classes of functions (a technicality required due to measure-zero sets). It is also possible to generalise this space to $L^2(\mathcal{X}; \Pi)$ where Π is a probability measure, in which case we have $d_3(f, g) = \sqrt{\int_{\mathcal{X}} (f(x) - g(x))^2 \, \Pi(\mathrm{d}x)}$.

In the theory of metric spaces, an important type of sequence $\{\mathbf{x}_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$ are Cauchy sequences. A sequence is said to be Cauchy if for every $\epsilon > 0, \exists N(\epsilon)$ such that $d(\mathbf{x}_m, \mathbf{x}_n) < \epsilon$ for every $m, n > N$. The metric space $\mathcal{X}$ is then said to be complete if every Cauchy sequence in $\mathcal{X}$ converges (i.e $\exists \mathbf{x} \in \mathcal{X}$ called limit such that $\lim_{n \to \infty} d(\mathbf{x}_n, \mathbf{x}) = 0$).

Let us now consider functions mapping from some metric space $\mathcal{X}$ to itself. We say a sequence $\{f_n\}_{n \in \mathbb{N}}$ converges pointwise to the function f if and only if $\lim_{n \to \infty} f_n(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Furthermore, we say the sequence $\{f_n\}_{n \in \mathbb{N}}$ converges uniformly to f if and only if $\sup_{\mathbf{x} \in \mathcal{X}} |f_n(\mathbf{x}) - f(\mathbf{x})| \to 0$ as $n \to \infty$.

All of the examples above are examples of complete metric spaces, but not all metric spaces are complete. For example, $\mathbb{R} - \{a\}$ (for some $a \in \mathbb{R}$) equipped with the 2-norm metric is not a complete space. Completeness of the space means that the space is "well behaved" in many aspect, most notably in that we can establish notions of continuity of mappings from a complete metric space to another. We now move on to another important type of space:

Definition 5 (**Real vector space**. Kreyszig [1989], Definition 2.1-1)**.** *A real vector space is a non-empty set $\mathcal{X}$ of elements, called vectors, together with two algebraic operations called vector addition and multiplication of vectors by scalars.*

An important class of metric spaces are obtained by taking a vector space and inducing a metric on it using a norm; these are called Banach spaces:

Definition 6 (**Banach space**. Kreyszig [1989], Definition 2.2-1)**.** *A norm is a function $\| \cdot \| : \mathcal{X} \to \mathbb{R}$ with the following properties $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \alpha \in \mathbb{R}$: (i) $\|\mathbf{x}\| \geq 0$, (ii)$\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0$, (iii)$\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, and (iv) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.*

A Banach space $\mathcal{X}$ is a vector space with a norm $\| \cdot \|$ defined on it, such that the space is complete in the metric $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ induced by this norm.

We remark that the Euclidean space with 2-norm, $C[a, b]$ with metric d_1 and $L^2[a, b]$ with metric d_2 are all examples of Banach spaces. Although this is in no way a requirement, the thesis will mostly focus on Banach spaces whose elements are functions (or equivalence classes of functions) from some abstract domain $\mathcal{X}$ to $\mathbb{R}$.

On a Banach space, we say a sequence $\{\mathbf{x}_n\}_{n \in \mathbb{N}}$ is absolutely convergent if and only if $\sum_{n \in \mathbb{N}} \|\mathbf{x}_n\| < \infty$.

The norm of a Banach space generalises the elementary concept of the length of a vector. However, we are still missing a notion of angles, which is provided in the Euclidean context by a dot product, a special case of inner product:

Definition 7 (**Hilbert space**. Kreyszig [1989], Definition 3.1-1)**.** *We call a space $\mathcal{X}$ an inner product space if it has a function $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, called inner product, which satisfies: (i) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$, (ii) $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$, (iii) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$, and (iv) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = 0$.*

A Hilbert space is an inner product space such that the space is complete in the metric $d(\mathbf{x}, \mathbf{y}) = \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle}$ induced by the inner product.

Clearly, we hence have that Hilbert spaces are always Banach spaces with norm/metric induced by an inner product. Note that the converse is not always true as it is possible to have Banach spaces with norms defined without an inner product structure. The Euclidean space with 2-norm and $L^2[a,b]$ space with metric d_2 are both examples of Hilbert space, but $C[a,b]$ with metric d_1 is not equipped with a norm which can be written as an inner product, and therefore is not a Hilbert space.

An important example of Hilbert space are Sobolev space [Adams and Fournier, 2003]. Suppose $\mathcal{X} \subseteq \mathbb{R}^d$. When $\alpha \in \mathbb{N}$, these spaces are defined as:

$$W_2^\alpha(\mathcal{X}) \quad := \quad \left\{ f \in L^2(\mathcal{X}) : D^\nu f \in L^2(\mathcal{X}) \text{ exists } \forall \nu \in \mathbb{N}_0^d \text{ with } |\nu| \leq \alpha \right\},$$

with inner product $\langle f,g\rangle_{W_2^\alpha(\mathcal{X})} := \sum_{|\nu|\leq\alpha} \langle D^\nu f, D^\nu g\rangle_{L^2(\mathcal{X})}$ for all $f,g \in W_2^\alpha(\mathcal{X})$ where $D^\nu f = \partial^{|\nu|} f / \partial x_1^{\nu_1} \dots \partial x_d^{\nu_d}$ denotes the total derivative corresponding to the multi-index $\nu = (\nu_1, \dots, \nu_d) \in \mathbb{N}_0^p$. This means that all of the functions in this space will have smoothness α.

It is also possible to have fractional Sobolev spaces; i.e. the smoothness $\alpha > 0$ can take any positive real value. For $\mathcal{X} = \mathbb{R}^d$ and denoting by $\hat{f}(\xi) = \int_{\mathcal{X}^d} f(\mathbf{x}) \exp(-2\pi i \langle \mathbf{x}, \xi\rangle) \mathrm{d}\mathbf{x}$ the Fourier transform of f, these spaces are given by:

$$H^\alpha(\mathbb{R}^d) \quad := \quad \left\{ f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \left|\hat{f}(\xi)\right|^2 (1 + \|\xi\|^2)^\alpha \mathrm{d}\xi < \infty \right\},$$

with associated inner product $\langle f,g\rangle_{H^\alpha(\mathbb{R}^d)} := \int \hat{f}(\xi)\overline{\hat{g}(\xi)}(1+\|\xi\|^2)^\alpha \mathrm{d}\xi$ for all $f,g \in H^\alpha(\mathbb{R}^d)$ where $\overline{g}$ denoted the complex conjugate of $\hat{g}$.

A final interesting example are the Sobolev spaces of dominating mixed smoothness which are defined as:

$$\mathcal{S}_2^\alpha(\mathcal{X}) := \left\{ f \in L^2(\mathcal{X}) : \sum_{\forall j: \nu_j \leq \alpha} D^\nu f \in L^2(\mathcal{X}) \right\}$$

with inner product given by $\langle f,g\rangle_{\mathcal{S}_2^\alpha} := \sum_{\forall j:\nu_j \leq \alpha} \langle D^\nu f, D^\nu g\rangle_{L^2(\mathcal{X})}$. Clearly $\mathcal{S}_2^\alpha(\mathcal{X})$ requires α derivatives in each coordinate, a stronger assumption than for $W_2^\alpha(\mathcal{X})$ which only requires the sum of coordinate derivatives to be α.

Many functional approximation results in Sobolev spaces require more regularity from the domain. One type of domains which is commonly used is domains with Lipschitz boundaries, which we introduce below. To do so, we begin with special Lipschitz domains. For $d > 2$, we say that an open set $\mathcal{X} \subset \mathbb{R}^d$ is a special Lipschitz domain if there exists a rotation of $\mathcal{X}$, denoted by $\tilde{\mathcal{X}}$, and a function $\phi : \mathbb{R}^{d-1} \to \mathbb{R}$ that satisfy the following: (i) $\tilde{\mathcal{X}} = \{\mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \mathbf{y} > \phi(\mathbf{x})\}$, (ii) ϕ is a Lipschitz function such that $|\phi(\mathbf{x}) - \phi(\mathbf{x}')| \leq M\|\mathbf{x} - \mathbf{x}'\| \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d-1}$, where $M > 0$ is called the Lipschitz bound of $\mathcal{X}$.

With this definition now complete, we can define the notion of a domain with

Lipschitz boundary. Let $\mathcal{X} \subset \mathbb{R}^d$ be an open set and $\partial\mathcal{X}$ be its boundary. We say the boundary is Lipschitz $\exists \epsilon, M > 0, K \in \mathbb{N}$ and open sets $U_1, \ldots, U_L \subset \mathbb{R}^d$ where $L \in \mathbb{N} \cup \{\infty\}$ such that the following holds: (i) $\forall \mathbf{x} \in \partial\mathcal{X}$, $\exists i$ such that $B(\mathbf{x}, \epsilon)$, the ball centred at $\mathbf{x}$ of radius ϵ, satisfies $B(\mathbf{x}, \epsilon) \subset U_i$, (ii) $U_{i_1} \cap \ldots \cap U_{i_{K+1}} = \emptyset$ for any distinct indices $\{i_1, \ldots, i_{K+1}\}$, and (iii) for each index i, there exists a special Lipschitz domain $\mathcal{X}_i \subset \mathbb{R}^d$ with Lipschitz bound b such that $U_i \cap \mathcal{X} = U_i \cap \mathcal{X}_i$ and $b \leq M$.

Going back to Hilbert spaces, an important property which we will make extensive use of is the Cauchy-Schwarz inequality, which states that:

Lemma 4 (**Cauchy-Schwarz inequality**. [Kreyszig, 1989], Lemma 3.2-1)**.** *For all* $\mathbf{x}, \mathbf{x}'$ *in some inner-product space* $\mathcal{X}$*, the following holds:* $|\langle \mathbf{x}, \mathbf{x}' \rangle| \leq \|\mathbf{x}\| \|\mathbf{x}'\|$.

We now conclude this section with important definitions and properties of operators. Let $\mathcal{X}, \mathcal{Y}$ be vector spaces. We say an operator $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ is a linear operator if and only if $\mathcal{A}[\alpha\mathbf{x} + \beta\mathbf{x}'] = \alpha\mathcal{A}[\mathbf{x}] + \beta\mathcal{A}[\mathbf{x}']$ for all $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. We also say that the linear operator $\mathcal{A}$ is bounded if and only if $\exists C > 0$ such that $\|\mathcal{A}[\mathbf{x}]\|_{\mathcal{Y}} \leq C\|\mathbf{x}\|_{\mathcal{X}}$ $\forall \mathbf{x} \in \mathcal{X}$. Finally, we call eigenvector (or eigenfunction in the case where $\mathcal{X}$ is a function space) any non-zero $\mathbf{v} \in \mathcal{X}$ that only changes by a constant factor when applying the operator. That is, we call eigenvector any $\mathbf{v}$ such that $\mathcal{A}[\mathbf{v}] = \lambda\mathbf{v}$, and $\lambda \in \mathbb{R}$ is then called the eigenvalue corresponding to $\mathbf{v}$. We call linear functional any linear operator $\mathcal{A} : \mathcal{X} \to \mathbb{R}$. We say that a linear functional is continuous if and only if it is bounded. The set of all bounded linear functionals on some normed space $\mathcal{X}$ constitutes a normed space itself, called the dual space and denoted $\mathcal{X}^*$. It has norm defined as: $\|\mathcal{A}\|_{\mathcal{X}^*} = \sup_{\mathbf{x} \in \mathcal{X} : \|\mathbf{x}\| = 1} |\mathcal{A}(\mathbf{x})|$ and is itself a Banach space.

A.2 Measure and Probability Theory

We have now completed our basic introduction to functional analysis. In this section, we recall definitions and theorems in measure and probability theory. The reader is referred to Williams [1991]; Grimmett and Stirzaker [2001]; Dudley [2002]; Koralov and Sinai [2007] for an in-depth introduction. Note that some of this section relies on the section above, and so the reader should read these two sections in the order they appear if unfamiliar with functional analysis.

In probability theory, we are interested in formalising the notion of random events on some abstract set $\mathcal{X}$. This is done by considering a basic collection of events $\mathcal{F}$ closed under a countable number of elementary set operations, called σ-algebra, and imposing a notion of size on these, called a probability measure and usually denoted $\mathbb{P}$. The pair $(\mathcal{X}, \mathcal{F})$ is called a measurable space and any element of $\mathcal{F}$ is called a $\mathcal{F}$-measurable subset of $\mathcal{X}$. The triplet $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ is called a probability space. We now recall the definitions of each of these:

Definition 8 (**σ-algebra**, Williams [1991] p15-16)**.** *A collection $\mathcal{F}$ of subsets of some abstract set $\mathcal{X}$ is called a sigma-algebra of subsets of $\mathcal{X}$ if: (i) $\mathcal{X} \in \mathcal{F}$, (ii) $A \in \mathcal{F} \Rightarrow A^{\mathsf{c}} \in \mathcal{F}$ and (iii) $A_i \in \mathcal{F}$ for all $i \in \mathbb{N} \Rightarrow \bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.*

Note that using properties in the definition above, we can also show that a σ-algebra $\mathcal{F}$ satisfies the following property: $A_i \in \mathcal{F}$ for all $i \in \mathbb{N} \Rightarrow \bigcap_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

When $\mathcal{X}$ is a topological space, the most common example of σ-algebra is the Borel σ-algebra, denoted $\mathcal{B}(\mathcal{X})$, which consist of the σ-algebra generated by the family of open subsets of $\mathcal{X}$. This is the smallest σ-algebra of $\mathcal{X}$ such that the open subsets of $\mathcal{X}$ are included.

Now that we have our basic collection of subsets, we can construct a notion of size called a measure, in order to get a measure space:

Definition 9 (**Measure and probability space**, Williams [1991] p18)**.** *Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. A map $\mathbb{P} : \mathcal{F} \to [0, \infty]$ is called a measure on $(\mathcal{X}, \mathcal{F})$ if $\mathbb{P}$ is countably additive (or $\sigma-$additive, that is, satisfies: (i) $\mathbb{P}(\emptyset) = 0$, and (ii) $\{A_i\}_{i\in\mathbb{N}} \in \mathcal{F}$ are disjoint sets with $A = \bigcup_{i\in\mathbb{N}} A_i$, then $\mathbb{P}(A) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$. Then $\mathbb{P}$ is a measure and the triple $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ is a measure space. Furthermore, if $\mathbb{P}(\mathcal{X}) = 1$, $\mathbb{P}$ is called a probability measure and $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ is called a probability space.*

If $A = \bigcup_{i=1}^n A_i$ implies $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A_i)$ only for finite n, then $\mathbb{P}$ is called a finitely additive measure.

In the case where $\mathcal{X}$ is a subset of $\mathbb{R}$ and $\mathcal{F} = \mathcal{B}(\mathcal{X})$, the most common example of measure is the Lebesgue measure. In general, when $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ is a probability space, the set $\mathcal{X}$ is often referred to as the sample space and any element $A \in \mathcal{F}$ is called an event. We say that an event $A \in \mathcal{F}$ happens almost surely if $\mathbb{P}(A) = 1$.

Let $\mathbb{P}_1, \mathbb{P}_2$ be two probability measures on the same measurable space $(\mathcal{X}, \mathcal{F})$. $\mathbb{P}_1$ is said to be absolutely continuous with respect to $\mathbb{P}_2$ if $\forall A \in \mathcal{F}$, $\mathbb{P}_2(A) = 0$ implies $\mathbb{P}_1(A) = 0$. If $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$ and $\mathbb{P}_2$ is absolutely continuous with respect to $\mathbb{P}_1$, then the two probability measures are said to be equivalent. Finally, $\mathbb{P}_1$ and $\mathbb{P}_2$ are said to be orthogonal if $\exists A \in \mathcal{F}$ such that $\mathbb{P}_1(A) = 1$ and $\mathbb{P}_2(A) = 0$.

Now that we have defined probability spaces, we can discuss the most important property of functions defined on these spaces:

Definition 10 (**Measurable function**, Williams [1991] p29-31)**.** *Let $(\mathcal{X}_1, \mathcal{F}_1)$ and $(\mathcal{X}_2, \mathcal{F}_2)$ be two measurable spaces. Suppose that $h : \mathcal{X}_1 \to \mathcal{X}_2$ and for $A \subseteq \mathcal{X}_2$,* *define* $h^{-1}(A) := \{\mathbf{x} \in \mathcal{X}_1 : h(\mathbf{x}) \in A\}$*. Then* h *is called* $\mathcal{F}_1/\mathcal{F}_2$*-measurable if* $h^{-1} : \mathcal{F}_2 \to \mathcal{F}_1$*, that is,* $h^{-1}(A) \in \mathcal{F}_1$*,* $\forall A \in \mathcal{F}_2$*.*

Note that the σ-algebra with respect to which a function is measurable is usually obvious from the context and we simply refer to the function h as being measurable rather than $\mathcal{F}_1/\mathcal{F}_2$-measurable. A useful property is that the sum, product, composition, infimum and supremum of a sequence of measurable functions is also measurable.

When $\mathcal{X}$ is a topological space, a common example is the class of Borel functions, which consists of all $\mathcal{B}(\mathcal{X})$-measurable functions. However, a much more important example of measurable function are random variables, which are simply measurable functions on probability spaces:

Definition 11 (**Random variable**, Williams [1991] p31)**.** *Let* $(\mathcal{X}_1, \mathcal{F}_1, \mathbb{P})$ *be a probability space and* $(\mathcal{X}_2, \mathcal{F}_2)$ *be a measurable space. We call random variable any function* $X : \mathcal{X}_1 \to \mathcal{X}_2$ *which is* $\mathcal{F}_1/\mathcal{F}_2$*-measurable.*

When $(\mathcal{X}_2, \mathcal{F}_2)$ is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we call the random variable a real-valued random variable. Let $(\mathcal{X}, \mathcal{F}, \mathbb{P})$ be a probability space and denote by X some real-valued random variable on this space. We call $\mathcal{L}_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ defined as $\mathcal{L}_X := \mathbb{P} \circ X^{-1}$ the law of the random variable X, and this is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The function $F_X : \mathbb{R} \to [0,1]$ defined as $F_X(c) := \mathcal{L}_X(-\infty, c] = \mathbb{P}(X \leq c)$ is then called the cumulative distribution function of the random variable X.

Finally, we say the sub-σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of a σ-algebra $\mathcal{F}$ are independent if, whenever $A_i \in \mathcal{F}_i$ for $i \in \mathbb{N}$ and $i_1, \ldots, i_n$ are distinct, then: $\mathbb{P}(\bigcap_{k=1}^n A_{i_k}) = \prod_{k=1}^n \mathbb{P}(A_{i_k})$. We also say that random variables are independent if the σ-algebras generated by these random variable are independent.

An important notion for this thesis is that of the Lebesgue integral $\int f \mathrm{d}\mu$ of a measurable function f against a measure μ. The definition is separated in three parts: (i) show that the integral of simple functions can be easily obtained in closed form, (ii) show that the integral of positive functions can be defined as the limit of integrals of simple functions, and finally (iii) write the integral of the function of interest as the sum of integrals of positive functions. Each step is highlighted below:

- Simple measurable functions can be defined as functions of the form $\tilde{f}(\mathbf{x}) = \sum_{k=1}^m a_k \delta(A_k)$ for $a_k \in [0, \infty]$ and $A_k \in \mathcal{F}$ (where $\delta(A) = 1$ if $\mathbf{x} \in A$, and $\delta(A) = 0$ otherwise). The integral of a simple function is then given by: $\int \tilde{f} \mathrm{d}\mu = \sum_{k=1}^m a_k \mu(A_k)$.

- By the monotone convergence theorem (p51 of Williams [1991]), we have that if a sequence of positive measurable functions f_n converges pointwise to f from below then $\int f_n \mathrm{d}\mu \to \int f \mathrm{d}\mu$. Note that this is meaningful even if the limit is infinite.

- For the integrand f of interest, define $f^+(\mathbf{x}) = \max(f(\mathbf{x}), 0)$ and $f^-(\mathbf{x}) = \max(-f(\mathbf{x}), 0)$. Then clearly f^+ and f^- are both positive measurable functions, and we can write $\int f \mathrm{d}\mu = \int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu$ (note that this only makes sense if we do not have both $\int f^+ \mathrm{d}\mu$ and $\int f^- \mathrm{d}\mu$ being infinite).

These three steps combined allow us to define the Lebesgue integral $\int f \mathrm{d}\mu$ in terms of limits of integrals of simple functions. A special case of Lebesgue integrals are expectations, in which case the integrand f is a random variable and the measure μ is a probability measure. These are sometimes denoted $\mathbb{E}_\mu[f]$.

An important notion for this thesis, defined as a Lebesgue integral, is that of the Radon-Nikodym derivative. Suppose that μ_1, μ_2 are two σ-finite measures on some measurable space $(\mathcal{X}, \mathcal{F})$ and assume that μ_2 is absolutely continuous with respect to μ_1. Then there exists a measurable function $f : \mathcal{X} \to [0, \infty)$, called Radon-Nikodym derivative, defined such that for any event A: $\mu_2(A) = \int_A f \mathrm{d}\mu_1$. The probability density function p of a probability measure $\mathbb{P}$ corresponds to the Radon-Nikodym derivative of this measure with respect to the Lebesgue measure.

Appendix B

Proofs of Theoretical Results

The second appendix contains the proofs of all of the main results in the thesis (which were omitted from the main text for brevity). These are classified by chapters, and presented in order of appearance in the thesis.

B.1 Proofs of Chapter 3

Proof of Proposition 1

Proof. Repeated application of Fubini's theorem on the expressions for the mean and covariance of g_n produces:

$$\begin{aligned}
\mathbb{E}[\Pi[g_n]] &= \int_\Omega \int_{\mathcal{X}} g_n(\mathbf{x},\omega)\Pi(\mathrm{d}\mathbf{x})\mathbb{P}(\mathrm{d}\omega) = \int_{\mathcal{X}} m_n(\mathbf{x})\Pi(\mathrm{d}\mathbf{x}). \\
\mathbb{V}[\Pi[g_n]] &= \int_\Omega \left[\int_{\mathcal{X}} g_n(\mathbf{x},\omega)\Pi(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} m_n(\mathbf{x})\Pi(\mathrm{d}\mathbf{x})\right]^2 \mathbb{P}(\mathrm{d}\omega) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_\Omega [g(\mathbf{x},\omega) - m_n(\mathbf{x})][g(\mathbf{x}',\omega) - m_n(\mathbf{x}')]\mathbb{P}(\mathrm{d}\omega)\Pi(\mathrm{d}\mathbf{x})\Pi(\mathrm{d}\mathbf{x}') \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} c_n(\mathbf{x},\mathbf{x}')\Pi(\mathrm{d}\mathbf{x})\Pi(\mathrm{d}\mathbf{x}').
\end{aligned}$$

The proof is completed by substituting the expressions for m_n and c_n into these equations. □

Proof of Proposition 2

Proof. From Equation 3.5 in Chapter 3 of the main text $e(\hat{\Pi};\Pi,\mathcal{H}_k) \leq \|\hat{\Pi}[k(\mathbf{x},\cdot)] - \Pi[k(\mathbf{x},\cdot)]\|_{\mathcal{H}_k}$. For the converse inequality, consider the specific integrand $f = \hat{\Pi}[k(\mathbf{x},\cdot)] - \Pi[k(\mathbf{x},\cdot)]$. Then, from the supremum definition of the dual norm,

$e(\hat{\Pi};\Pi,\mathcal{H}_k) \geq |\hat{\Pi}[f]-\Pi[f]|/\|f\|_{\mathcal{H}_k}$. Now we use the reproducing property:

$$\frac{|\hat{\Pi}[f]-\Pi[f]|}{\|f\|_{\mathcal{H}_k}} = \frac{|\langle f, \hat{\Pi}[k(\mathbf{x},\cdot)] - \Pi[k(\mathbf{x},\cdot)]\rangle_{\mathcal{H}_k}|}{\|f\|_{\mathcal{H}_k}} = \frac{e(\hat{\Pi};\Pi,\mathcal{H}_k)^2}{e(\hat{\Pi};\Pi,\mathcal{H}_k)} = e(\hat{\Pi};\Pi,\mathcal{H}_k).$$

□

Proof of Lemma 3

Proof. Assume without loss of generality that $\delta < \infty$. The distribution of $\Pi[g_n]$ is Gaussian with mean u_n and variance v_n. Since $v_n = e(\hat{\Pi}_{\mathrm{BQ}};\Pi,\mathcal{H}_k)^2$ we have $v_n \leq \gamma_n^2$. Now the posterior probability mass on $I_\delta{}^{\mathsf{c}}$ is given by $\int_{I_\delta{}^{\mathsf{c}}} \phi(r|u_n,v_n)\mathrm{d}r$, where $\phi(r|u_n,v_n)$ is the probability density function of the $\mathcal{N}(u_n,v_n)$ distribution. Denote by Φ the cumulative distribution function of a $\mathcal{N}(0,1)$. From the definition of δ we get the upper bound

$$\begin{aligned}\mathbb{P}\{\Pi[g_n] \notin I_\delta\} \quad &\leq \quad \int_{-\infty}^{\Pi[f]-\delta} \phi(r|u_n,v_n)\mathrm{d}r + \int_{\Pi[f]+\delta}^{\infty} \phi(r|u_n,v_n)\mathrm{d}r \\ &= \quad 1 + \Phi\Big(\underbrace{\frac{\Pi[f]-u_n}{\sqrt{v_n}}}_{(*)} - \frac{\delta}{\sqrt{v_n}}\Big) - \Phi\Big(\underbrace{\frac{\Pi[f]-u_n}{\sqrt{v_n}}}_{(*)} + \frac{\delta}{\sqrt{v_n}}\Big).\end{aligned}$$

From the definition of the WCE we have that the terms $(*)$ are bounded by $\|f\|_{\mathcal{H}_k} < \infty$, so that asymptotically as $\gamma_n \to 0$ we have

$$\begin{aligned}\mathbb{P}\{\Pi[g_n] \notin I_\delta\} \quad &\lesssim \quad 1 + \Phi\big(-\delta/\sqrt{v_n}\big) - \Phi\big(\delta/\sqrt{v_n}\big) \\ &\lesssim \quad 1 + \Phi\big(-\delta/\gamma_n\big) - \Phi\big(\delta/\gamma_n\big) \quad \lesssim \quad \mathrm{erfc}\big(\delta/\sqrt{2}\gamma_n\big).\end{aligned}$$

where $\mathrm{erfc}(x)$ denotes the complementary error function. The result follows from the fact that $\mathrm{erfc}(x) \lesssim \exp(-x^2/2)$ for x sufficiently small. □

Proof of Theorem 8

Proof. The assumption $\sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x},\mathbf{x}) < \infty$ implies that all $f \in \mathcal{H}_k$ are bounded on $\mathcal{X}$. For MC estimators, Lemma 33 of Song [2008] show that for these functions, the WCE converges in probability at the classical rate $e(\hat{\Pi}_{\mathrm{MC}};\Pi,\mathcal{H}_k) = O_P(n^{-1/2})$. For their corresponding Bayesian estimators, it follows straightforwardly from Lemma 1 that the root-n rate is an upper bound, and we hence have: $e(\hat{\Pi}_{\mathrm{BMC}};\Pi,\mathcal{H}_k) = O_P(n^{-\frac{1}{2}})$. Furthermore, the above consistency result applied to Lemma 3 gives the contraction result. □

Proof of Theorem 9

Proof. Consider first the case of IS points (MC points are a special case). Initially consider fixed states $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ (i.e. fixing the random seed) and $\mathcal{H}_k = \mathcal{H}_\alpha$. From a standard result in functional approximation due to Wu and Schaback [1993], see also Wendland [2005, Theorem 11.13], there exists $C > 0$ and $h_0 > 0$ such that, for all $\mathbf{x} \in \mathcal{X}$ and $h_{\mathbf{X}} < h_0$, $|f(\mathbf{x}) - m_n(\mathbf{x})| \leq C h_{\mathbf{X}}^\alpha \|f\|_{\mathcal{H}_k}$; i.e. $v(h) = h^\alpha$ for Sobolev spaces of smoothness α. We augment $\mathbf{X}$ with a finite number of states $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^m$ to ensure that $h_{\mathbf{X}\cup\mathbf{Y}} < h_0$ always holds. From the regression bound (Lemma 2),

$$\begin{aligned}\left|\hat{\Pi}_{\mathrm{BIS}}[f] - \Pi[f]\right| \;&\leq\; \|f - m_n\|_2 \;=\; \left(\int_{\mathcal{X}} (f(\mathbf{x}) - m_n(\mathbf{x}))^2 \, \Pi'(\mathrm{d}\mathbf{x})\right)^{1/2} \\ &\leq\; \left(\int_{\mathcal{X}} (C h_{\mathbf{X}\cup\mathbf{Y}}^\alpha \|f\|_{\mathcal{H}_k})^2 \, \Pi'(\mathrm{d}\mathbf{x})\right)^{1/2} \;=\; C h_{\mathbf{X}\cup\mathbf{Y}}^\alpha \|f\|_{\mathcal{H}_k}.\end{aligned}$$

It follows that $e(\hat{\Pi}_{\mathrm{BIS}}; \Pi, \mathcal{H}_k) \leq C_1 h_{\mathbf{X}\cup\mathbf{Y}}^\alpha$ for some $C_1 > 0$. Now, taking an expectation $\mathbb{E}_{\mathbf{X}}$ over the samples $\mathbf{X}$ generated IID from the importance sampling distribution Π', we have:

$$\mathbb{E}_{\mathbf{X}}[e(\hat{\Pi}_{\mathrm{BIS}}; \Pi, \mathcal{H}_k)] \;\leq\; C\mathbb{E}_{\mathbf{X}}[h_{\mathbf{X}\cup\mathbf{Y}}^\alpha] \;\leq\; C\mathbb{E}_{\mathbf{X}}[h_{\mathbf{X}}^\alpha]. \tag{B.1}$$

From Lemma 2 in Oates et al. [2018], we have a scaling relationship such that, for $h_{\mathbf{X}\cup\mathbf{Y}} < h_0$, we have $\mathbb{E}_{\mathbf{X}}[h_{\mathbf{X}}^\alpha] = O(n^{-\alpha/d+\epsilon})$ for $\epsilon > 0$ arbitrarily small. From Markov's inequality, convergence in mean implies convergence in probability and thus, using Equation B.1, we have $e(\hat{\Pi}_{\mathrm{BIS}}; \Pi, \mathcal{H}_k) = O_P(n^{-\alpha/d+\epsilon})$. This completes the proof for $\mathcal{H}_k = \mathcal{H}_\alpha$. More generally, if $\mathcal{H}_k$ is norm-equivalent to $\mathcal{H}_\alpha$ then the result follows from the fact that $e(\hat{\Pi}_{\mathrm{BIS}}; \Pi, \mathcal{H}_k) \leq \lambda e(\hat{\Pi}_{\mathrm{BIS}}; \Pi, \mathcal{H}_\alpha)$ for some $\lambda > 0$. Note that the same arguments follow for BMCMC, except that Lemma 3 in [Oates et al., 2018] should be used instead of Lemma 2. □

Proof of Theorem 10

Proof. The proof follows that of Theorem 9, but uses a different power function. From Table 11.1 in Wendland [2005], we obtain upper bounds on the power function for the Gaussian RBF, multiquadric and inverse-multiquadric kernels. In the case of the Gaussian RBF kernel, this is given by $v_1(h_{\mathbf{X}}) = \exp(-C_1|\log(h_{\mathbf{X}})|/h_{\mathbf{X}}) = \exp(-C_1/h_{\mathbf{X}}^{1-\epsilon'})$ for some $C_1 > 0$ and $\epsilon' > 0$ arbitrarily small. For the multiquadric and inverse-multiquadric kernels this is $v_2(h_{\mathbf{X}}) = \exp(-C_2/h_{\mathbf{X}})$ for some $C_2 > 0$. We are now interested in the behaviour of the WCE. For the Gaussian RBF kernel, this is given by $e(\hat{\Pi}_{\text{BIS}};\Pi,\mathcal{H}_k) = O_P(\exp(-Cn^{1/d-\epsilon})) = O_P(v_1(n^{-1/d+\epsilon})) = O_P(\exp(-C_1 n^{1/d-\epsilon''}))$, where $\epsilon'' > 0$ can be arbitrarily small, whilst for the multiquadric and inverse-multiquadric we have $e(\hat{\Pi}_{\text{BIS}};\Pi,\mathcal{H}_k) = O_P\big(\exp(-Cn^{1/d-\epsilon})\big) = O_P(v_2(n^{-1/d+\epsilon})) = O_P(\exp(-C_2 n^{1/d-\epsilon}))$. This completes the proof. Similarly to Theorem 9, the proof also follows for MC and MCMC points. □

Proof of Theorem 11

Proof. The Koksma-Hlawka inequality (Theorem 2.9 in Niederreiter [1992]) states that $|\Pi[f] - \Pi_{\text{QMC}}[f]| \leq D^*(\{\mathbf{x}_i\}_{i=1}^n) V_{\text{HK}}(f)$ where $V_{\text{HK}}(f)$ denotes total variation of f in the sense of Hardy and Krause, and $D^*(\{\mathbf{x}_i\}_{i=1}^n)$ is the star discrepancy. Taking the supremum over the unit ball of $\mathcal{H}_k$ and using Lemma 1:

$$\begin{aligned} e(\hat{\Pi}_{\text{BQMC}};\Pi,\mathcal{H}_k) &\leq e(\hat{\Pi}_{\text{QMC}};\Pi,\mathcal{H}_k) \leq e(\hat{\Pi}_{\text{QMC}};\Pi,\mathcal{H}_\alpha) \\ &\leq \sup_{\|f\|_{\mathcal{H}_\alpha}\leq 1} V_{\text{HK}}(f) D^*(\{\mathbf{x}_i\}_{i=1}^n) \end{aligned}$$

Now we have that $\exists C > 0$ such that $V_{\text{HK}}(f) \leq C$ for any f in a Sobolev space. We therefore have that

$$e(\hat{\Pi}_{\text{BQMC}};\Pi,\mathcal{H}_k) \leq C \times D^*(\{\mathbf{x}_i\}_{i=1}^n) \leq C_1 n^{-1+\epsilon}$$

for some $C_1 > 0$ since low-discrepancy sequence satisfy $D^*(\{\mathbf{x}_i\}_{i=1}^n) = O(n^{-1+\epsilon})$. □

Proof of Theorem 12

Proof. From Theorem 15.21 of Dick and Pillichshammer [2010], which assumes $\alpha \geq 2$, $\alpha \in \mathbb{N}$, the QMC rule $\hat{\Pi}_{\text{QMC}}$ based on a higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ net over $\mathbb{Z}_b$ for some prime b satisfies $e(\hat{\Pi}_{\text{QMC}};\Pi,\mathcal{H}_k) \leq C_{d,\alpha}(\log n)^{d\alpha} n^{-\alpha} = O(n^{-\alpha+\epsilon})$ for $\mathcal{S}^\alpha$ the Sobolev space of dominating mixed smoothness order α, where $C_{d,\alpha} > 0$ is a constant that depends only on d and α (but not on n). The result follows immediately from norm equivalence and Lemma 1. The contraction rate follows from Lemma 3. □

Proof of Proposition 3

Proof. Conditional on a value of λ and following Proposition 1, $\Pi[g_n]$ is is a Gaussian distribution with mean and variance given by $\mathbb{E}[\Pi[g_n]] = \Pi[c_0(\cdot,\mathbf{X})]\mathbf{C}_0^{-1}\mathbf{f}$ and:

$$\mathbb{V}[\Pi[g_n]] = \lambda\{\Pi\Pi[c_0(\cdot,\cdot)] - \Pi[c_0(\cdot,\mathbf{X})]\mathbf{C}_0^{-1}\Pi[c_0(\mathbf{X},\cdot)]\}$$

Furthermore, the posterior on the amplitude parameter satisfies

$$p(\lambda|\mathbf{f}) \quad \propto \quad p(\mathbf{f}|\lambda)p(\lambda) \;=\; \frac{1}{(2\pi)^{n/2}\lambda^{\frac{n}{2}+1}|\mathbf{C}_0|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\lambda}\mathbf{f}^\top\mathbf{C}_0^{-1}\mathbf{f}\right)$$

which corresponds to an inverse-gamma distribution with parameters $\alpha = \frac{n}{2}$ and $\beta = \frac{1}{2}\mathbf{f}^\top\mathbf{C}_0^{-1}\mathbf{f}$. We therefore have that $(\Pi[g_n], \lambda)$ is distributed as normal-inverse-gamma and the marginal distribution for $\Pi[g_n]$ is a Student-t distribution, as claimed. □

Proof of Proposition 4

Proof. Define $\mathbf{z} = \Pi[c(\mathbf{X},\cdot)]$ and ${}_a\mathbf{z} = {}_a\Pi[c(\mathbf{X},\cdot)]$. Let $\epsilon = {}_a\mathbf{z} - \mathbf{z}$, write ${}_a\hat{\Pi}_{\mathrm{BQ}} = \sum_{i=1}^n {}_a w_i^{\mathrm{BQ}}\delta(\mathbf{x}_i)$ and consider

$$\begin{aligned}
e({}_a\hat{\Pi}_{\mathrm{BQ}};\Pi,\mathcal{H}_c)^2 &= \|{}_a\hat{\Pi}_{\mathrm{BQ}}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)]\|_{\mathcal{H}_c}^2 \\
&= {}_a\mathbf{w}_{\mathrm{BQ}}^\top\mathbf{C}_a\mathbf{w}_{\mathrm{BQ}} - 2_a\mathbf{w}_{\mathrm{BQ}}^\top\mathbf{z} + \Pi\Pi[c] \\
&= (\mathbf{C}^{-1}{}_a\mathbf{z})^\top\mathbf{C}(\mathbf{C}^{-1}{}_a\mathbf{z}) - 2(\mathbf{C}^{-1}{}_a\mathbf{z})^\top\mathbf{z} + \Pi\Pi[c] \\
&= (\mathbf{z}+\epsilon)^\top\mathbf{C}^{-1}(\mathbf{z}+\epsilon) - 2(\mathbf{z}+\epsilon)^\top\mathbf{C}^{-1}\mathbf{z} + \Pi\Pi[c] \\
&= e(\hat{\Pi}_{\mathrm{BQ}};\Pi,\mathcal{H}_c)^2 + \epsilon^\top\mathbf{C}^{-1}\epsilon.
\end{aligned}$$

Use $\otimes$ to denote the tensor product of RKHS. Now, since

$$\epsilon_i \;=\; {}_a z_i - z_i \;=\; {}_a\hat{\Pi}[c(\mathbf{x},\mathbf{x}_i)] - \Pi[c(\mathbf{x},\mathbf{x}_i)] \;=\; \langle {}_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)], c(\cdot,\mathbf{x}_i)\rangle_{\mathcal{H}_c},$$

we have that:

$$\begin{aligned}
\epsilon^\top\mathbf{C}^{-1}\epsilon &= \sum_{i,i'}(\mathbf{C}^{-1})_{i,i'}\big\langle {}_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)], c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} \\
&\qquad\qquad \times\big\langle {}_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)], c(\cdot,\mathbf{x}_{i'})\big\rangle_{\mathcal{H}_c} \\
&= \Big\langle {}_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)]\big) \otimes \big({}_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)]\big), \\
&\qquad \sum_{i,i'}(\mathbf{C}^{-1})_{i,i'}c(\cdot,\mathbf{x}_i)\otimes c(\cdot,\mathbf{x}_{i'})\Big\rangle_{\mathcal{H}_c\otimes\mathcal{H}_c}
\end{aligned}$$

and hence

$$\epsilon^\top\mathbf{C}^{-1}\epsilon \;\leq\; \|{}_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)]\|_{\mathcal{H}_c}^2\Big\|\sum_{i,i'}(\mathbf{C}^{-1})_{i,i'}c(\cdot,\mathbf{x}_i)\otimes c(\cdot,\mathbf{x}_{i'})\Big\|_{\mathcal{H}_c\otimes\mathcal{H}_c}.$$

From Proposition 2 we have $\|_a\hat{\Pi}[c(\mathbf{x},\cdot)] - \Pi[c(\mathbf{x},\cdot)]\|_{\mathcal{H}_c} = e({}_a\hat{\Pi};\Pi,\mathcal{H}_c)$ so it remains to show that the second term is equal to $\sqrt{n}$. Indeed,

$$\begin{aligned}
&\Big\|\sum_{i,i'}(\mathbf{C}^{-1})_{i,i'}c(\cdot,\mathbf{x}_i)\otimes c(\cdot,\mathbf{x}_{i'})\Big\|^2_{\mathcal{H}_c}\\
&\quad= \sum_{i,i',l,l'}(\mathbf{C}^{-1})_{i,i'}(\mathbf{C}^{-1})_{l,l'}\big\langle c(\cdot,\mathbf{x}_i)\otimes c(\cdot,\mathbf{x}_{i'}), c(\cdot,\mathbf{x}_l)\otimes c(\cdot,\mathbf{x}_{l'})\big\rangle_{\mathcal{H}_c}\\
&\quad= \sum_{i,i',l,l'}(\mathbf{C}^{-1})_{i,i'}(\mathbf{C}^{-1})_{l,l'}(\mathbf{C})_{il}(\mathbf{C})_{i',l'} \;=\; \mathrm{Tr}(\mathbf{C}\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}) \;=\; n.
\end{aligned}$$

This completes the proof. □

Proof of Proposition 5

Proof. The proof follows by combining Theorem 15.21 of Dick and Pillichshammer [2010] with Lemma 1. □

Proof of Proposition 6

Proof. The first result follows from the regression bound argument (Lemma 2) together with a functional approximation result in Le Gia et al. [2012, Theorem 3.2].

The result for QMC with spherical t-designs follows from combining Hesse and Sloan [2005]; Bondarenko et al. [2013] and Lemma 1. □

B.2 Proofs of Chapter 4

Proof of Proposition 8

Proof. Denote by $\mathbf{e}_p$ the vertical vector of length P with d^{th} entry taking value 1 and all other entries taking value 0, and by $\mathbf{C}^p_{\mathbf{x}}(\mathbf{y}) = \mathbf{C}(\mathbf{y},\mathbf{x})\mathbf{e}_p$ the d^{th} column of $\mathbf{C}(\mathbf{y},\mathbf{x})$. We notice that the representer of the integral is given by:

$$\Pi[f_p] \;=\; \Pi[\mathbf{f}^\top\mathbf{e}_p] \;=\; \Pi\left[\langle\mathbf{f},\mathbf{C}(\cdot,\mathbf{x})\mathbf{e}_p\rangle_{\mathcal{H}_\mathbf{C}}\right] \;=\; \langle\mathbf{f},\Pi\left[\mathbf{C}(\cdot,\mathbf{x})\mathbf{e}_p\right]\rangle_{\mathcal{H}_\mathbf{C}} \;=\; \langle\mathbf{f},\Pi\left[\mathbf{C}^p_{\mathbf{x}}\right]\rangle_{\mathcal{H}_\mathbf{C}}$$

and so, using the Cauchy-Schwartz inequality, we get: $|\Pi[f_p]-\hat{\Pi}[f_p]| \leq \|\mathbf{f}\|_{\mathcal{H}_\mathbf{C}}\|\Pi[\mathbf{C}^p_{\mathbf{x}}]-\hat{\Pi}[\mathbf{C}^p_{\mathbf{x}}]\|_{\mathcal{H}_\mathbf{C}}$. Taking supremums, we obtain the following expression for the worst-case integration error:

$$\sup_{\|f\|_{\mathcal{H}_\mathbf{C}}\leq 1}\left|\Pi[f_p]-\hat{\Pi}[f_p]\right| \;=\; \left\|\Pi[\mathbf{C}^p_{\mathbf{x}}]-\hat{\Pi}[\mathbf{C}^p_{\mathbf{x}}]\right\|_{\mathcal{H}_\mathbf{C}}$$

We note that $\Pi[\mathbf{C}^p_\mathbf{x}] \in \mathcal{H}_\mathbf{C}$ and that the multi-output BQ rule is given by $\hat{\Pi}_{\text{BQ}}[\mathbf{C}^p_\mathbf{x}] = \Pi[\mathbf{C}(\cdot,\mathbf{X})]\mathbf{C}(\mathbf{X},\mathbf{X})^{-1}\mathbf{C}^p_\mathbf{x}(\mathbf{X})$ and corresponds to an optimal interpolant in the sense of Theorem 3.1 in Micchelli and Pontil [2005]. We must therefore have that, for fixed quadrature points $\mathbf{X}$, any quadrature rule $\hat{\Pi}[\mathbf{C}^p_\mathbf{x}]$ satisfies:

$$\left\|\Pi[\mathbf{C}^p_\mathbf{x}] - \hat{\Pi}_{\text{BQ}}[\mathbf{C}^p_\mathbf{x}]\right\|_{\mathcal{H}_\mathbf{C}} \quad \leq \quad \left\|\Pi[\mathbf{C}^p_\mathbf{x}] - \hat{\Pi}[\mathbf{C}^p_\mathbf{x}]\right\|_{\mathcal{H}_\mathbf{C}}.$$

Combining the equation above with the expression for the worst-case integration error of f_p gives us our final result.

□

Proof of Theorem 13

Proof. For the sake of clarity, we will distinguish between uni-output BQ and multi-output BQ rules and weights by adding subscripts corresponding to their kernel; i.e. $\Pi^\mathbf{C}_{\text{BQ}}[f]$ and $\mathbf{W}^\mathbf{C}_{\text{BQ}}$ denote the multi-output case and $\Pi^c_{\text{BQ}}[f]$ and $\mathbf{W}^c_{\text{BQ}}$ denote the uni-output case. We start this proof by expressing the weights of the multi-output BQ algorithm in terms of weights for the uni-output BQ algorithm:

$$\begin{aligned}
\mathbf{W}^\mathbf{C}_{\text{BQ}} &= \Pi[\mathbf{C}(\cdot,\mathbf{X})]\mathbf{C}(\mathbf{X},\mathbf{X})^{-1} \;=\; (\Pi[\mathbf{B}\otimes\mathbf{c}(\cdot,\mathbf{X})])\,(\mathbf{B}\otimes c(\mathbf{X},\mathbf{X}))^{-1} \\
&= (\mathbf{B}\otimes\Pi[\mathbf{c}(\cdot,\mathbf{X})])\left(\mathbf{B}^{-1}\otimes c(\mathbf{X},\mathbf{X})^{-1}\right) \;=\; \mathbf{B}\mathbf{B}^{-1}\otimes\Pi[\mathbf{c}(\cdot,\mathbf{X})]c(\mathbf{X},\mathbf{X})^{-1} \\
&= \mathbf{I}_D\otimes\mathbf{w}^c_{\text{BQ}}.
\end{aligned}$$

Using the above, we can find an expression for the multi-output BQ approximation with some kernel $\mathbf{C}_1 = \mathbf{B}c_1$ of the project mean element with respect to kernel $\mathbf{C}_2 = \mathbf{B}c_2$ in terms of the uni-output BQ approximation with kernel c_1 of the kernel mean of c_2.

$$\begin{aligned}
\hat{\Pi}^{\mathbf{C}_1}_{\text{BQ}}[(\mathbf{C}_2)^p_\mathbf{x}] &= (\mathbf{C}_2)^p_\mathbf{x}(\mathbf{X})\mathbf{W}^{C_1}_{\text{BQ}} \;=\; (\mathbf{C}_2)^p_\mathbf{x}(\mathbf{X})(I\otimes\mathbf{w}^{c_1}_{\text{BQ}}) \;=\; (\mathbf{B}\mathbf{e}_p\otimes c_2(\mathbf{X},\mathbf{x}))(\mathbf{I}\otimes\mathbf{w}^{c_1}_{\text{BQ}}) \\
&= \mathbf{B}\mathbf{e}_p\mathbf{I}\otimes c_2(\mathbf{X},\mathbf{x})\mathbf{w}^{c_1}_{\text{BQ}} \;=\; \mathbf{B}\mathbf{e}_p\hat{\Pi}^{c_1}_{\text{BQ}}[c_2(\cdot,\mathbf{x})].
\end{aligned}$$

As discussed, taking both kernels to be the same, the integration error for each individual integrand can be bounded as follows:

$$\begin{aligned}
\sup_{\|f\|_{\mathcal{H}_{\mathbf{C}_2}}\leq 1} \left|\Pi[f_p] - \hat{\Pi}_{\mathrm{BQ}}^{\mathbf{C}_1}[f_p]\right|^2 &= \left\|\Pi\left[(\mathbf{C}_2)_{\mathbf{x}}^p\right] - \hat{\Pi}_{\mathrm{BQ}}^{\mathbf{C}_1}\left[(\mathbf{C}_2)_{\mathbf{x}}^p\right]\right\|_{\mathcal{H}_{\mathbf{C}_2}}^2 \\
&= \left\|(\mathbf{B}\mathbf{e}_p)\left(\Pi\left[c_2(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}\left[c_2(\cdot,\mathbf{x})\right]\right)\right\|_{\mathcal{H}_{\mathbf{C}_2}}^2 \\
&= \sum_{i,j=1}^{P} (\mathbf{B}^{-1})_{ij} \times \Big\langle \mathbf{B}_{ip}(\Pi\left[c_2(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}\left[c_2(\cdot,\mathbf{x})\right]), \\
&\qquad \mathbf{B}_{jp}(\Pi\left[c_2(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}\left[c_2(\cdot,\mathbf{x})\right])\Big\rangle_{\mathcal{H}_{c_2}} \\
&= \sum_{i,j=1}^{P} (\mathbf{B}^{-1})_{ij}\mathbf{B}_{ip}\mathbf{B}_{jp} \left\|\Pi\left[c_2(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}\left[c_2(\cdot,\mathbf{x})\right]\right\|_{\mathcal{H}_{c_2}}^2 \\
&\leq K \left\|\Pi\left[c_2(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}\left[c_2(\cdot,\mathbf{x})\right]\right\|_{\mathcal{H}_{c_2}}^2 .
\end{aligned}$$

Here, we first used the definition of worst-case error, then the definition of the $\mathcal{H}_{\mathbf{C}_2}$ norm in terms of $\mathcal{H}_{c_2}$ norm (as given for the separable kernel in Álvarez and Lawrence [2011]), and the final inequality follows by taking $K > 0$ to be $K = |\sum_{i,j=1}^{P}(\mathbf{B}^{-1})_{ij}\mathbf{B}_{ip}\mathbf{B}_{jp}|$. Taking the square-root on either side gives us:

$$\begin{aligned}
\sup_{\|f\|_{\mathcal{H}_{\mathbf{C}_2}}\leq 1} \left|\Pi[f_p] - \hat{\Pi}_{\mathrm{BQ}}^{\mathbf{C}_1}[f_p]\right| &\leq \sqrt{K}\left\|\Pi\left[c_2(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}\left[c_2(\cdot,\mathbf{x})\right]\right\|_{\mathcal{H}_{c_2}} \\
&= \sqrt{K} \sup_{\|f\|_{\mathcal{H}_{c_2}}\leq 1} \left|\Pi[f_p] - \hat{\Pi}_{\mathrm{BQ}}^{c_1}[f_p]\right| .
\end{aligned}$$

We can take $\mathbf{C}_1$ equal to $\mathbf{C}_2$ to get:

$$\sup_{\|f\|_{\mathcal{H}_{\mathbf{C}}}\leq 1} \left|\Pi[f_p] - \hat{\Pi}_{\mathrm{BQ}}^{\mathbf{C}}[f_p]\right| \leq \sqrt{K} \sup_{\|f\|_{\mathcal{H}_{c}}\leq 1} \left|\Pi[f_p] - \hat{\Pi}_{\mathrm{BQ}}^{c}[f_p]\right| .$$

The convergence for the separable kernel case is therefore driven by the convergence of the scalar-valued kernel. We can therefore use results from the uni-output case in the previous chapter or in Briol et al. [2015b]; Oates et al. [2018]; Briol et al. [2017]; Kanagawa et al. [2017] to complete the proof. □

Proof of Proposition 9

Proof. Note that if the kernel is actually of the form $\mathbf{C}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q}\mathbf{B}_q c_q(\mathbf{x},\mathbf{x}')$, we can use the triangle inequality satisfied by the norm of $\mathcal{H}_{\mathbf{C}}$ to show that (for some $C_2 > 0$):

$$\sup_{\|f\|_{\mathcal{H}_{\mathbf{C}}}\leq 1} \left|\Pi[f_p] - \hat{\Pi}_{\mathrm{BQ}}[f_p]\right| \;\leq\; C_2 \sum_{q=1}^{Q} \left\|\Pi\left[c_q(\cdot,\mathbf{x})\right] - \hat{\Pi}_{\mathrm{BQ}}\left[c_q(\cdot,\mathbf{x})\right]\right\|^2_{\mathcal{H}_c},$$

so that the overall convergence is dominated by the slowest decaying term. □

Proof of Theorem 14

Proof. Recall that $h_{\mathbf{X}}$ denotes the fill distance and $\rho_{\mathbf{X}}$ denotes the mesh ratio. Denote by $\hat{\Pi}^{\mathbf{C}_\alpha}_{BQ}[\mathbf{f}]$ the multi-output BQ rule based on $\mathbf{C}_\alpha$, $\hat{\Pi}^{c_\alpha}_{BQ}[f]$ the uni-output BQ rule based on c_α and $\hat{f}^\alpha_p$ the interpolant corresponding this rule. We start by upper bounding the integration error in the uni-output case:

$$\begin{aligned}\left|\Pi[f] - \hat{\Pi}^{c_\alpha}_{BQ}[f]\right| \;&\leq\; K_1\|\pi\|_{L_\infty(\mathcal{X})}\|f-\hat{f}^\alpha\|_{L_1(\mathcal{X})} \;\leq\; K_2\|f-\hat{f}^\alpha\|_{L^2(\mathcal{X})} \\ &\leq\; K_3 h^\beta_{\mathbf{X}}\rho^\alpha_{\mathbf{X}}\|f\|_{L^2(\mathcal{X})} \;\leq\; K_4 h^\beta_{\mathbf{X}}\rho^\alpha_{\mathbf{X}}\|f\|_{W^\beta_2(\mathcal{X})} \;\leq\; K_5 h^\beta_{\mathbf{X}}\rho^\alpha_{\mathbf{X}}\|f\|_{\mathcal{H}_{c_\beta}},\end{aligned}$$

for some $K_1, \ldots, K_5 > 0$. Note that this argument closely follows Kanagawa et al. [2017]. The first and second inequality correspond to Holder's inequality and the third inequality follows from Theorem 4.2 in Narcowich et al. [2006]. Finally, the fourth and fifth inequalities follow from the definition the Sobolev norm and the norm-equivalence of $\mathcal{H}_{c_\beta}$ and $W^\beta_2(\mathcal{X})$.

Dividing the above by $\|f_p\|_\beta$ on both sides and taking supremums over the unit ball of $\mathcal{H}_{c_\beta}$ we get a result for the worst-case error in the uni-output case: $e(\mathcal{H}_{c_\beta}, \hat{\Pi}^{c_\alpha}_{BQ}, \mathbf{X}) \leq K_6 h^\beta_{\mathbf{X}}\rho^\alpha_{\mathbf{X}}$. We can then upper bound the integration error in the multi-output case using Theorem 13 as follows:

$$\begin{aligned}\left|\Pi[f_p] - \hat{\Pi}^{\mathbf{C}_\alpha}_{BQ}[f_p]\right| \;&\leq\; \|\mathbf{f}\|_{\mathbf{C}_\beta} e(\mathcal{H}_{\mathbf{C}_\beta}, \hat{\Pi}^{\mathbf{C}_\alpha}_{BQ}, \mathbf{X}, p) \;\leq\; K_6\|\mathbf{f}\|_{\mathbf{C}_\beta} e(\mathcal{H}_{c_\beta}, \hat{\Pi}^{c_\alpha}_{BQ}, \mathbf{X}) \\ &\leq\; K_7\|\mathbf{f}\|_{\mathbf{C}_\beta} h^\beta_{\mathbf{X}}\rho^\alpha_{\mathbf{X}},\end{aligned}$$

for some $K_6, K_7 > 0$. When using a quasi-uniform grid, then we can use the assumption that $h_{\mathbf{X}} \leq C q_{\mathbf{X}}$ for some constant $C > 0$ and the fact that $h_{\mathbf{X}}$ converges as $n^{-\frac{1}{d}}$ to show that the integration error satisfies:

$$\left|\Pi[f_p] - \hat{\Pi}^{\mathbf{C}_\alpha}_{BQ}[f_p]\right| \;\leq\; K_7\|\mathbf{f}\|_{\mathbf{C}_\beta} h^\beta_{\mathbf{X}}\rho^\alpha_{\mathbf{X},\mathcal{X}} \;\leq\; K_8\|\mathbf{f}\|_{\mathbf{C}_\beta} h^\beta_{\mathbf{X}} \;=\; O\left(n^{-\frac{\beta}{d}}\right),$$

for some $K_8 > 0$. □

Proof of Proposition 10

Proof. The results follows from combining Theorem 13 with the rate for the scalar-valued Matérn $\frac{3}{2}$ covariance function provided in Proposition 6. □

Proof of Proposition 11

Proof. First, from the definition of J:

$$
\begin{aligned}
J\big((1-\rho)g_{i-1}+\rho c(\cdot,\mathbf{x}_i)\big) \quad &= \quad \frac{1}{2}\Big\langle(1-\rho)g_{i-1}+\rho c(\cdot,\mathbf{x}_i)-\Pi[c(\cdot,\mathbf{x})], \\
&\qquad\qquad (1-\rho)g_{i-1}+\rho c(\cdot,\mathbf{x}_i)-\Pi[c(\cdot,\mathbf{x})]\Big\rangle_{\mathcal{H}_c} \\
&= \quad \frac{1}{2}\Big[(1-\rho)^2\big\langle g_{i-1},g_{i-1}\big\rangle_{\mathcal{H}_c}+2(1-\rho)\rho\big\langle g_{i-1},c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} \\
&\qquad +2\rho^2\big\langle c(\cdot,\mathbf{x}_i),c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c}-2(1-\rho)\big\langle g_{i-1},\Pi[c(\cdot,\mathbf{x})]\big\rangle_{\mathcal{H}_c} \\
&\qquad -2\rho\big\langle c(\cdot,\mathbf{x}_i),\Pi[c(\cdot,\mathbf{x})]\big\rangle_{\mathcal{H}_c}+\big\langle \Pi[c(\cdot,\mathbf{x})],\Pi[c(\cdot,\mathbf{x})]\big\rangle_{\mathcal{H}_c}\Big].
\end{aligned}
$$

Taking the derivative of this expression with respect to ρ, we get:

$$
\begin{aligned}
\frac{\partial J\big((1-\rho)g_{i-1}+\rho c(\cdot,\mathbf{x}_i)\big)}{\partial\rho} \quad &= \quad \frac{1}{2}\Big[-2(1-\rho)\big\langle g_{i-1},g_{i-1}\big\rangle_{\mathcal{H}_c}+2(1-2\rho)\big\langle g_{i-1},c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} \\
&\qquad +2\rho\big\langle c(\cdot,\mathbf{x}_i),c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c}+2\big\langle g_{i-1},\Pi[c(\cdot,\mathbf{x})]\big\rangle_{\mathcal{H}_c} \\
&\qquad -2\big\langle c(\cdot,\mathbf{x}_i),\Pi[c(\cdot,\mathbf{x})]\big\rangle_{\mathcal{H}_c}\Big] \\
&= \quad \rho\Big[\big\langle g_{i-1},g_{i-1}\big\rangle_{\mathcal{H}_c}-2\big\langle g_{i-1},c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c}+\big\langle c(\cdot,\mathbf{x}_i),c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} \\
&= \quad \rho\big\|g_{i-1}-c(\cdot,\mathbf{x}_i)\big\|^2_{\mathcal{H}_c}-\big\langle g_{i-1}-c(\cdot,\mathbf{x}_i),g_{i-1}-\Pi[c(\cdot,\mathbf{x})]\big\rangle_{\mathcal{H}_c}.
\end{aligned}
$$

Setting this derivative to zero gives us the following optimum:

$$
\rho^* \;=\; \frac{\Big\langle g_{i-1}-\Pi[c(\cdot,\mathbf{x})],g_{i-1}-c(\cdot,\mathbf{x}_i)\Big\rangle_{\mathcal{H}_c}}{\Big\|g_{i-1}-c(\cdot,\mathbf{x}_i)\Big\|^2_{\mathcal{H}_c}}.
$$

Clearly, differentiating a second time with respect to ρ gives $\|g_{i-1}-c(\cdot,\mathbf{x}_i)\|^2_{\mathcal{H}_c}$, which is non-negative and so ρ^* is a minimum. One can show using geometrical arguments about the marginal polytope $\mathcal{M}$ that ρ^* will be in $[0,1]$ [Jaggi, 2013].

The numerator of this line-search expression is

$$\begin{aligned}
&\Big\langle g_{i-1} - \Pi[c(\cdot,\mathbf{x})], g_{i-1} - c(\cdot,\mathbf{x}_i)\Big\rangle_{\mathcal{H}_c} \\
&= \big\langle g_{i-1}, g_{i-1}\big\rangle_{\mathcal{H}_c} - \big\langle \Pi[c(\cdot,\mathbf{x})], g_{i-1}\big\rangle_{\mathcal{H}_c} - \sum_{l=1}^{i-1} w_l^{(i-1)} c(\mathbf{x}_l,\mathbf{x}_i) + \Pi[c(\cdot,\mathbf{x}_i)] \\
&= \sum_{l=1}^{i-1}\sum_{m=1}^{i-1} w_l^{(i-1)} w_m^{(i-1)} c(\mathbf{x}_l,\mathbf{x}_m) - \sum_{l=1}^{i-1} w_l^{(i-1)} \Big[c(\mathbf{x}_l,\mathbf{x}_i) + \Pi[c(\mathbf{x}_l,\mathbf{x})]\Big] + \Pi[c(\mathbf{x}_i,\mathbf{x})].
\end{aligned}$$

Similarly the denominator is

$$\begin{aligned}
\big\|g_{i-1} - c(\cdot,\mathbf{x}_i)\big\|_{\mathcal{H}_c}^2 &= \big\langle g_{i-1} - c(\cdot,\mathbf{x}_i), g_{i-1} - c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} \\
&= \big\langle g_{i-1}, g_{i-1}\big\rangle_{\mathcal{H}_c} - 2\big\langle g_{i-1}, c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} + \big\langle c(\cdot,\mathbf{x}_i), c(\cdot,\mathbf{x}_i)\big\rangle_{\mathcal{H}_c} \\
&= \sum_{l=1}^{i-1}\sum_{m=1}^{i-1} w_l^{(i-1)} w_m^{(i-1)} c(\mathbf{x}_l,\mathbf{x}_m) - 2\sum_{l=1}^{i-1} w_l^{(i-1)} c(\mathbf{x}_l,\mathbf{x}_i) + c(\mathbf{x}_i,\mathbf{x}_i).
\end{aligned}$$

□

Proof of Theorem 15

Proof. Using Lemma 1 from Chapter 3, we have that BQ rules are optimally weighted in $\mathcal{H}_c$ and so we have that $e(\hat{\Pi}_{\text{FWBQ}}; \Pi, \mathcal{H}_c) \leq e(\hat{\Pi}_{\text{FW}}; \Pi, \mathcal{H}_c)$ and $e(\hat{\Pi}_{\text{FWLSBQ}}; \Pi, \mathcal{H}_c) \leq e(\hat{\Pi}_{\text{FWLS}}; \Pi, \mathcal{H}_c)$. Now, the values attained by the objective function J along the path $\{g_i\}_{i=1}^n$ determined by the FW and FWLS algorithm can be expressed in terms of the half the WCE squared. We therefore have that: $e(\hat{\Pi}_{\text{FWBQ}}; \Pi, \mathcal{H}_c)^2 \|f\|_{\mathcal{H}_c} \leq 2^{1/2} J_{\text{FW}}^{1/2}(g_n)$ and $e(\hat{\Pi}_{\text{FWLSBQ}}; \Pi, \mathcal{H}_c)^2 \|f\|_{\mathcal{H}_c} \leq 2^{1/2} J_{\text{FWLS}}^{1/2}(g_n)$, since $\|f\|_{\mathcal{H}_c} \leq 1$. To complete the proof we leverage recent analysis of the FW algorithm with steps $\rho_i = 1/(n+1)$ and the FWLS algorithm. Specifically, from [Bach et al., 2012, Proposition 1] we have that:

$$J(g_n) \leq \begin{cases} \frac{2\text{diam}(\mathcal{M})^4}{R^2} n^{-2} & \text{for FW with step size } \rho_i = 1/(i+1) \\ \text{diam}(\mathcal{M})^2 \exp(-R^2 n / \text{diam}(\mathcal{M})^2) & \text{for FWLS} \end{cases}$$

where $\text{diam}(\mathcal{M})$ is the diameter of the marginal polytope $\mathcal{M}$ and R is the radius of the smallest ball centered at $\Pi[c(\cdot,\mathbf{x})]$ included in $\mathcal{M}$. This proves our consistency result, and the contraction result follows from Lemma 3. □

B.3 Proofs of Chapter 5

Proof of Proposition 12

Proof of Proposition 12. Consider the diffusion Stein discrepancy, obtained by combining the expression for the Stein discrepancy with the diffusion-based Stein operator $\mathcal{S}_{\mathbb{P}}[g](\mathbf{x})$ and the function class $\mathcal{G}$. We first note that

$$\mathcal{S}_{\mathbb{P}}[g](\mathbf{x}) = \langle m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), g(\mathbf{x})\rangle + \langle \nabla_{\mathbf{x}}, m(\mathbf{x})g(\mathbf{x})\rangle.$$

The discrepancy with this operator is then given by

$$\begin{aligned} D(\mathbb{P}_1||\mathbb{P}_2) &= \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}_2}[g](\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}) \right| = \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} (\mathcal{S}_{\mathbb{P}_2}[g](\mathbf{x}) - \mathcal{S}_{\mathbb{P}_1}[g](\mathbf{x}))\mathbb{P}_1(\mathrm{d}\mathbf{x}) \right| \\ &= \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} \langle m(\mathbf{x})^\top (\nabla_{\mathbf{x}} \log p_2(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_1(\mathbf{x})), g(\mathbf{x})\rangle p_1(\mathbf{x})\mathrm{d}\mathbf{x} \right|, \end{aligned}$$

Using the Cauchy-Schwarz inequality, we get:

$$D(\mathbb{P}_1||\mathbb{P}_2) \leq \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} \left\| m(\mathbf{x})^\top \left(\nabla_{\mathbf{x}} \log p_2(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_1(\mathbf{x})\right) \right\|_2 \|g(\mathbf{x})\|_2 \, p_1(\mathbf{x})\mathrm{d}\mathbf{x} \right|.$$

This inequality is tight, and attained when $g(\mathbf{x}) = m(\mathbf{x})^\top (\nabla_{\mathbf{x}} \log p_2(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_1(\mathbf{x}))$, so that the supremum is attained at that point. We therefore end up with a discrepancy of the form:

$$D(\mathbb{P}_1||\mathbb{P}_2) = \int_{\mathcal{X}} \left\| m(\mathbf{x})^\top \left(\nabla_{\mathbf{x}} \log p_2(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_1(\mathbf{x})\right) \right\|_2^2 \mathbb{P}_1(\mathrm{d}\mathbf{x}).$$

In order to obtain a computable estimator, we will follow the proof of Theorem 1 in Hyvärinen [2006] and use an integration-by-part trick. To do so, we first expand the integrand in the expression for the discrepancy and take $p_1(\mathbf{x}) = q(\mathbf{x})$ (the density of the data-generating model $\mathbb{Q}$) and $p_2(\mathbf{x}) = p(\mathbf{x}|\theta)$:

$$\begin{aligned} & \|m(\mathbf{x})^\top (\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) - \nabla_{\mathbf{x}} \log q(\mathbf{x}))\|_2^2 \\ = \; & \|m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta)\|_2^2 + \|m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log q(\mathbf{x})\|_2^2 \\ & -2\langle m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log q(\mathbf{x})\rangle \end{aligned}$$

When integrating the above, the second term does not depend on θ and can hence be ignored for the purpose of minimisation over parameters. We then end up with:

$$\int_{\mathcal{X}} \left[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) m(\mathbf{x})\|_2^2 - 2 \left\langle m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right\rangle \right] \mathbb{Q}(\mathrm{d}\mathbf{x})$$

Using integration-by-parts, we can get obtain an expression for the second term which does not depend on the density q:

$$\begin{aligned}
& \int_{\mathcal{X}} \left\langle m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right\rangle \mathbb{Q}(\mathrm{d}\mathbf{x}) \\
= & \int_{\mathcal{X}} \left\langle \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta), m(\mathbf{x}) m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right\rangle \mathbb{Q}(\mathrm{d}\mathbf{x}) \\
= & -\int_{\mathcal{X}} \left\langle \nabla_{\mathbf{x}}, m(\mathbf{x}) m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) \right\rangle q(\mathbf{x}) \mathrm{d}\mathbf{x} \\
= & -\int_{\mathcal{X}} \left\langle \nabla_{\mathbf{x}}, m(\mathbf{x}) m(\mathbf{x})^\top \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) \right\rangle \mathbb{Q}(\mathrm{d}\mathbf{x})
\end{aligned}$$

Combing this equation with the previous one completes the proof. □

Proof of Proposition 13

Proof. The information metric is defined as: $g(\theta) = -\frac{\partial^2 \mathrm{KSD}(\mathbb{P}_\alpha || \mathbb{P}_\beta)^2}{\partial \alpha \partial \beta}\big|_{\alpha=\beta=\theta}$. We hence require the following expression, where the Stein reproducing kernel is adapted to the measure $\mathbb{P}_\alpha$:

$$\begin{aligned}
\frac{\partial^2 \mathrm{KSD}(\mathbb{P}_\alpha || \mathbb{P}_\beta)^2}{\partial \alpha \partial \beta} &= \frac{\partial^2}{\partial \alpha \partial \beta} \Big[\int_{\mathcal{X}} \int_{\mathcal{X}} k_{\mathbb{P}_\beta}(\mathbf{x}, \mathbf{y}) p(\mathbf{x}|\alpha) p(\mathbf{y}|\alpha) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \\
& \quad -2 \int_{\mathcal{X}} \int_{\mathcal{X}} k_{\mathbb{P}_\beta}(\mathbf{x}, \mathbf{y}) p(\mathbf{x}|\alpha) p(\mathbf{y}|\beta) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \\
& \quad + \int_{\mathcal{X}} \int_{\mathcal{X}} k_{\mathbb{P}_\beta}(\mathbf{x}, \mathbf{y}) p(\mathbf{x}|\beta) p(\mathbf{y}|\beta) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \Big] \\
&= -2 \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{\partial^2 [k_{\mathbb{P}_\beta}(\mathbf{x}, \mathbf{y}) p(\mathbf{x}|\alpha) p(\mathbf{y}|\beta)]}{\partial \alpha \partial \beta} \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \\
&= -2 \sum_{l=1}^{d} \int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) \frac{\partial^2 \log p(\mathbf{x}|\alpha)}{\partial x_l \partial \alpha_j} \frac{\partial^2 \log p(\mathbf{y}|\alpha)}{\partial y_l \partial \alpha_k} \mathbb{P}_\alpha(\mathrm{d}\mathbf{x}) \mathbb{P}_\beta(\mathrm{d}\mathbf{y}).
\end{aligned}$$

The proof is completed by taking $\alpha = \beta$ in the expression above. □

Proof of Proposition 14

Proof. In the case of exponential families, the score function can be expressed as: $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta) = \theta^\top \nabla T(\mathbf{x}) + \nabla b(\mathbf{x})$. Combining this expression together with Equations 5.25 and 5.26 gives us our first result:

$$\mathrm{KSD}_U(\mathbb{Q}^m||\mathbb{P}_\theta)^2 \quad = \quad \theta^\top A(\{\mathbf{y}_j\}_{j=1}^m)\theta + B(\{\mathbf{y}_j\}_{j=1}^m)\theta + C(\{\mathbf{y}_j\}_{j=1}^m)$$

Taking the derivative with respect to the parameter vector θ and setting to zero gives $\theta^\top A(\{\mathbf{y}_j\}_{j=1}^m) + B(\{\mathbf{y}_j\}_{j=1}^m) = 0$. Solving this system of linear equations then gives $\theta = -B(\{\mathbf{y}_j\}_{j=1}^m)A(\{\mathbf{y}_j\}_{j=1}^m)^{-1}$, which concludes the proof. □

Proof of Proposition 15

Proof. The information metric is defined as: $g(\theta) = -\frac{\partial^2 \mathrm{MMD}(\mathbb{P}_\alpha||\mathbb{P}_\beta)^2}{\partial\alpha\partial\beta}\big|_{\alpha=\beta=\theta}$. We hence require the following expression:

$$\begin{aligned}
\frac{\partial^2 \mathrm{MMD}(\mathbb{P}_\alpha||\mathbb{P}_\beta)^2}{\partial\alpha\partial\beta} \quad &= \quad \frac{\partial^2}{\partial\alpha\partial\beta}\Big[\int_{\mathcal{U}}\int_{\mathcal{U}} k(G_\alpha(\mathbf{u}), G_\alpha(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) \\
&\qquad -2\int_{\mathcal{U}}\int_{\mathcal{U}} k(G_\alpha(\mathbf{u}), G_\beta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) \\
&\qquad +\int_{\mathcal{U}}\int_{\mathcal{U}} k(G_\beta(\mathbf{u}), G_\beta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v})\Big] \\
&= \quad -2\frac{\partial^2}{\partial\alpha\partial\beta}\int_{\mathcal{U}}\int_{\mathcal{U}} k(G_\alpha(\mathbf{u}), G_\beta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) \\
&= \quad -2\int_{\mathcal{U}}\int_{\mathcal{U}} (\nabla_\alpha G_\alpha(\mathbf{u}))^\top \nabla_1\nabla_2 k(G_\alpha(\mathbf{u}), G_\beta(\mathbf{v}))\nabla_\beta G_\beta(\mathbf{v})\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v})
\end{aligned}$$

The proof is completed by taking $\alpha = \beta$ in the expression above. □

Proof of Proposition 16

Proof. Consider the influence function obtained from the kernel scoring rule as given in Equation 5.28:

$$\mathrm{IF}_{\mathrm{MMD}}(\mathbf{z}, \mathbb{P}_\theta) \quad = \quad \left(\int_{\mathcal{X}} \nabla_\theta\nabla_\theta S_{\mathrm{MMD}}(\mathbf{x}, \mathbb{P}_\theta)\mathbb{P}_\theta(\mathrm{d}\mathbf{x})\right)^{-1} \nabla_\theta S_{\mathrm{MMD}}(\mathbf{z}, \mathbb{P}_\theta).$$

It is straightforward to show that under assumptions (i-iv) in Proposition 16, the influence function is bounded in $\mathbf{z}$, which directly implies that the estimator is bias-robust. □

Proof of Proposition 17

Proof. Consider the expression for the Langevin Stein operator KSD obtained by combining Equation 5.9 with Equation 5.10. Taking the derivative with respect to the parameters of the model, we get:

$$\begin{aligned}
\nabla_\theta L^{\text{KSD}}(\theta) &= \int_{\mathcal{X}}\int_{\mathcal{X}} \nabla_\theta k_{\mathbb{P}_\theta}(\mathbf{x},\mathbf{y})\mathbb{Q}(\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{y}) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \Big[k(\mathbf{x},\mathbf{y})\nabla_\theta\nabla_{\mathbf{x}}\log p(\mathbf{x}|\theta)\nabla_{\mathbf{y}}\log p(\mathbf{y}|\theta) \\
&\quad +k(\mathbf{x},\mathbf{y})\nabla_\theta\nabla_{\mathbf{y}}\log p(\mathbf{y}|\theta)\nabla_{\mathbf{x}}\log p(\mathbf{x}|\theta) + \big(\nabla_\theta\nabla_{\mathbf{x}}\log p(\mathbf{x}|\theta)\big)\nabla_2 k(\mathbf{x},\mathbf{y}) \\
&\quad +\nabla_\theta\nabla_{\mathbf{y}}\log p(\mathbf{y}|\theta)\nabla_1 k(\mathbf{x},\mathbf{y})\Big]\mathbb{Q}(\mathrm{d}\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{y}).
\end{aligned}$$

Let us now consider the loss function based on the MMD squared. The loss function and it's gradient are given by

$$\begin{aligned}
\nabla_\theta L^{\text{MMD}}(\theta) &= \nabla_\theta\Big[\int_{\mathcal{U}}\int_{\mathcal{U}} k(G_\theta(\mathbf{u}),G_\theta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) - 2\int_{\mathcal{U}}\int_{\mathcal{X}} k(G_\theta(\mathbf{u}),\mathbf{x})\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{Q}(\mathrm{d}\mathbf{x}) \\
&\quad + \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})\mathbb{Q}(\mathrm{d}\mathbf{x})\mathbb{Q}(\mathrm{d}\mathbf{y})\Big] \\
&= \int_{\mathcal{U}}\int_{\mathcal{U}} \nabla_\theta k(G_\theta(\mathbf{u}),G_\theta(\mathbf{v}))\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) - 2\int_{\mathcal{U}}\int_{\mathcal{X}} \nabla_\theta k(G_\theta(\mathbf{u}),\mathbf{x})\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{Q}(\mathrm{d}\mathbf{x}) \\
&= \int_{\mathcal{U}}\int_{\mathcal{U}} \nabla_\theta G_\theta(\mathbf{u})\left(\nabla_1 k(G_\theta(\mathbf{u}),G_\theta(\mathbf{v})) + \nabla_2 k(G_\theta(\mathbf{v}),G_\theta(\mathbf{u}))\right)\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{U}(\mathrm{d}\mathbf{v}) \\
&\quad -2\int_{\mathcal{U}}\int_{\mathcal{X}} \nabla_\theta G_\theta(\mathbf{u})\nabla_1 k(G_\theta(\mathbf{u}),\mathbf{y})\mathbb{U}(\mathrm{d}\mathbf{u})\mathbb{Q}(\mathrm{d}\mathbf{y})
\end{aligned}$$

□