Department of Electrical Engineering and Automation

Kernel-Based and Bayesian Methods for Numerical Integration

Toni Karvonen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 160/2019

Kernel-Based and Bayesian Methods for Numerical Integration

Toni Karvonen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall TU2 of the school on 8 November 2019 at 12.

Aalto University
School of Electrical Engineering
Department of Electrical Engineering and Automation
Sensor Informatics and Medical Technology

Supervising professor
Prof. Simo Särkkä, Aalto University, Finland

Thesis advisor
Prof. Simo Särkkä, Aalto University, Finland

Preliminary examiners
Prof. Dino Sejdinovic, University of Oxford, United Kingdom
Prof. Emtiyaz Khan, RIKEN, Japan

Opponent
Prof. Roman Garnett, Washington University in St. Louis, United States

Aalto University publication series
DOCTORAL DISSERTATIONS 160/2019

ISBN 978-952-60-8703-0 (printed)
ISBN 978-952-60-8704-7 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-8704-7

Unigrafia Oy
Helsinki 2019

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Toni Karvonen

Name of the doctoral dissertation
Kernel-Based and Bayesian Methods for Numerical Integration

Publisher School of Electrical Engineering

Unit Department of Electrical Engineering and Automation

Series Aalto University publication series DOCTORAL DISSERTATIONS 160/2019

Field of research Automation, Systems and Control Engineering

Manuscript submitted 6 May 2019 **Date of the defence** 8 November 2019

Permission for public defence granted (date) 20 August 2019 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Kernel-based methods provide a flexible toolkit for approximation of linear functionals. Importantly, these methods carry a probabilistic interpretation: a worst-case optimal method in the reproducing kernel Hilbert space induced by the kernel being used can be equivalently formulated as the posterior mean of a Gaussian process (GP) with the same covariance kernel; the worst-case error corresponds to the GP posterior standard deviation. This connection makes it possible to speak of and quantify, in a statistically principled way, uncertainty in the approximation provided by a kernel method. Consequently, these methods can be viewed as probabilistic numerical methods that interpret numerical approximation as a statistical inference problem and attempt to endow the solution of a numerical problem with a full non-degenerate posterior probability distribution. Unfortunately, both the kernel and GP formulations suffer from cubic computational and quadratic memory cost in the number of data points. Furthermore, because standard versions of kernel-based methods do not typically coincide with any "classical" method of numerical analysis in a way that would give rise to a corresponding non-degenerate GP posterior (with the exception of spline methods), there has been no straightforward way to interpret classical methods as useful statistical inference procedures within the GP framework.

This thesis studies numerical approximation of analytically intractable integrals by the means of kernel and Bayesian cubature rules, the former worst-case optimal cubature methods and the latter posteriors of GPs. The first contribution is the development of algorithms that significantly reduce the computational cost of these cubature methods by employing point sets that can be expressed as unions of fully symmetric sets. The resulting algorithm is non-approximate, computationally competitive and scalable, flexible, and algorithmically simple, enabling application of kernel and Bayesian cubature rules to integration problems involving up to millions of points. Additionally, a closed-form approximation linked to the Gauss–Hermite quadrature is proposed for the special case of one-dimensional integration against the standard Gaussian measure. The second main contribution is the study of different kernels and GP modelling choices that yield Bayesian cubature rules corresponding to classical cubature methods, such as Gaussian cubatures and uniformly weighted Monte Carlo rules. The proposed Bayes–Sard framework is general and capable of probabilistic reproduction of virtually any cubature rule.

Keywords Numerical integration, reproducing kernel Hilbert spaces, Gaussian processes, probabilistic numerics

ISBN (printed) 978-952-60-8703-0 **ISBN (pdf)** 978-952-60-8704-7

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2019

Pages 186 **urn** http://urn.fi/URN:ISBN:978-952-60-8704-7

Tekijä
Toni Karvonen

Väitöskirjan nimi
Ydinperusteiset ja bayesilaiset menetelmät numeerisessa integroinnissa

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Sähkötekniikan ja automaation laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 160/2019

Tutkimusala Automaatio, systeemit ja säätötekniikka

Käsikirjoituksen pvm 06.05.2019 **Väitöspäivä** 08.11.2019

Väittelyluvan myöntämispäivä 20.08.2019 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Ydinperusteiset menetelmät ovat joustava joukko lineaarifunktioiden approksimointia varten kehitettyjä algoritmeja. Mikä tärkeintä, näillä menetelmillä on todennäköisyysteoreettinen tulkinta: menetelmä, joka on pahimmassa tapauksessa optimaalinen reproduktiivisen ytimen Hilbertin avaruudessa, on mahdollista ilmaista täysin ekvivalentista gaussisen prosessin posteriorikeskiarvona ja sen pahimman tapauksen virhe vastaavana posteriorikeskihajontana. Tämä yhteys mahdollistaa ydinmenetelmän tuottaman approksimaation epävarmuudesta puhumisen ja mallintamisen tilastollisesti merkityksellisesti. Täten ydinmenetelmät voi nähdä probabilistisina numeerisina menetelminä, jotka käsittelevät numeerista approksimaatiota tilastollisena päättelynä ja pyrkivät varustamaan ongelman ratkaisun epädegeneroituneella posterioritodennäköisyysjakaumalla. Sekä ytimiin että gaussisiin prosesseihin perustuvat approksimaatiomenetelmät kärsivät kuutiollisesta aika- ja neliöllisestä tilavaativuudesta. Ydinperusteiset menetelmät eivät myöskään tavanomaisessa muodossaan tyypillisesti samanaikaisesti vastaa "klassisia" numeerisen analyysin menetelmiä ja epädegeneroituneita gaussisten prosessien posteriorijakaumia (splinimenetelmiä lukuunottamatta), joten klassisia menetelmiä ei ole ollut mahdollista tulkita hyödyllisinä tilastollisen päättelyn menetelminä gaussisia prosesseja käyttäen.

Tämä väitöskirja tutkii suljettua muotoa vailla olevien integraalien approksimointia käyttäen ydinperusteisia ja bayesilaisia kubatuurisääntöjä, joista edelliset ovat pahimmassa tapauksessa optimaalisia ja jälkimmäiset gaussisten prosessien posteriorijakaumia. Väitöskirjan ensimmäinen kontribuutio on laskennallisesti tehokkaiden algoritmien kehittäminen näitä kubatuurimenetelmiä varten käyttäen pistejoukkoja, jotka ovat täysin symmetristen pistejoukkojen yhdisteitä. Näin kehitetty algoritmi ei hyödynnä approksimaatioita, on laskennallisesti kilpailukykyinen, skaalautuva, joustava ja toteutukseltaan yksinkertainen, mahdollistaen ydinperusteisten ja bayesilaisten kubatuurimenetelmien soveltamisen integrointiongelmiin, joissa on käytettävä jopa miljoonia datapisteitä. Lisäki yksiulotteiseen integrointiin gaussisen mitan suhteen kehitetään erilainen suljetussa muodossa ilmaistavissa oleva approksimaatio, joka perustuu Gaussin–Hermiten kvadratuuriin. Väitöskirjan toinen pääkontribuutio liittyy erilaisten ydinten ja gaussisten prosessimallien käyttöön klassisten kubatuurisääntöjen, kuten Gaussin kubatuurien ja Monte Carlo -sääntöjen, tulkitsemisessa bayesilaisina kubatuurisääntöinä. Väitöskirjassa tätä varten kehitetyllä Bayesin–Sardin menetelmällä voi tulkita probabilistisesti lähes minkä tahansa kubatuurisäännön.

Avainsanat numeerinen integrointi, reproduktiivisen ytimen Hilbertin avaruudet, gaussiset prosessit, probabilistinen numeriikka

ISBN (painettu) 978-952-60-8703-0 **ISBN (pdf)** 978-952-60-8704-7

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2019

Sivumäärä 186 **urn** http://urn.fi/URN:ISBN:978-952-60-8704-7

Preface

This research has been carried out in the research group of sensor informatics and medical technology at the Department of Electrical Engineering and Automation at Aalto University from 2016 to 2019. During this time, my doctoral studies and work have been generously supported by the Aalto ELEC Doctoral School. I am also grateful to the Finnish Foundation for Technology Promotion, the Foundation for Aalto University Science and Technology, the Vilho, Yrjö and Kalle Väisälä Foundation, the Magnus Ehrnrooth Foundation, Oskar Öflunds Stiftelse, the Academy of Finland, the Society for Industrial and Applied Mathematics, the Neural Information Processing Systems Foundation, the International Society for Magnetic Resonance in Medicine, the Institute of Electrical and Electronics Engineers, and the Johann Radon Institute for Computational and Applied Mathematics for plentiful travel funding. The Aalto Direct service of the Aalto Learning Centre has been an indispensable tool in tracking down and accessing some of the more obscure publications.

In the beginning of my doctoral studies in early 2016 the plan was that I would continue working on stability analysis of non-linear Kalman filters, the topic of my master's thesis advised by Prof. Simo Särkkä at the Department of Biomedical Engineering and Computational Science (BECS). As may be obvious by now, this was not to be. This I would attribute to the mathematics involved being tedious and uninspiring, the results one can expect to obtain somewhat unimpressive, to me not being skillful enough a mathematician, and, most importantly, to my emerging interest, origins of which can be dated to Friday, May 29, 2015, in numerical integration and probabilistic numerics. Throughout this time, I have been supported and encouraged by my supervisor, Prof. Simo Särkkä, and I wish to express my gratitude to him for this. This thesis would not have been possible—or perhaps the process would have merely been much more agonising—without his constant support and latitude given to me to freely work on topics of my own choosing. I also thank Prof. Dino Sejdinovic and Prof. Emtiyaz Khan for pre-examining this thesis.

There are a number of people who have, at different times, made Rakentajanaukio 2C an enjoyable and inspiring environment to work at. These people include Prof. Arno Solin, Dr. Juho Kokkala, Filip Tronarp, Marko Mikkonen,

Dr. Roland Hostettler, Kimmo Suotsalo, Rui Gao, Zheng Zhao, Marco Soldati, Juha Sarmavuori, Dr. Zenith Purisha, Prof. Ángel García-Fernández, Prof. Ilkka Laakso, Dr. Lauri Palva, Dr. Sara Sommariva, Prof. Ivan Vujaklija, and Dennis Yeung.

Of my collaborators special thanks are due to Prof. Chris Oates, Jakub Prüher, Dr. Motonobu Kanagawa, Prof. Eric Moulines, and Prof. Silvère Bonnabel, all of whom have been kind enough to host me at their home institutions, some of them several times, during the past four years. The probabilistic numerics research community is still compact enough for one to meet almost everybody during the short span of a few years. The various conferences, workshops, and visits have been made much more enjoyable and productive by the presence, often recurring, of in particular Dr. Jon Cockayne and Dr. François-Xavier Briol from Warwick and London and Hans Kersting, Alexandra Gessner, Simon Bartels, Dr. Maren Mahsereci, Dr. Michael Schober, and Prof. Philipp Hennig from Tübingen. Other people with whom it has been a pleasure to work and interact with include George Wynne, Leah South, Tui Nolan, Matthew Fisher, Prof. Mark Girolami, Prof. Fred Hickernell, Prof. Takeru Matsuda, Prof. Ken'ichiro Tanaka, Prof. Yuto Miyatake, Dr. Sho Sonoda, Prof. Elodie Vernet, Prof. Tim Sullivan, Dr. Jordan Franks, and Dr. Jana de Wiljes.

Finally, it was only at Laura's suggestion that I successfully applied for a summer research assistant position at BECS in the spring of 2014 and so began the journey culminating in this thesis.

Helsinki, August 28, 2019,

Toni Karvonen

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Toni Karvonen and Simo Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720, 2018.

II Toni Karvonen, Simo Särkkä, and Chris J. Oates. Symmetry exploits for Bayesian cubature methods. Accepted for publication in *Statistics and Computing*, 2019.

III Toni Karvonen and Simo Särkkä. Gaussian kernel quadrature at scaled Gauss–Hermite nodes. *BIT Numerical Mathematics*, doi:10.1007/s10543-019-00758-3, 2019.

IV Toni Karvonen, Chris J. Oates, and Simo Särkkä. A Bayes–Sard cubature method. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 5882–5893, 2018.

V Toni Karvonen and Simo Särkkä. Classical quadrature rules via Gaussian processes. In *27th IEEE International Conference on Machine Learning for Signal Processing*, Tokyo, Japan, doi:10.1109/MLSP.2017.8168195, 2017.

Author's Contribution

Publication I: "Fully symmetric kernel quadrature"

Karvonen derived the main results and algorithms, implemented the numerical experiments and wrote the article, incorporating comments by Särkkä.

Publication II: "Symmetry exploits for Bayesian cubature methods"

Karvonen had the main responsibility in writing the article. Oates provided numerous constructive suggestions and comments and wrote the introduction. The experiments were implemented by Karvonen.

Publication III: "Gaussian kernel quadrature at scaled Gauss–Hermite nodes"

Karvonen derived the main results and algorithms, implemented the numerical experiments and wrote the article, incorporating comments by Särkkä.

Publication IV: "A Bayes–Sard cubature method"

Karvonen and Oates wrote the article and designed the experiments in close collaboration. The experiments were implemented by Karvonen.

Publication V: "Classical quadrature rules via Gaussian processes"

Karvonen derived the main results, implemented the numerical example and wrote the article.

Abbreviations

RKHS reproducing kernel Hilbert space

WCE worst-case error

GP Gaussian process

i.i.d. independent and identically distributed

WSABI warped sequential active Bayesian integration

Symbols

B_r Bernoulli polynomial of degree $r \in \mathbb{N}_0$

$C^m(\Omega)$ space of $m \in \mathbb{N}_0$ times continuously differentiable functions on domain Ω

$\mathbb{C}$ set of complex numbers; covariance of random variables

D^m mth weak derivative

d space dimension

δ_{ij} delta function ($\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$)

$\widehat{f}$ Fourier transform of function f

$\partial\Omega$ boundary of set Ω

∂_{x_i} partial derivative with respect to ith coordinate

$\partial^{\boldsymbol{\alpha}}$ multiple partial derivative

$\mathbb{E}$ expectation of a random variable

e exponential constant

$e_K(X, \boldsymbol{w})$ worst-case error of cubature rule $Q(X, \boldsymbol{w})$ in $\mathscr{H}_K$

$e^{\mathbf{av}}_{\mathscr{F},\mathbb{P}}(X, \boldsymbol{w})$ average-case error of cubature rule $Q(X, \boldsymbol{w})$

ε_i Gaussian noise variable

$\varepsilon, \delta, \beta$ constants in the Mercer eigendecomposition of the Gaussian kernel

f function of interest or integrand

$f_{\mathbf{GP}}$ Gaussian process

$\boldsymbol{f}_X$ vector of function evaluations at points X

$\boldsymbol{\Phi}_X$ alternant matrix at points X

φ_p Mercer eigenfunction of a kernel

$\varphi_{f,X}$ interpolant to function f at points X based on a set of functions $\{\varphi_i\}_{i=1}^n$

$\boldsymbol{\varphi}$ vector function formed by stacking $\{\varphi_i\}_{i=1}^Q$

$\boldsymbol{\varphi}_\mu$ vector of integrals of φ_p

$\mathrm{GP}(m, K)$ distribution of Gaussian process with mean function m and covariance kernel K

$g|_A$ restriction of function g on set A

Γ Gamma function

$\mathscr{H}$ generic reproducing kernel Hilbert space

$\mathscr{H}_K$ reproducing kernel Hilbert space of kernel K

$\mathscr{H}_{K,0}$ pre-Hilbert space of kernel K

$\mathbb{H}_r$ Hardy space on the disk of radius $r > 0$

H_p pth unnormalised probabilists' Hermite polynomial

$h_{X,\Omega}$ fill-distance of point set X

h_0 maximal fill-distance for error estimates

I integration functional

$\boldsymbol{I}_n$ n-dimensional identity matrix

$\mathbb{1}_n$ n-vector of ones

J number of fully symmetric sets

K positive-definite kernel

$K_{\boldsymbol{x}}$ kernel translate function $K(\boldsymbol{x}, \cdot)$

K_ν Matérn kernel of order $\nu > 0$

K_ν modified Bessel function of the second kind of order $\nu > 0$

K_0 Brownian motion kernel; Stein kernel; unparametrised stationary basis kernel

K_m $m \in \mathbb{N}_0$ times integrated Brownian motion kernel

K'_m released $m \in \mathbb{N}_0$ times integrated Brownian motion kernel

K_r Hardy kernel with parameter $r > 0$

K_k polyharmonic spline kernel with parameter $k \in \mathbb{N}$

$K_{d,k}$ Wendland kernel of dimension d with parameter $k \in \mathbb{N}$

K_{μ} kernel mean function for measure μ

K_X Gaussian process posterior covariance function

$K_{\sigma,\ell}$ kernel parametrised by magnitude and length-scale parameters σ and ℓ

$K^{\pi}_{X,\boldsymbol{\Sigma}}$ posterior covariance of a Gaussian process with a parametric prior mean function

K^{π}_{X} Gaussian process posterior covariance at the flat prior limit

$K^{\mathbf{pol}}_{m}$ polynomial kernel of degree m

$\boldsymbol{K}_X$ kernel matrix

$\boldsymbol{K}_{\sigma,\ell,X}$ kernel matrix at points X of a kernel parametrised by magnitude and length-scale parameters σ and ℓ

$\widetilde{K}^{\mathbf{pol}}_{m}$ standard polynomial kernel of degree $m \in \mathbb{N}_0$

$\boldsymbol{k}_{\mu,X}$ vector of kernel mean evaluations

$\boldsymbol{k}_X$ vector of kernel translates at points X

$L_{\boldsymbol{x}}$ point evaluation functional

$L^p(\Omega)$ space of p-integrable functions with respect to the Lebesgue measure on domain Ω

$L^p(\Omega, \mu)$ space of p-integrable functions with respect to measure μ on domain Ω

L_K kernel integral operator

$l(\sigma, \ell)$ marginal likelihood as a function of kernel parameters σ and ℓ

λ non-negative smoothing parameter

λ_p Mercer eigenvalue of a kernel

$\boldsymbol{\lambda}$ generator vector

$[\boldsymbol{\lambda}]$ fully symmetric set generated by $\boldsymbol{\lambda}$

$\boldsymbol{\Lambda}$ diagonal matrix containing Mercer eigenvalues of the Gaussian kernel

ℓ kernel length-scale

M^d_m dimension of Π^d_m

m_X Gaussian process posterior mean function

$\boldsymbol{m}_X$ vector of Gaussian process mean function evaluations

$m^{\pi}_{X,\boldsymbol{\Sigma}}$ posterior mean of a Gaussian process with a parametric prior mean function

m^{π}_{X} Gaussian process posterior mean at the flat prior limit

μ integration measure

$\mathbb{N}$ set of positive integers

$\mathbb{N}_0$ set of non-negative integers

$\mathbb{N}_0^d$ set of d-dimensional non-negative multi-indices

$\mathbb{N}^d$ set of d-dimensional positive multi-indices

$\mathrm{N}(\boldsymbol{m},\boldsymbol{K})$ Gaussian distribution with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{K}$

n number of points

ν smoothness parameter of a Matérn kernel

$\nabla_{\boldsymbol{x}}$ gradient with respect to $\boldsymbol{x}$

$\mathcal{O}$ big O notation (growth rate of a function)

Ω domain of interest

$\otimes$ direct product

P_X power function for points X

$\mathscr{P}_d$ collection of $d \times d$ permutation and sign-change matrices

$\boldsymbol{P}$ permutation and sign-change matrix

$\boldsymbol{P}_*\mu$ pushforward of measure μ by matrix $\boldsymbol{P}$

$p_{f,X}$ polynomial interpolant to function f at points X

p_μ density function of measure μ

$\boldsymbol{p}_\mu$ vector of integrals of polynomials

Π_d collection of permutations of the first d positive integers

Π^d_m space of d-variate polynomials of degree at most m

π pi; a linear space of real-valued functions

$Q(X,\boldsymbol{w})$ generic cubature rule with points X and weights $\boldsymbol{w}$

$Q(f;X,\boldsymbol{w})$ generic cubature approximation of function f

$Q_K(X)$ kernel cubature rule with points X

$Q_{K,\lambda}(X)$ smoothed kernel cubature rule with points X

$Q_{K,\lambda}(f;X)$ smoothed kernel cubature approximation of function f

Q number of functions in parametric prior mean function

$Q_{\mathbf{tra}}$ trapezoidal rule

q_X separation radius of point set X

$R_{X,\boldsymbol{w}}$ error representer of cubature rule $Q(X,\boldsymbol{w})$

$\mathbb{R}$ space of reals

r,s orders of Sobolev spaces

$s_{f,X}$ kernel interpolant to function f at points X

$s_{f,X,\lambda}$ smoothed kernel interpolant

s_{2m+1} natural spline interpolant of degree $2m+1$

S_d collection d-vectors with each element 1 or -1

$\boldsymbol{S}$ $J \times J$ matrix formed out of partial row sums of the kernel matrix

$\mathbb{S}^d$ d-dimensional unit sphere

$\boldsymbol{\Sigma}$ variance of random coefficients in a parametric prior mean function

σ kernel magnitude parameter

$\sigma_{\mathbf{ML}}$ maximum likelihood estimate of the kernel magnitude parameter

$T_{\mu,\boldsymbol{x}}$ operator used to define Stein kernels

$\top$ transpose of vector or matrix

θ_q random coefficient in a parametric prior mean function

$\boldsymbol{\theta}$ random coefficient vector in a parametric prior mean function

$u_{X,i}$ Lagrange cardinal function

$\boldsymbol{u}_X$ vector of Lagrange cardinal functions

$\boldsymbol{V}_X$ Vandermonde matrix at points X

$\boldsymbol{w}$ generic cubature weight vector

w_i generic cubature weight

$\boldsymbol{w}_{K,i}$ kernel cubature weight

$w_{K,j}^{\boldsymbol{\lambda}}$ weight of a fully symmetric kernel cubature rule

$w_i^{\mathbf{GH}}$ weight of the Gauss–Hermite quadrature rule

$\widetilde{w}_{K,i}$ approximate kernel quadrature weight

$\boldsymbol{w}_K$ kernel cubature weight vector

$\boldsymbol{w}_{K,\lambda}$ smoothed kernel cubature weights for points X

$\boldsymbol{w}_K^{\mathbf{BSC}}$ Bayes–Sard weights

$\boldsymbol{w}_\pi^{\mathbf{BSC}}$ Bayes–Sard weights for the basis functions in π

$\boldsymbol{w}_K^{\boldsymbol{\lambda}}$ vector of distinct weights of a fully symmetric kernel cubature rule

$\widetilde{\boldsymbol{w}}_K$ vector of approximate kernel quadrature weights

$\boldsymbol{w}^{\mathbf{GH}}$ vector of Gauss–Hermite quadrature weights

$\boldsymbol{w}_P$ weight vector of a polynomial quadrature rule

X evaluation point set

X_n evaluation point set with $n \in \mathbb{N}$ points

$x_i^{\mathbf{GH}}$ point of the Gauss–Hermite quadrature rule

$\boldsymbol{x}_i$ evaluation point

χ_A characteristic function of set A

$(x)_+$ $\max\{0, x\}$

y_i data value

$\boldsymbol{y}$ vector of data values

$\#$ number of elements in a set

$\lfloor \cdot \rfloor$ floor function, the largest integer not exceeding the argument

1. Introduction

This thesis concerns numerical approximation of intractable integrals using positive-definite kernels. Even though kernel-based methods for numerical integration have been around since the 1970s (Larkin, 1970), it has been only during the past 30 years or so that they have begun to attract considerable interest, chiefly because *kernel cubature rules*, worst-case optimal numerical integration methods in the reproducing kernel Hilbert space induced by the chosen kernel, can be interpreted as *Bayesian cubature rules*, perhaps the most prominent examples of *probabilistic numerical methods* (Hennig et al., 2015). The probabilistic interpretation, and in particular its use in modelling epistemic uncertainty inherent to numerical approximations with partial information of continuous objects, is the main motivator behind this thesis. The most compelling applications of these methods are in complicated and computationally intensive computer models that can benefit from prior information afforded by a probabilistic model and where, due to the limited computational budget, numerical uncertainties remain significant—and need to be accounted for. Recent application areas of this type include industrial (Briol et al., 2019) and cardiac (Oates et al., 2017b) models and electrical impedance tomography (Oates et al., 2019a).

Our principal aim is to present an overview of both kernel and Bayesian cubature from a *kernel-centric* point of view that is dominant in scattered data approximation (Wendland, 2005) and Gaussian process regression (Rasmussen and Williams, 2006) communities. This is in some contrast to how kernels are used in fields such as information-based complexity (Traub et al., 1988) and parts of approximation theory concerned with error analysis. In these fields kernels are often seen merely as convenient *tools* due to them inducing many classical function spaces of interest. We also make use of the opportunity and include a comprehensive list of references on the theory, methodology, and applications of kernel and Bayesian cubature rules in approximation theory, scattered data approximation, statistics, and machine learning literature.

The thesis consists of five publications and this overview. The publications are grouped thematically into two categories: Publications **I**–**III** mainly develop non-approximate computational strategies to alleviate the cubic computational cost

in the number of data points associated to kernel and Bayesian cubature while Publications **IV** and **V** discuss, among some other things, how to interpret many classical methods of numerical integration as Bayesian cubature rules. As such, the novel contributions are overwhelmingly methodological. Chapters 2 and 3 consists of a concise but fairly comprehensive overview of kernel and Bayesian cubature. Chapter 4 reviews the fully symmetric fast kernel cubature algorithms developed in Publications **I** and **II** and the Mercer expansion based explicit and numerically stable weight approximation introduced in Publication **III**. Chapter 5 reviews the technical contributions of Publications **IV** and **V** as well as two other approaches to reproducing classical methods as Bayesian cubatures. One of these is the well-known connection between splines and finitely smooth Gaussian processes (Kimeldorf and Wahba, 1970a; Lee and Wasilkowski, 1986) and the other a more recent, less well-known, and less interesting approach based on increasingly flat isotropic kernels (Driscoll and Fornberg, 2002). The publications are summarised and their significance evaluated in Chapter 6. The last chapter also briefly discusses some issues that we believe currently present the main obstacle to widespread adoption of kernel and Bayesian cubature methods.

Finally, it is a pleasure to acknowledge the influence the superb thesis of Jens Oettershagen (Oettershagen, 2017) and the recent review article by Kanagawa et al. (2018) have had on the selection for presentation and organisation of much of the material in Chapters 2 and 3.

2. Kernel Cubature

This chapter begins with a review of the theory and basic properties of positive-definite kernels and reproducing kernel Hilbert spaces they induce. Then we characterise kernel cubature rules by their worst-case optimality among all cubature rules with fixed points, discuss connections to kernel interpolation and smoothing, and review a number of important convergence results. The equivalent probabilistic interpretation via Gaussian processes of kernel cubature rules is the topic of Chapter 3. We also make an attempt at providing an exhaustive list of references to works wherein kernel cubature rules are analysed, discussed, or applied.

2.1 Reproducing Kernel Hilbert Spaces

Let Ω be a subset of $\mathbb{R}^d$ (more general domains are possible but not considered here). A *kernel* is any symmetric bivariate function $K\colon \Omega \times \Omega \to \mathbb{R}$. Unless stated otherwise, kernels in this thesis are always *positive-definite*, which means that for any $n \in \mathbb{N}$ and any distinct points $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \Omega$ the inequality

$$\sum_{i=1}^{n} \sum_{j=1}^{n} z_i z_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) > 0 \tag{2.1}$$

is satisfied for every non-zero vector $\boldsymbol{z} \in \mathbb{R}^n$. Equivalently, the *kernel matrix* $[\boldsymbol{K}_X]_{ij} \coloneqq K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive-definite. When the kernel matrix is allowed to be merely positive-semidefinite (i.e., the inequality in (2.1) is not necessarily strict), we call the kernel *positive-semidefinite*.

A Hilbert space $\mathcal{H}$ that consists of functions $f\colon \Omega \to \mathbb{R}$ and is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a norm $\|\cdot\|_{\mathcal{H}}$ is a *reproducing kernel Hilbert space* (RKHS) on Ω if the point evaluation functional $L_{\boldsymbol{x}}(f) \coloneqq f(\boldsymbol{x})$ is bounded for any $\boldsymbol{x} \in \Omega$. That is, for every $\boldsymbol{x} \in \Omega$ there exists $C_{\boldsymbol{x}} \geq 0$ such that $|L_{\boldsymbol{x}}(f)| \leq C_{\boldsymbol{x}} \|f\|_{\mathcal{H}}$ for every $f \in \mathcal{H}$. The Riesz representation theorem then implies that for every $\boldsymbol{x} \in \Omega$ there is a *reproducing kernel* $K_{\boldsymbol{x}} \coloneqq K(\boldsymbol{x}, \cdot) \in \mathcal{H}$, a *representer* of the linear functional $L_{\boldsymbol{x}}$, such that the *reproducing property* $f(\boldsymbol{x}) = L_{\boldsymbol{x}}(f) = \langle f, K_{\boldsymbol{x}} \rangle_{\mathcal{H}}$ holds for every $f \in \mathcal{H}$. As perhaps suggested by our notation, there is one-to-one corre-

spondence between reproducing kernel Hilbert spaces and positive-semidefinite kernels, summarised in the following theorem; see Aronszajn (1950, pp. 342–343) or Berlinet and Thomas-Agnan (2004, Sec. 1.3).

Theorem 2.1 (Moore–Aronszajn)**.** *If $\mathscr{H}$ is a reproducing kernel Hilbert space, then its reproducing kernel is positive-semidefinite. Conversely, every positive-semidefinite kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the reproducing kernel of a unique reproducing kernel Hilbert space on $\Omega \subset \mathbb{R}^d$, henceforth denoted $\mathscr{H}_K$.*

As mentioned earlier, we take a kernel-centric approach and accordingly always identify a reproducing kernel Hilbert space with its positive-semidefinite reproducing kernel and say that a given positive-semidefinite kernel K *induces* the space $\mathscr{H}_K$. A different perspective, common in, for example, the study of information-based complexity (Traub et al., 1988; Novak and Woźniakowski, 2008), is to start with a classical function space of interest, such as a Sobolev space, verify that it is an RKHS and find out its reproducing kernel. The origins of this approach appear to be in the work of Sard (1949).

Some basic arithmetic operations between kernels yield new kernels and RKHSs (Berlinet and Thomas-Agnan, 2004, Sec. 1.4).

Proposition 2.2 (Sums of kernels)**.** *If K_1 and K_2 are two positive-semidefinite kernels with RKHSs $\mathscr{H}_{K_1}$ and $\mathscr{H}_{K_2}$, then also $K = K_1 + K_2$ is positive-semidefinite and its RKHS is the direct sum*

$$\mathscr{H}_K = \left\{ f_1 + f_2 : f_1 \in \mathscr{H}_{K_1},\, f_2 \in \mathscr{H}_{K_2} \right\}.$$

The norm is

$$\|f\|_{\mathscr{H}_K}^2 = \min_{\substack{f = f_1 + f_2 \\ f_1 \in \mathscr{H}_{K_1},\, f_2 \in \mathscr{H}_{K_2}}} \left(\|f_1\|_{\mathscr{H}_{K_1}}^2 + \|f_2\|_{\mathscr{H}_{K_2}}^2 \right).$$

Proposition 2.3 (Products of kernels)**.** *If K_1 and K_2 are two positive-semidefinite kernels on domains $\Omega_1 \subset \mathbb{R}^{d_1}$ and $\Omega_2 \subset \mathbb{R}^{d_2}$ with RKHSs $\mathscr{H}_{K_1}$ and $\mathscr{H}_{K_2}$, then the kernel*

$$K\big((\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{x}_1', \boldsymbol{x}_2')\big) = K_1(\boldsymbol{x}_1, \boldsymbol{x}_1') K_2(\boldsymbol{x}_2, \boldsymbol{x}_2')$$

is positive-semidefinite on $\Omega = \Omega_1 \times \Omega_2 \subset \mathbb{R}^{d_1 + d_2}$. Its RKHS on Ω is the completion of the space

$$\mathscr{H}_{K,0} = \left\{ f_1 \otimes f_2 : f_1 \in \mathscr{H}_{K_1},\, f_2 \in \mathscr{H}_{K_2} \right\},$$

where $(f_1 \otimes f_2)(\boldsymbol{x}_1, \boldsymbol{x}_2) = f_1(\boldsymbol{x}_1) f_2(\boldsymbol{x}_2)$, equipped with the inner product

$$\left\langle f_1 \otimes f_2, f_1' \otimes f_2' \right\rangle_{\mathscr{H}_{K,0}} = \langle f_1, f_1' \rangle_{\mathscr{H}_{K_1}} \langle f_2, f_2' \rangle_{\mathscr{H}_{K_2}}.$$

It is often helpful to know which functions are contained in the RKHS induced by a given *positive-semidefinite* kernel and how the RKHS norm behaves. In principle, there is a simple and universal characterisation (typically used to prove Theorem 2.1): the RKHS is the completion of the pre-Hilbert space

$$\mathscr{H}_{K,0} = \left\{ \sum_{i=1}^{n} a_i K_{\boldsymbol{x}_i} : n \in \mathbb{N},\, a_i \in \mathbb{R},\, \boldsymbol{x}_i \in \Omega \right\} \tag{2.2}$$

with respect to the inner product

$$\left\langle \sum_{i=1}^{n} a_i K_{\boldsymbol{x}_i}, \sum_{j=1}^{m} b_j K_{\boldsymbol{x}'_j} \right\rangle_{\mathscr{H}_{K,0}} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(\boldsymbol{x}_i, \boldsymbol{x}'_j).$$

Unfortunately, this characterisation is typically of little help in actually determining if a given function—even a simple one, such as a polynomial—lives in the RKHS. However, from (2.2) one would expect that many properties of the kernel be inherited by the functions in its RKHS. The following inheritance results can be found in Steinwart and Christmann (2008, Sec. 4.3).

Proposition 2.4 (Boundedness and continuity)**.** *If K is positive-semidefinite, every $f \in \mathscr{H}_K$ is bounded if and only if K is bounded on Ω. Moreover, if $K_{\boldsymbol{x}} = K(\boldsymbol{x}, \cdot)$ is in addition continuous for every $\boldsymbol{x} \in \Omega$, then every $f \in \mathscr{H}_K$ is continuous.*

For the following definition, recall that a *multi-index* $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ is a non-negative integer vector. Recall also the following standard notational conventions: (i) $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^d$, $\boldsymbol{\alpha} \leq \boldsymbol{\beta}$ means that $\alpha_i \leq \beta_i$ for every $0 \leq i \leq d$, (ii) $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$, and (iii) $\boldsymbol{\alpha}! = \alpha_1! \times \cdots \times \alpha_d!$.

Definition 2.5. A positive-semidefinite kernel K on an open subset Ω of $\mathbb{R}^d$ is *m times continuously differentiable* if

$$\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} K(\boldsymbol{x}, \boldsymbol{x}') := \frac{\partial^{2|\boldsymbol{\alpha}|}}{\partial^{\boldsymbol{\alpha}} \boldsymbol{z} \, \partial^{\boldsymbol{\alpha}} \boldsymbol{z}'} K(\boldsymbol{z}, \boldsymbol{z}') \bigg|_{\substack{\boldsymbol{z}=\boldsymbol{x} \\ \boldsymbol{z}'=\boldsymbol{x}'}}$$

exists and is continuous for every multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ such that $|\boldsymbol{\alpha}| \leq m$. The kernel is *infinitely differentiable* (or *infinitely smooth*) if $\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} K$ exists for every $\boldsymbol{\alpha} \in \mathbb{N}_0^d$.

Proposition 2.6 (Differentiability)**.** *If Ω is an open subset of $\mathbb{R}^d$ and the positive-semidefinite kernel K is m-times differentiable on Ω, then every $f \in \mathscr{H}_K$ is m times continuously differentiable. Furthermore, $\partial^{\boldsymbol{\alpha}} K_{\boldsymbol{x}} \in \mathscr{H}_K$ and*

$$|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x})| := \left| \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d} f(\boldsymbol{x}) \right| = \left| \langle f, \partial^{\boldsymbol{\alpha}} K_{\boldsymbol{x}} \rangle_{\mathscr{H}_K} \right| \leq \|f\|_{\mathscr{H}_K} \sqrt{\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} K(\boldsymbol{x}, \boldsymbol{x})}$$

for every $|\boldsymbol{\alpha}| \leq m$ and $\boldsymbol{x} \in \Omega$.

Proposition 2.7 (Measurability and integrability)**.** *If $(\Omega, \mathscr{A}, \mu)$ is a measure space and K is positive-semidefinite, every $f \in \mathscr{H}_K$ is measurable if and only if $K_{\boldsymbol{x}}$ is measurable for every $\boldsymbol{x} \in \Omega$. Moreover, if $\int_\Omega K(\boldsymbol{x}, \boldsymbol{x})^{p/2} \, \mathrm{d}\mu(\boldsymbol{x}) < \infty$ for $p \geq 1$, then every $f \in \mathscr{H}_K$ is p-integrable.*

Analyticity is also inherited (Sun and Zhou, 2008, Thm. 1).

Proposition 2.8 (Analyticity)**.** *If a positive-semidefinite kernel is of the form $K(\boldsymbol{x}, \boldsymbol{x}') = \varphi(\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$ for a real analytic function φ, then every function in $\mathscr{H}_K$ is real analytic.*

The following result is a consequence of Theorem 12 in Section 4.5 of Berlinet and Thomas-Agnan (2004). Its particular implication is that constant functions are in $\mathcal{H}_K$ if and only if $K - c$ is a positive-semidefinite kernel for some constant $c > 0$.

Proposition 2.9. *If K is positive-semidefinite, a function f is in $\mathcal{H}_K$ if and only if there exists $c > 0$ such that $K(\boldsymbol{x},\boldsymbol{x}') - cf(\boldsymbol{x})f(\boldsymbol{x}')$ is a positive-semidefinite kernel.*

By the Riesz representation theorem, representers of positive linear functionals are in the RKHS (e.g., Muandet et al., 2017, Lem. 3.1).

Proposition 2.10. *Let K be positive-semidefinite and $L\colon \mathcal{H}_K \to \mathbb{R}$ a positive linear functional. The representer $K_L(\boldsymbol{x}) \coloneqq L(K(\boldsymbol{x},\cdot))$ of L is in $\mathcal{H}_K$ and $L(f) = \langle f, K_L\rangle_{\mathcal{H}_K}$ for any $f \in \mathcal{H}_K$ if $L(g) < \infty$ for the function $g(\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{x})$.*

Unfortunately, these properties alone are rarely enough to completely characterise the functions lying in an RKHS or to provide sufficient insight into the structure of the RKHS norm and inner product. Next we review characterisations of RKHSs induced by a variety of popular kernels.

2.2 Examples of Kernels and RKHSs

This section introduces most of the kernels that appear in the remainder of this thesis and characterises the RKHSs they induce. It is useful to introduce some standard terminology: a kernel is *stationary* if $K(\boldsymbol{x},\boldsymbol{x}')$ depends only on $\boldsymbol{x} - \boldsymbol{x}'$ and *isotropic*[1] if the dependency is on $\|\boldsymbol{x} - \boldsymbol{x}'\|$. Translates of some of the kernels reviewed in this section are depicted in Figure 2.1. We occasionally engage in slight notational abuse and use $K(\boldsymbol{x} - \boldsymbol{x}')$ or $K(\|\boldsymbol{x} - \boldsymbol{x}'\|)$ in the place of $K(\boldsymbol{x},\boldsymbol{x}')$ if K is stationary or isotropic, respectively.

2.2.1 Matérn and Sobolev Kernels

A *Matérn kernel* with smoothness parameter $\nu > 0$ and length-scale $\ell > 0$ is

$$K_\nu(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell}\right)^{\nu} \mathrm{K}_\nu\left(\frac{\sqrt{2\nu}\,\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell}\right), \tag{2.3}$$

where Γ is the Gamma function and K_ν the modified Bessel function of the second kind of order ν.[2] These kernels originate in Matérn (1960, Sec. 2.4) and have been extensively covered by Stein (1999). For half-integers values of the smoothness parameter, $\nu = m + 1/2$ for $m \in \mathbb{N}_0$, the Matérn kernel has the

[1] In which case it is often called a *radial basis function*.

[2] Note that dependence of the kernel on ℓ is suppressed in the notation. In general, here and later on only dependencies on parameters relevant to the immediate discussion and results are made explicit. For example, dependence on ℓ is explicitly denoted in Sections 3.3 and 5.4.

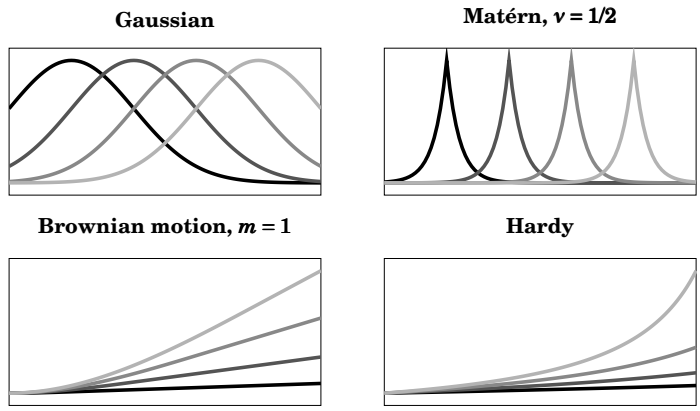

Figure 2.1. Four translates on the interval $[0,1]$ of the (i) Gaussian kernel (2.4) with $\ell = 0.2$, (ii) Matérn kernel (2.3) with $\nu = 1/2$ and $\ell = 0.2$, (iii) Brownian motion kernel (2.7) with $m = 1$, and (iv) Hardy kernel (2.11) with $r = 1$. The Brownian motion kernel and the Hardy kernel are non-stationary.

expression

$$K_\nu(\boldsymbol{x},\boldsymbol{x}') = \frac{m!}{(2m)!}\exp\left(-\frac{\sqrt{2\nu}\,\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)\sum_{p=0}^{m}\frac{2^{m-p}(m+p)!}{p!(m-p)!}\left(\frac{\sqrt{2\nu}\,\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)^{m-p}$$

in terms of elementary functions. Most commonly used special cases are

$$K_{1/2}(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right),$$

$$K_{3/2}(\boldsymbol{x},\boldsymbol{x}') = \left(1+\frac{\sqrt{3}\,\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)\exp\left(-\frac{\sqrt{3}\,\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right),$$

$$K_{5/2}(\boldsymbol{x},\boldsymbol{x}') = \left(1+\frac{\sqrt{5}\,\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}+\frac{5\,\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}\,\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right),$$

the first of which often goes by the name *exponential kernel*.

From the asymptotic behaviour of the modified Bessel function K_ν at the origin it can be concluded that a Matérn kernel of order $\nu > m$ for $m \in \mathbb{N}_0$ is m times continuously differentiable (Stein, 1999, p. 32) in the sense of Definition 2.5. The corresponding RKHS thus consists of m times differentiable functions. A more complete characterisation of the RKHS is that it is norm-equivalent to a Sobolev space (Wendland, 2005, Ch. 10).

Definition 2.11 (Norm-equivalence)**.** Two normed vector spaces $\mathscr{F}_1$ and $\mathscr{F}_2$ are *norm-equivalent* if they are identical as sets and there exist constants $C_1, C_2 > 0$ such that

$$C_1\|f\|_{\mathscr{F}_1} \leq \|f\|_{\mathscr{F}_2} \leq C_2\|f\|_{\mathscr{F}_1}$$

for every $f \in \mathscr{F}_1 = \mathscr{F}_2$.

Theorem 2.12 (RKHSs of Matérns)**.** *If $\nu = r - d/2$ for $r > d/2$, the RKHS $\mathcal{H}_{K_\nu}$ induced by the Matérn kernel* (2.3) *of order ν is norm-equivalent to the Sobolev space $H^r(\Omega)$.*

For a concise treatment of Sobolev spaces in this context, see for example Kanagawa et al. (2019, Sec. 2.2). In short, $H^r(\Omega)$ is the restriction onto Ω of the space

$$H^r(\mathbb{R}^d) := \left\{ f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\xi}\|^2\right)^r \left|\widehat{f}(\boldsymbol{\xi})\right|^2 \mathrm{d}\boldsymbol{\xi} < \infty \right\}$$

equipped with the inner product

$$\langle f, g \rangle_{H^r(\mathbb{R}^d)} := \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\xi}\|^2\right)^r \widehat{f}(\boldsymbol{\xi}) \overline{\widehat{g}(\boldsymbol{\xi})} \, \mathrm{d}\boldsymbol{\xi}.$$

Here

$$\widehat{f}(\boldsymbol{\xi}) := \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{x}^\mathsf{T} \boldsymbol{\xi}} \, \mathrm{d}\boldsymbol{x}$$

is the Fourier transform of $f \in L^1(\mathbb{R}^d)$. If $r \in \mathbb{N}_0$ and Ω has a Lipschitz boundary, then $H^r(\Omega)$ can be (up to equivalent norms) defined as the collection of functions whose *weak* derivatives up to order r exist and are in $L^2(\Omega)$; the inner product takes the form of a sum of $L^2(\Omega)$ inner products between weak derivatives. That the boundary $\partial\Omega$ of Ω is Lipschitz essentially means that the boundary is sufficiently regular. For example, there being a continuously differentiable embedding of $\partial\Omega$ in $\mathbb{R}^d$ suffices; see Stein (1970, Sec. 3.3) for a formal definition and a larger collection of examples. Every domain discussed in this thesis has a Lipschitz boundary. For later purposes, it is convenient to define a general class of kernels akin to Matérns in that their RKHSs are Sobolev spaces.

Definition 2.13 (Sobolev kernel)**.** A stationary kernel is a *Sobolev kernel*[3] *of order $r > 0$* if its RKHS is norm-equivalent to the Sobolev space $H^r(\Omega)$.

2.2.2 Gaussian Kernel

For a length-scale parameter $\ell > 0$, the famous *Gaussian kernel*[4] is

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2} \right). \tag{2.4}$$

The particular parametrisation of the Matérn kernels (2.3) implies that the Gaussian kernel can be obtained as "an infinitely differentiable limit": $\lim_{\nu \to \infty} K_\nu(\boldsymbol{x}, \boldsymbol{x}') = K(\boldsymbol{x}, \boldsymbol{x}')$ (Stein, 1999, p. 50). The properties listed in Section 2.1 tell us only that the induced RKHS consists of bounded and infinitely smooth analytic functions. A more illustrative characterisation is due to Steinwart et al. (2006) and Minh (2010, Thm. 1); see also Steinwart and Christmann (2008, Sec. 4.4) and De Marchi and Schaback (2009, Ex. 3).

[3] This term is occasionally attached to certain specific kernels (Fasshauer and McCourt, 2015, p. 43).

[4] The kernel also goes under the names *squared exponential*, *exponentiated quadratic* and *radial basis function kernel*.

Theorem 2.14 (RKHSs of Gaussians)**.** *Suppose that Ω has a non-empty interior. Then the RKHS $\mathcal{H}_K$ induced by the Gaussian kernel* (2.4) *consists of the functions*

$$f(\boldsymbol{x}) = \mathrm{e}^{-\|\boldsymbol{x}\|^2/(2\ell^2)} \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} f_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \quad \text{such that} \quad \|f\|_{\mathcal{H}_K}^2 = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} \ell^{2|\boldsymbol{\alpha}|} \boldsymbol{\alpha}! f_{\boldsymbol{\alpha}}^2 < \infty. \tag{2.5}$$

The inner product is

$$\langle f, g\rangle_{\mathcal{H}_K} = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^d} \ell^{2|\boldsymbol{\alpha}|} \boldsymbol{\alpha}! f_{\boldsymbol{\alpha}} g_{\boldsymbol{\alpha}},$$

where g is defined via a series analogous to that defining f.

This RKHS is not a large one; the coefficients $f_{\boldsymbol{\alpha}}$ essentially have to decay at least like $(\ell^{|\boldsymbol{\alpha}|}\sqrt{\boldsymbol{\alpha}!})^{-1/2}$. For example, no polynomial (besides $f \equiv 0$) or a function $f(\boldsymbol{x}) = \mathrm{e}^{-\rho\|\boldsymbol{x}\|^2/(2\ell^2)}$ for $\rho \geq 2$ belongs to $\mathcal{H}_K$ (Minh, 2010, Thms. 2 and 3). Naturally, every exponentially damped polynomial $\mathrm{e}^{-\|\boldsymbol{x}\|^2/(2\ell^2)} \sum_{\boldsymbol{\alpha}\in\mathscr{I}} f_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$ defined by a finite multi-index set $\mathscr{I} \subset \mathbb{N}_0^d$ is in $\mathcal{H}_K$ because the sum in (2.5) terminates. As a slightly more complicated example, consider the univariate function

$$f(x) = \mathrm{e}^{-x^2/(2\ell^2)} \sin(x) = \mathrm{e}^{-x^2/(2\ell^2)} \sum_{\alpha=0}^{\infty} \frac{(-1)^\alpha}{(2\alpha+1)!} x^{2\alpha+1}.$$

This function has coefficients $f_{2\alpha+1}^2 = ((2\alpha+1)!)^{-2}$. Thus

$$\|f\|_{\mathcal{H}_K}^2 = \sum_{\alpha=0}^{\infty} \ell^{2(2\alpha+1)} (2\alpha+1)! \frac{1}{((2\alpha+1)!)^2} = \sum_{\alpha=0}^{\infty} \frac{\ell^{2(2\alpha+1)}}{(2\alpha+1)!} = \sinh(\ell^2) < \infty.$$

The Gaussian RKHS can also be characterised as a space of functions with Fourier transforms that decay sufficiently fast (Kanagawa et al., 2018, Ex. 2.7).

2.2.3 Brownian Motion Kernels

Let $d = 1$ and $\Omega = [0,1] \subset \mathbb{R}$. The *Brownian motion kernel* is

$$K_0(x,x') = \min\{x,x'\}. \tag{2.6}$$

For $m \in \mathbb{N}$, the *m times integrated Brownian motion kernel* is

$$K_m(x,x') = \int_0^x \int_0^{x'} K_{m-1}(z,z')\,\mathrm{d}z\,\mathrm{d}z' = \int_0^1 \frac{(x-t)_+^m (x'-t)_+^m}{(m!)^2}\,\mathrm{d}t, \tag{2.7}$$

where $(x)_+ = \max\{0,x\}$ and the last expression is valid also for $m = 0$ if the convention $0^0 = 0$ is used. The kernel K_0 describes the covariance structure of the standard Brownian motion and K_m for $m \geq 1$ that of the m times integrated Brownian motion (Wahba, 1990, Sec. 1.5). For example, explicit expressions for K_1 and K_2 are (e.g., Schober et al., 2014, Secs. 3.2 & 3.3)

$$K_1(x,x') = \frac{1}{3}\min\{x,x'\}^3 + \frac{1}{2}|x-x'|\min\{x,x'\}^2,$$

$$K_2(x,x') = \frac{1}{20}\min\{x,x'\}^5 + \frac{1}{12}|x-x'|\left((x+x')\min\{x,x'\}^3 - \frac{1}{2}\min\{x,x'\}^4\right).$$

The RKHSs induced by Brownian motion kernels are characterised in Theorem 2.15. Derivation of this result can be found in Adler and Taylor (2007, Sec. 3.1) for $m = 0$ and in van der Vaart and van Zanten (2008, Sec. 10) for general $m \in \mathbb{N}$. See also Wahba (1990, Sec. 1.2).

Theorem 2.15 (RKHSs of Brownian motion kernels)**.** *The RKHS $\mathcal{H}_{K_m}$ of the $m \in \mathbb{N}_0$ times integrated Brownian motion kernel* (2.7) *consists of functions $f \in C^m([0,1])$ such that (i) $f^{(q)}(0) = 0$ for $q = 0, \dots, m$, (ii) $f^{(m)}$ is absolutely continuous, and (iii) the weak derivative $\mathrm{D}^{m+1} f$ is in $L^2([0,1])$. The RKHS inner product is*

$$\langle f, g \rangle_{\mathcal{H}_{K_m}} = \left\langle \mathrm{D}^{m+1} f, \mathrm{D}^{m+1} g \right\rangle_{L^2([0,1])}.$$

The restriction that the functions and their first m derivatives vanish at the origin can be relaxed and the RKHS made into the full Sobolev space $H^{m+1}([0,1])$ by using the *released* Brownian motion kernel

$$K'_m(x, x') = \sum_{q=0}^{m} \frac{(xx')^q}{(q!)^2} + K_m(x, x'). \tag{2.8}$$

The inner product then becomes

$$\langle f, g \rangle_{\mathcal{H}_{K'_m}} = \sum_{q=0}^{m} f^{(q)}(0) g^{(q)}(0) + \left\langle \mathrm{D}^{m+1} f, \mathrm{D}^{m+1} g \right\rangle_{L^2([0,1])}. \tag{2.9}$$

All these results on Brownian motion kernels are related to the integral form

$$f(x) = \sum_{q=0}^{m} \frac{x^q}{q!} f^{(q)}(0) + \int_0^1 \frac{(x-t)_+^m}{m!} f^{(m+1)}(t) \, \mathrm{d}t$$

of Taylor's theorem, from which the identity (for $x \neq x'$)

$$\partial_{x'}^{m+1} K_m(x, x') = \frac{(x - x')_+^m}{m!} \tag{2.10}$$

can be derived by selecting $f(x) = K_m(x, x')$. Inserting (2.10) into the inner product (2.9) gives

$$\left\langle f, K'_m(x, \cdot) \right\rangle_{\mathcal{H}_{K'_m}} = \sum_{q=0}^{m} \frac{x^q}{q!} f^{(q)}(0) + \int_0^1 \frac{(x-t)_+^m}{m!} f^{(m+1)}(t) \, \mathrm{d}t = f(x),$$

which is the reproducing property.

2.2.4 Other Kernels

There is a large number of kernels other than the ones reviewed above that are used in kernel and Bayesian cubature. Note that some of the kernels listed below are only *conditionally positive-definite* (of order $m \in \mathbb{N}_0$), which is to say that (2.1) needs to hold only for all $\boldsymbol{z} \in \mathbb{R}^n$ such that $\sum_{i=1}^n z_i p(z_i) = 0$ for every polynomial p of degree less than m (Wendland, 2005, Ch. 8). This relaxation introduces some complications in the RKHS construction for which we refer the reader to Wendland (2005, Sec. 10.3).

Hardy Kernel

The RKHS of the *Hardy kernel*

$$K_r(x,x') = \frac{r^2}{r^2 + xx'} = \sum_{p=0}^{\infty} r^{-2p}(xx')^p \tag{2.11}$$

consists of functions whose complex extensions are in the *Hardy space* $\mathbb{H}_r$, a Hilbert space of functions analytic on the disc $\{z \in \mathbb{C} : |z| < r\}$ equipped with the inner product

$$\langle f, g \rangle_{\mathbb{H}_r} = \frac{1}{2\pi} \int_0^{2\pi} f(r\,\mathrm{e}^{\mathrm{i}\theta}) \overline{g(r\,\mathrm{e}^{\mathrm{i}\theta})}\,\mathrm{d}\theta.$$

See Zwicknagl and Schaback (2013, Sec. 5.1) and Oettershagen (2017, Sec. 3.6.2) for more details. This kernel has been used in kernel quadrature by Larkin (1970, Sec. 3), Minka (2000, Sec. 4), and Oettershagen (2017, Sec. 6.2), who has undertaken an extensive numerical study on optimal and greedy selection of the integration points. Other related kernels have been analysed in the context of quadrature in Richter (1970) and Richter-Dyn (1971b,a). The Hardy kernel is a member of the general class of Taylor space kernels (Dick, 2006; Zwicknagl and Schaback, 2013).

Multiquadrics and Inverse Multiquadrics

Multiquadric and inverse multiquadric kernels, popular in scattered data approximation literature, have been used by Sommariva and Vianello (2006b) for integration on $[0,1]^2 \subset \mathbb{R}^2$. These infinitely smooth kernels are given by

$$K(\boldsymbol{x},\boldsymbol{x}') = \left(1 + \frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\ell^2}\right)^{1/2} \qquad \text{and} \qquad K(\boldsymbol{x},\boldsymbol{x}') = \left(1 + \frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\ell^2}\right)^{-1/2},$$

respectively, where $\ell > 0$ is a length-scale parameter. Multiquadrics are only conditionally positive-definite.

Polyharmonic Splines

Numerical integration with conditionally positive-definite *polyharmonic spline kernels*

$$K_k(\boldsymbol{x},\boldsymbol{x}') = \begin{cases} \left(\dfrac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)^k & \text{if } k \in \mathbb{N} \text{ is odd,} \\ \left(\dfrac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)^k \log\left(\dfrac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right) & \text{if } k \in \mathbb{N} \text{ is even} \end{cases}$$

has been analysed by Bezhaev (1991). Numerical results for the special case $k = 2$, called *thin plate spline kernel*, appear in Sommariva and Vianello (2006b,a); Punzi et al. (2008) and Fuselier et al. (2014) for different domains, including the sphere. The RKHS of a thin plate spline is a Beppo-Levi space (Wendland, 2005, Thm. 10.43).

Wendland Kernels

The *Wendland kernel*

$$K_{d,k}(\boldsymbol{x},\boldsymbol{x}') = \varphi_{d,k}\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right) \quad \text{for} \quad \varphi_{d,k}(r) = \frac{1}{\Gamma(k)2^{k-1}}\int_r^1 t(1-t)^l(t^2-r^2)^{k-1}\,\mathrm{d}t,$$

where k is such that $2k \in \mathbb{N}$,

$$l = \left\lfloor \frac{d}{2} + k \right\rfloor + 1,$$

and the convention $\varphi_{d,k}(r) = 0$ if $r > 1$ is used, has been applied to numerical integration by Sommariva and Vianello (2006b) with $(k,l) = (1,3)$. These kernels are similar to Matérn kernels (2.3) in that $K_{d,k}$ is a Sobolev kernel of order $d/2 + k + 1/2$ (Wendland, 2005, Thm. 10.35) and that when appropriate scaling of the variables is used, it can be shown that the Wendland kernel converges to the Gaussian as k, a smoothness parameter, is increased (Chernih et al., 2014).

Distance Kernels on the Sphere

The *distance kernel* $K(\boldsymbol{x},\boldsymbol{x}') = 8/3 - \|\boldsymbol{x}-\boldsymbol{x}'\|$ induces the Sobolev space $H^{3/2}(\mathbb{S}^2)$ of functions defined on the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ and is a member of a larger class of kernels that have Sobolev spaces on spheres as their RKHSs (Brauchart and Dick, 2012; Brauchart et al., 2014). This kernel has been used in an application of Bayesian cubature to a computer graphics problem in Briol et al. (2019) and Xi et al. (2018) and in Publication **II**.

Shift-Invariant Kernels

Shift-invariant kernels of the form

$$K(\boldsymbol{x},\boldsymbol{x}') = \prod_{i=1}^{d}\left[1 - (-1)^r\gamma \mathrm{B}_{2r}(|x_i - x_i'|)\right], \qquad \boldsymbol{x},\boldsymbol{x}' \in [0,1]^d,$$

where $r \in \mathbb{N}$ and $\gamma > 0$ are parameters and B_{2r} are the even-degree Bernoulli polynomials, have been exploited by Rathinavel and Hickernell (2018) to reduce the computational complexity in n, the number of evaluation points, of Bayesian cubature and associated kernel scale parameter selection from $\mathcal{O}(n^3)$ to $\mathcal{O}(n \log n)$. Kernels of this type are useful in the error analysis of quasi Monte Carlo rules (Hickernell, 1998).

Stein Kernels

A recurring challenge in kernel cubature is that integrals $\int_\Omega K(\boldsymbol{x},\boldsymbol{x}')\,\mathrm{d}\mu(\boldsymbol{x}')$, with μ a potentially complicated measure, need to be computed. *Stein kernels* are a particular solution to this problem. One way to define a Stein kernel is as follows. First, suppose that μ is probability measure that admits a Lebesgue density function p_μ and define the operator

$$T_{\mu,\boldsymbol{x}} f(\boldsymbol{x}) := \frac{\nabla_{\boldsymbol{x}} \cdot [p_\mu(\boldsymbol{x})\nabla_{\boldsymbol{x}} f(\boldsymbol{x})]}{p_\mu(\boldsymbol{x})},$$

acting on twice differentiable functions f. Here $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = (\partial_{x_1} f(\boldsymbol{x}) \cdots \partial_{x_d} f(\boldsymbol{x})) \in \mathbb{R}^d$. Given a twice differentiable positive-definite kernel K, a Stein kernel is then

$$K_0(\boldsymbol{x},\boldsymbol{x}') = T_{\mu,\boldsymbol{x}} T_{\mu,\boldsymbol{x}'} K(\boldsymbol{x},\boldsymbol{x}').$$

If the density p_μ vanishes sufficiently fast at the boundary of Ω, the divergence theorem ensures that $\int_\Omega K_0(\boldsymbol{x},\boldsymbol{x}') p_\mu(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}' = 0$ for every $\boldsymbol{x} \in \Omega$ (Oates et al., 2017a). Barp et al. (2018) construct kernel cubature rules using Stein kernels.

2.3 Optimal Cubature Rules in RKHSs

Let $\Omega \subset \mathbb{R}^d$ be Borel measurable, μ a finite Borel measure on Ω, and $f : \Omega \to \mathbb{R}$, the *integrand*, a μ-integrable function. A *cubature rule* (or, in one dimension, *quadrature rule*) $Q(X,\boldsymbol{w})$ approximates the integral of f using a weighted sum of function evaluations:

$$Q(f;X,\boldsymbol{w}) := \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i) \approx I(f) := \int_\Omega f \,\mathrm{d}\mu \tag{2.12}$$

for some *points* (or *nodes*) $X = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\} \subset \Omega$, assumed distinct throughout this thesis, and *weights* $\boldsymbol{w} = (w_1,\ldots,w_n) \in \mathbb{R}^n$. The challenge is to design the points and weights such that the approximation (2.12) is, by some relevant criterion, "good", "best", or "optimal". This section defines and reviews the most important properties of kernel cubature rules that are worst-case optimal in reproducing kernel Hilbert spaces. For a more complete and general modern review that includes proofs, see Oettershagen (2017, Ch. 3). A Bayesian probabilistic interpretation of kernel cubature rules is introduced in Chapter 3.

2.3.1 Worst-Case Error

The *worst-case error* (WCE) in an RKHS $\mathcal{H}_K$ of a cubature rule with points X and weights $\boldsymbol{w}$ is

$$e_K(X,\boldsymbol{w}) := \sup_{\|f\|_{\mathcal{H}_K} \leq 1} |I(f) - Q(f;X,\boldsymbol{w})|.$$

Such a quantity could be of course defined in any normed function space, but computing it is typically possible only if the space is an RKHS. For this purpose, define the *kernel mean* function

$$K_\mu := \int_\Omega K(\cdot,\boldsymbol{x}) \,\mathrm{d}\mu(\boldsymbol{x}).$$

That is, kernel mean at $\boldsymbol{x}$ is the integral of the corresponding point representer: $K_\mu(\boldsymbol{x}) = I(K_{\boldsymbol{x}})$. Practical computation of kernel means is an issue that will be discussed more in Section 6.2.2. If $\int_\Omega K(\boldsymbol{x},\boldsymbol{x}) \mathrm{d}\mu(\boldsymbol{x}) < \infty$, then Proposition 2.10 implies that $K_\mu \in \mathcal{H}_K$ and $I(f) = \langle f, K_\mu \rangle_{\mathcal{H}_K}$. Note that this condition is satisfied by all stationary kernels since for them $K(\boldsymbol{x},\boldsymbol{x})$ does not depend on $\boldsymbol{x}$

and consequently $\int_\Omega K(\boldsymbol{x},\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}) = K(\boldsymbol{0})\mu(\Omega)$. The following result provides a closed-form expression for the worst-case error in an RKHS. For its proof, see for instance Oettershagen (2017, Sec. 3.2).

Proposition 2.16. *The worst-case error* (2.14) *can be expressed as*

$$\begin{aligned} e_K(X,\boldsymbol{w})^2 &= \left\| K_\mu - \sum_{i=1}^n w_i K_{\boldsymbol{x}_i} \right\|_{\mathscr{H}_K}^2 \\ &= \int_\Omega K(\boldsymbol{x},\boldsymbol{x}')\,\mathrm{d}\mu(\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}') - 2\sum_{i=1}^n w_i K_\mu(\boldsymbol{x}_i) + \sum_{i,j=1}^n w_i w_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ &= I(K_\mu) - 2\boldsymbol{w}^\top \boldsymbol{k}_{\mu,X} + \boldsymbol{w}^\top \boldsymbol{K}_X \boldsymbol{w}, \end{aligned}$$

where $[\boldsymbol{K}_X]_{ij} = K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ *is the* $n \times n$ *positive-definite kernel matrix and* $[\boldsymbol{k}_{\mu,X}]_i = K_\mu(\boldsymbol{x}_i)$ *is an* n*-vector.*

The first expression in Proposition 2.16 gives the worst-case error in terms of the RKHS norm of the *error representer* $R_{X,\boldsymbol{w}} \coloneqq K_\mu - \sum_{i=1}^n w_i K_{\boldsymbol{x}_i}$. This terminology derives from the fact that the reproducing property gives

$$|I(f) - Q(f;X,\boldsymbol{w})| = |\langle f, R_{X,\boldsymbol{w}}\rangle_{\mathscr{H}_K}|.$$

The Cauchy–Schwarz inequality then yields

$$|I(f) - Q(f;X,\boldsymbol{w})| \leq \|f\|_{\mathscr{H}_K} \|R_{X,\boldsymbol{w}}\|_{\mathscr{H}_K} = \|f\|_{\mathscr{H}_K}\, e_K(X,\boldsymbol{w}). \tag{2.13}$$

Inequalities of this form are fundamental tools in error and convergence analysis, including that presented in Section 2.3.4, of many types of cubature rules since they effectively decouple the properties of the integrand from those of the cubature rule.[5]

Alternatively, one could consider the *average-case error* by placing a probability distribution $\mathbb{P}$ over some separable Banach space $\mathscr{F}$ of functions:

$$e^{\mathrm{av}}_{\mathscr{F},\mathbb{P}}(X,\boldsymbol{w}) \coloneqq \left(\int_{\mathscr{F}} [I(f) - Q(f;X,\boldsymbol{w})]^2\,\mathrm{d}\mathbb{P}(f) \right)^{1/2}.$$

The standard reference on average-case analysis is the monograph of Ritter (2000). Crucially, if the distribution $\mathbb{P}$ is zero-mean Gaussian with the covariance kernel

$$K_{\mathbb{P}}(\boldsymbol{x},\boldsymbol{x}') = \int_{\mathscr{F}} f(\boldsymbol{x}) f(\boldsymbol{x}')\,\mathrm{d}\mathbb{P}(f),$$

then the notions of worst-case error in the RKHS of $K_{\mathbb{P}}$ and average-case error coincide (Novak and Woźniakowski, 2010, Sec. 13.4):

$$e_{K_{\mathbb{P}}}(X,\boldsymbol{w}) = e^{\mathrm{av}}_{\mathscr{F},\mathbb{P}}(X,\boldsymbol{w}).$$

[5] For instance, in quasi Monte Carlo literature a similar inequality is termed the *Koksma–Hlawka inequality* (Dick et al., 2013, Sec. 3.4).

2.3.2 Kernel Cubature and Interpolation

The *kernel cubature rule* (or *kernel-based cubature rule*) with points X, denoted $Q_K(X) := Q(X, \boldsymbol{w}_K)$, is the cubature rule with minimal worst-case error among all possible cubature rules using these points. That is,

$$e_K(X, \boldsymbol{w}_K) = \inf_{\boldsymbol{w} \in \mathbb{R}^n} e_K(X, \boldsymbol{w}) = \inf_{\boldsymbol{w} \in \mathbb{R}^n} \sup_{\|f\|_{\mathcal{H}_K} \leq 1} |I(f) - Q(f; X, \boldsymbol{w})|. \tag{2.14}$$

From Proposition 2.16 it easy to see that the kernel cubature weights $\boldsymbol{w}_K$ are unique and given by $\boldsymbol{w}_K = \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\mu,X}$. Consequently, the kernel cubature approximation of $I(f)$ is

$$Q_K(f; X) = Q(f; X, \boldsymbol{w}_K) = \sum_{i=1}^{n} w_{K,i} f(\boldsymbol{x}_i) = \boldsymbol{f}_X^\top \boldsymbol{w}_K = \boldsymbol{f}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\mu,X}, \tag{2.15}$$

where $\boldsymbol{f}_X = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) \in \mathbb{R}^n$ is the vector of integrand evaluations. It also follows from the expression for the weights that

$$e_K(X, \boldsymbol{w}_K)^2 = I(K_\mu) - \boldsymbol{w}_K^\top \boldsymbol{k}_{\mu,X} = I(K_\mu) - Q_K(K_\mu; X),$$

which means that the squared worst-case error of the kernel cubature rule is equal to the cubature approximation error for the kernel mean function. Because the kernel cubature weights solve the linear system of equations $\boldsymbol{K}_X \boldsymbol{w}_K = \boldsymbol{k}_{\mu,X}$, ith row of which is

$$Q_K(K_{\boldsymbol{x}_i}; X) = \sum_{j=1}^{n} w_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K_\mu(\boldsymbol{x}_i) = I(K_{\boldsymbol{x}_i}),$$

we see that the kernel cubature rule is completely determined by the fact that it integrates exactly each of the n kernel translates $K_{\boldsymbol{x}_i}$ at the cubature points.

It is often instructive to think of kernel cubature rules in terms of interpolation. The *kernel interpolant*[6] $s_{f,X}$ is the unique function in the span of the kernel translates $\{K_{\boldsymbol{x}_i}\}_{i=1}^n$ that interpolates f at points X:

$$s_{f,X} = \sum_{i=1}^{n} c_i K_{\boldsymbol{x}_i}, \tag{2.16}$$

for coefficients $\boldsymbol{c} = (c_1, \ldots, c_n)$ selected so that the interpolation condition $s_{f,X}|_X = f|_X$ is satisfied. Here $g|_A$ denotes the restriction on set A of a function g, so that $s_{f,X}|_X = f|_X$ means that $s_{f,X}(\boldsymbol{x}) = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in X$. The interpolation property implies that $\boldsymbol{c} = \boldsymbol{K}_X^{-1} \boldsymbol{f}_X$. Defining the vector function $\boldsymbol{k}_X = (K_{\boldsymbol{x}_1}, \ldots, K_{\boldsymbol{x}_n})$, the kernel interpolant is thus

$$s_{f,X} = \boldsymbol{c}^\top \boldsymbol{k}_X = \boldsymbol{f}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X = \boldsymbol{f}_X^\top \boldsymbol{u}_X, \tag{2.17}$$

[6] Also known as *spline interpolant* (Oettershagen, 2017, Def. 3.9) and, if the kernel is isotropic, *radial basis function interpolant* (Fasshauer, 2007, Ch. 2).

where $\boldsymbol{u}_X = (u_{X,1},\dots,u_{X,n}) = \boldsymbol{K}_X^{-1}\boldsymbol{k}_X$ are called *Lagrange cardinal functions*. They satisfy the cardinality property $u_{X,i}(\boldsymbol{x}_j) = \delta_{ij}$. An equivalent definition of the kernel interpolant is that its RKHS norm is minimal among all functions $g \in \mathscr{H}_K$ such that $g|_X = f|_X$ (Fasshauer and McCourt, 2015, Sec. 9.1):

$$s_{f,X} = \arg\min\left\{ \|g\|_{\mathscr{H}_K} : g \in \mathscr{H}_K \text{ and } g|_X = f|_X \right\}. \tag{2.18}$$

Since $I(\boldsymbol{k}_X) = \boldsymbol{k}_{\mu,X}$, comparison to (2.15) reveals that the kernel cubature approximation is obtained by integrating the kernel interpolant:

$$Q_K(f;X) = I(s_{f,X}). \tag{2.19}$$

Moreover, the weights are integrals of the Lagrange cardinal functions, $w_{K,i} = I(u_{X,i})$. In Section 2.3.4, the relation (2.19) makes it possible to leverage results on convergence of kernel interpolants in error analysis of kernel cubature rules.

In interpolation, a concept roughly analogous to the worst-case error (2.14) is that of the *power function* P_X (Schaback, 1993). At point $\boldsymbol{x} \in \Omega$, this function is defined as the RKHS norm of the error in interpolating $K_{\boldsymbol{x}}$ with the kernel interpolant based on points X:

$$\begin{aligned} P_X(\boldsymbol{x}) &:= \left\| K_{\boldsymbol{x}} - s_{K_{\boldsymbol{x}},X} \right\|_{\mathscr{H}_K} = \left\| K(\boldsymbol{x},\cdot) - \boldsymbol{k}_X(\boldsymbol{x})^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\cdot) \right\|_{\mathscr{H}_K} \\ &= \left(K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}_X(\boldsymbol{x})^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x}) \right)^{1/2}. \end{aligned} \tag{2.20}$$

The Lagrange form (2.17), the reproducing property, and the Cauchy–Schwarz inequality then yield the error formula

$$|f(\boldsymbol{x}) - s_{f,X}(\boldsymbol{x})| = \left| \left\langle f, K_{\boldsymbol{x}} - \boldsymbol{k}_X(\boldsymbol{x})^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X \right\rangle_{\mathscr{H}_K} \right| \leq \|f\|_{\mathscr{H}_K} P_X(\boldsymbol{x}),$$

which is analogous to (2.13).

2.3.3 Smoothing

One can also consider *smoothed* kernel interpolants and cubature rules (Rieger and Zwicknagl, 2010, Sec. 6). A smoothed kernel interpolant (that is actually *not* an interpolant), $s_{f,X,\lambda}$, solves the regularised least-squares problem

$$s_{f,X,\lambda} = \operatorname*{arg\,min}_{g \in \mathscr{H}_K} \left[\sum_{i=1}^{n} (g(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))^2 + \lambda \|g\|_{\mathscr{H}_K}^2 \right], \tag{2.21}$$

for a smoothing parameter $\lambda > 0$. This is a generalisation of (2.18). The unique solution to (2.21) is

$$s_{f,X,\lambda} = \boldsymbol{f}_X^\top (\boldsymbol{K}_X + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{k}_X, \tag{2.22}$$

where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. This approximant is exactly of the same form as the kernel interpolant (2.17) and interpolation is recovered

(i.e., $s_{f,X,0} = s_{f,X}$) by setting $\lambda = 0$ in (2.22). The corresponding smoothed kernel cubature rule is the integral of the smoothed interpolant:

$$Q_{K,\lambda}(f;X) \coloneqq Q(f;X,\boldsymbol{w}_{K,\lambda}) = I(s_{f,X,\lambda}) = \boldsymbol{f}_X^{\mathrm{T}}(\boldsymbol{K}_X + \lambda \boldsymbol{I}_d)^{-1}\boldsymbol{k}_{\mu,X}. \tag{2.23}$$

Its weights are $\boldsymbol{w}_{K,\lambda} = (\boldsymbol{K}_X + \lambda \boldsymbol{I}_d)^{-1}\boldsymbol{k}_{\mu,X}$. Non-zero values of λ are often useful for improving the condition number of the matrix that defines the weights.

2.3.4 Convergence Results

It is often useful to know how accurate kernel cubature rules are. This is typically addressed using convergence analysis that seeks to quantify the rate of decay of the integration error when more points are added. For this purpose, define the *fill-distance*

$$h_{X,\Omega} \coloneqq \sup_{\boldsymbol{x}\in\Omega} \min_{i=1,\dots,n} \|\boldsymbol{x} - \boldsymbol{x}_i\|$$

and the *separation radius*

$$q_X \coloneqq \frac{1}{2}\min_{i\neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq h_{X,\Omega}.$$

Fill-distance is the radius of the largest ball in Ω that does not contain any of the points in X while separation radius is half the minimal distance between points in X. When $h_{X,\Omega}$ is small, X covers the domain Ω well and it is to be expected that the integration error $|I(f) - Q_K(f;X)|$ is also small. By saying that a statement is true for *sufficiently dense* X we mean that there is $h_0 > 0$ such that the claim holds for all point sets with $h_{X,\Omega} \leq h_0$. If the number of points needs to be emphasised, we denote an n-point set by X_n. A sequence $\{X_n\}_{n=1}^{\infty}$ of such sets is said to be *quasi-uniform* if $h_{X_n,\Omega}$ and q_{X_n} remain roughly equal. Formally,

$$q_{X_n} \leq h_{X_n,\Omega} \leq \gamma q_{X_n}$$

for some constant $\gamma \geq 1$ and every $n \in \mathbb{N}$. Quasi-uniformity implies that $h_{X_n,\Omega} = \mathcal{O}(n^{-1/d})$ (Wendland, 2005, Prop. 14.1). For example, a simple calculation shows that uniform Cartesian grids of $n = N^d$ points (with the end-points included) on the hyper-cube $\Omega = [0,1]^d$ have $h_{X_n,\Omega} = \sqrt{d/4}(N-1)^{-1} = \mathcal{O}(n^{-1/d})$.

With these preliminaries we are ready to present convergence results for (smoothed) kernel cubature rules based on Sobolev kernels and the Gaussian kernel. For the most part, these results follow from the bounds on

$$\|f - s_{f,X,\lambda}\|_{L^1(\Omega)} = \int_\Omega |f(\boldsymbol{x}) - s_{f,X,\lambda}(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x}$$

in Wendland and Rieger (2005) and Rieger and Zwicknagl (2010) by using the

following argument: if μ has a bounded Lebesgue density p_μ on Ω, then

$$\begin{aligned} |I(f) - Q_{K,\lambda}(f;X)| &= \left| \int_\Omega [f(\boldsymbol{x}) - s_{f,X,\lambda}(\boldsymbol{x})] p_\mu(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \right| \\ &\leq \|p_\mu\|_{L^\infty(\Omega)} \int_\Omega |f(\boldsymbol{x}) - s_{f,X,\lambda}(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} \\ &= \|p_\mu\|_{L^\infty(\Omega)} \|f - s_{f,X,\lambda}\|_{L^1(\Omega)}. \end{aligned} \tag{2.24}$$

Theorems 2.18 and 2.20 apply to smoothed kernel cubature rules of Section 2.3.3. As such, when the smoothing parameter λ is positive, $h_{X,\Omega} \to 0$ does not guarantee convergence. However, convergence rates of the interpolation case (i.e., $\lambda = 0$) can be achieved by specifying that λ should be a function of $h_{X,\Omega}$ that decays sufficiently fast; see Wendland and Rieger (2005, Prop. 3.6) and Rieger and Zwicknagl (2010, Cor. 6.3). All the results require that the domain satisfy an interior cone condition. In short, this condition prohibits presence of "pinch points" on the boundary of Ω.

Definition 2.17 (Interior cone condition)**.** A domain $\Omega \subset \mathbb{R}^d$ satisfies an *interior cone condition* if there exists an angle $\theta \in (0, 2\pi)$ and a radius $r > 0$ such that for every $\boldsymbol{x} \in \Omega$ there is a unit vector $\boldsymbol{\xi}(\boldsymbol{x})$ such that the cone

$$\left\{ \boldsymbol{x} + \lambda \boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^d, \|\boldsymbol{y}\| = 1, \boldsymbol{y}^\mathsf{T} \boldsymbol{\xi}(\boldsymbol{x}) \geq \cos\theta, \lambda \in [0, r] \right\}$$

is contained in Ω.

The first result, for Sobolev kernels, is a consequence of Proposition 3.6 in Wendland and Rieger (2005), a more conventional version of which for $\lambda = 0$ is given in Wendland (2005, Cor. 11.33). This theorem appears explicitly in a slightly less general form in Kanagawa et al. (2019, Sec. 3).

Theorem 2.18. *Consider the smoothed kernel cubature rule* (2.23) *and let the domain Ω be bounded, have a Lipschitz boundary, and satisfy an interior cone condition. Suppose that the measure μ admits a bounded Lebesgue density function and K is a Sobolev kernel of order $r \in \mathbb{R}_+$ such that $\lfloor r \rfloor > d/2$. Then there is a constant $C > 0$, independent of f and X, such that*

$$|I(f) - Q_{K,\lambda}(f;X)| \leq C \left(h_{X,\Omega}^r + \sqrt{\lambda} \right) \|f\|_{\mathcal{H}_K} \tag{2.25}$$

for any $f \in \mathcal{H}_K = H^r(\Omega)$ and all sufficiently dense X. If X_n are quasi-uniform, this becomes

$$|I(f) - Q_{K,\lambda}(f;X_n)| \leq C \left(n^{-r/d} + \sqrt{\lambda} \right) \|f\|_{\mathcal{H}_K}$$

for a potentially different constant C and all sufficiently large $n \in \mathbb{N}$.

Proof. Equation (2.24) gives

$$|I(f) - Q_{K,\lambda}(f;X)| \leq \|p_\mu\|_{L^\infty(\Omega)} \|f - s_{f,X,\lambda}\|_{L^1(\Omega)}.$$

The bound (2.25) for $\|f - s_{f,X,\lambda}\|_{L^1(\Omega)}$ follows from setting $\tau = r$, $q = 1$, and $j = 0$ in Proposition 3.6 of Wendland and Rieger (2005). □

The rate $n^{-r/d}$ that is attained in the non-smoothed case with quasi-uniform point sets is the optimal rate of convergence of a deterministic cubature rule on a hypercube (Novak, 1988, Sec. 1.3.12). Similar rates of convergence in different spaces of finitely smooth functions are attained by, for instance, quasi Monte Carlo rules (Dick and Pillichshammer, 2010, Ch. 15) and certain sparse grid rules (Novak and Ritter, 1996, 1997).

Kanagawa et al. (2019, Secs. 4 and 5) have used results by Narcowich and Ward (2004) to show that Sobolev kernels are *adaptive* to misspecification of Sobolev smoothness. That is, even if the integrand lives in a Sobolev space rougher than the one induced by the kernel, the kernel cubature method attains the optimal rate of convergence. For similar results for randomly selected points, see Kanagawa et al. (2016). The following theorem is a generalisation based on Narcowich et al. (2006) of the results in Kanagawa et al. (2019).

Theorem 2.19. *Consider the kernel cubature rule* (2.15) *and let the domain* Ω *be bounded, have a Lipschitz boundary, and satisfy an interior cone condition. Let* $0 < s \leq r$ *be such that* $\lfloor s \rfloor > d/2$*. Suppose that the measure* μ *admits a bounded Lebesgue density function and* K *is a Sobolev kernel of order* r*. If* X_n *are quasi-uniform and* $f \in H^s(\Omega)$*, then there is a constant* $C > 0$*, independent of* f *and* X*, such that*

$$|I(f) - Q_K(f; X_n)| \leq C \|f\|_{H^s(\Omega)} n^{-s/d}$$

for all sufficiently large $n \in \mathbb{N}$*.*

Proof. By setting $\tau = r$, $\beta = s$, and $\mu = 0$ in Theorem 4.2 of Narcowich et al. (2006) we obtain

$$\|f - s_{f,X}\|_{L^2(\Omega)} \leq C \|f\|_{H^s(\Omega)} n^{-s/d} \tag{2.26}$$

for some constant $C > 0$ because the mesh ratio $\rho_{X,\Omega} = h_{X,\Omega}/q_X$ is bounded under the quasi-uniformity assumption. Because Ω is bounded, $f - s_{f,X} \in L^1(\Omega)$ and Hölder's inequality produces the standard bound

$$\|f - s_{f,X}\|_{L^1(\Omega)} \leq \sqrt{\mu(\Omega)} \|f - s_{f,X}\|_{L^2(\Omega)}.$$

The claim now follows by combining (2.24) and (2.26). □

The RKHS of the Gaussian kernel, a subset of analytic functions (recall Theorem 2.14), is much smaller than any Sobolev space. As such, it is perhaps unsurprising that exponential rates of convergence can be attained. The following theorem is an amalgamation of the results in Rieger and Zwicknagl (2010, Sec. 6); a non-smoothed version can be found in Wendland (2005, Thm. 11.22). The results can be improved if the point set is denser near the boundary of Ω than in its interior (Rieger and Zwicknagl, 2014).

Theorem 2.20. *Suppose that the measure* μ *admits a bounded Lebesgue density function and* K *is the Gaussian kernel* (2.4)*. Then there are constants* $c_1, c_2 > 0$*, independent of* f *and* X*, such that, for all sufficiently dense* X*,*

$$|I(f) - Q_{K,\lambda}(f; X)| \leq \left(2\,\mathrm{e}^{c_1 \log h_{X,\Omega}/h_{X,\Omega}} + \mathrm{e}^{c_2/h_{X,\Omega}} \sqrt{\lambda}\right) \|f\|_{\mathcal{H}_K}$$

if Ω is a hypercube and

$$|I(f) - Q_{K,\lambda}(f;X)| \leq \left(2\mathrm{e}^{c_1 \log h_{X,\Omega}/h_{X,\Omega}^{1/2}} + c_2\sqrt{\lambda}\right) \|f\|_{\mathcal{H}_K}$$

if Ω is bounded, has a Lipschitz boundary, and satisfies an interior cone condition. If X_n are quasi-uniform, these bounds become

$$|I(f) - Q_{K,\lambda}(f;X_n)| \leq \left(2\mathrm{e}^{-c_1 n^{1/d}\log n} + \mathrm{e}^{c_2 n^{1/d}}\sqrt{\lambda}\right) \|f\|_{\mathcal{H}_K}$$

and

$$|I(f) - Q_{K,\lambda}(f;X_n)| \leq \left(2\mathrm{e}^{-c_1 n^{1/(2d)}\log n} + c_2\sqrt{\lambda}\right) \|f\|_{\mathcal{H}_K},$$

respectively, for potentially different constants c_1 and c_2 and all sufficiently large $n \in \mathbb{N}$.

As opposed to Sobolev spaces induced by Sobolev kernels, the Gaussian RKHS characterised in Theorem 2.14 does not correspond to any classical function space. Consequently convergence of other cubature rules in this space has attracted little attention besides recent work by Kuo and Woźniakowski (2012) and Kuo et al. (2017) on the Gauss–Hermite rule and its tensor-product extensions.

Some final remarks are in order:

- The convergence results in Wendland (2005, Ch. 11) and Rieger and Zwicknagl (2010) apply to many other infinitely smooth kernels besides Gaussians, such as inverse multiquadrics. Zwicknagl and Schaback (2013) provide rates of convergence for Taylor space kernels, including the Hardy kernel.

- The assumption that the integrand lives in the RKHS of the kernel is problematic, particularly so for the Gaussian kernel whose RKHS is rather restricted (recall Section 2.2.2). Results like Theorem 2.19 that are valid also for functions outside the RKHS are therefore valuable. Unfortunately, we are not aware of any analogue of Theorem 2.19 that would hold for the Gaussian kernel.

- All error estimates contain the ominous stipulation that the estimates are valid only for sufficiently dense point sets. This can be problematic in high dimensions where an ever-increasing number of points is required for adequate coverage of the space.

- Even though exponential rates of convergence are obtained for quasi-uniform point sets with the Gaussian kernel, some care should be taken in selecting the points because uniformly placed points induce the famous Runge phenomenon (Platte and Driscoll, 2005; Platte et al., 2011) and consequent impairment of stability of the approximation process. Finitely smooth kernels do not exhibit this problem (De Marchi and Schaback, 2010). See Oettershagen (2017, Sec. 4.3) and Karvonen et al. (2019, Sec. 4) for further discussion on this in the context of cubature.

2.3.5 Historical Notes

It seems that kernel cubature rules were first defined in a form recognisable to a modern practitioner by Larkin (1970), though his interest was still mainly in kernels that induce classical function spaces. Properties of these rules were then studied in a number of articles during the 1970s (Richter, 1970; Richter-Dyn, 1971b,a; Larkin, 1972, 1974; Barrar et al., 1974; Barrar and Loeb, 1975; Bojanov, 1979). The early research was mostly concerned with optimal placement of the points of kernel cubature rules for totally positive kernels (Karlin, 1968) in one dimension; for more modern reviews of the topic, still incompletely understood, see Bojanov (1994), Karvonen et al. (2019), and in particular Oettershagen (2017, Sec. 5.1). As far as we are aware of, the only related results on the multivariate case appear in Gavrilov (1998, 2007).

First convergence results were obtained by Bezhaev (1991) for polyharmonic spline kernels. He used essentially the same argument that was outlined in Section 2.3.4. He also appears to have been the first to consider kernel cubature from a completely kernel-centric point of view, an approach that was later revitalised by Sommariva and Vianello (2006a,b). Since then, the work on error estimates has been largely inspired by applications in statistics and machine learning and the interpretation of kernel cubature as a probabilistic numerical algorithm (Kanagawa et al., 2016, 2019; Briol et al., 2019).

3. Bayesian Cubature

This chapter reviews Gaussian process regression and Bayesian cubature that provide an equivalent probabilistic perspective to kernel interpolation and cubature introduced in Chapter 2. The equivalences are well-known and frequently discussed, see for instance Schaback and Wendland (2006); Fasshauer (2011); Scheuerer et al. (2013); and Kanagawa et al. (2018) for recent surveys.

3.1 Gaussian Process Regression

In *Gaussian process* (GP) *regression* (or *kriging*) (O'Hagan, 1978; Rasmussen and Williams, 2006) the function $f : \Omega \to \mathbb{R}$ of interest—in a sense unknown until evaluated—is modelled as a Gaussian process $f_{\mathrm{GP}} \sim \mathrm{GP}(m, K)$ with a mean function $m : \Omega \to \mathbb{R}$ and a covariance kernel $K : \Omega \times \Omega \to \mathbb{R}$. That is, f_{GP} is a stochastic process characterised by the fact that for any $n \in \mathbb{N}$ and any distinct points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \Omega$ the joint distribution of the random variables $f_{\mathrm{GP}}(\boldsymbol{x}_1), \ldots, f_{\mathrm{GP}}(\boldsymbol{x}_n)$ is Gaussian with mean and covariance specified by m and K:

$$\begin{bmatrix} f_{\mathrm{GP}}(\boldsymbol{x}_1) \\ \vdots \\ f_{\mathrm{GP}}(\boldsymbol{x}_n) \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_n) \end{bmatrix}, \begin{bmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \right).$$

A number of sample paths of four different Gaussian processes are depicted in Figure 3.1. For a comprehensive technical review, see Bogachev (1998). Suppose then that we obtain noisy "data" $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ of the function of interest f:

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \qquad \text{for i.i.d.} \qquad \varepsilon_i \sim \mathrm{N}(0, \lambda).$$

The conditional process $f_{\mathrm{GP}} \mid \boldsymbol{y}$ is a Gaussian process with the mean and covariance functions

$$m_X(\boldsymbol{x}) := \mathbb{E}\left[f_{\mathrm{GP}}(\boldsymbol{x}) \mid \boldsymbol{y} \right] = m(\boldsymbol{x}) + \boldsymbol{k}_X(\boldsymbol{x})^{\mathsf{T}} (\boldsymbol{K}_X + \lambda \boldsymbol{I}_n)^{-1} (\boldsymbol{y} - \boldsymbol{m}_X), \tag{3.1}$$

$$K_X(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{C}\left[f_{\mathrm{GP}}(\boldsymbol{x}), f_{\mathrm{GP}}(\boldsymbol{x}') \mid \boldsymbol{y} \right] = K(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_X(\boldsymbol{x})^{\mathsf{T}} (\boldsymbol{K}_X + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{k}_X(\boldsymbol{x}'), \tag{3.2}$$

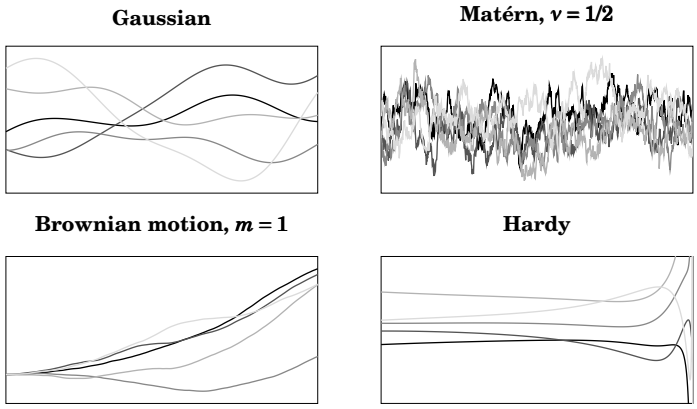

Figure 3.1. Sample paths on $[0,1]$ of zero-mean Gaussian processes with the covariance kernels plotted in Figure 2.1.

where the vector function $\boldsymbol{k}_X : \mathbb{R}^d \to \mathbb{R}^n$ and the kernel matrix $\boldsymbol{K}_X \in \mathbb{R}^{n\times n}$ have been defined in Section 2.3 and $[\boldsymbol{m}_X]_i = m(\boldsymbol{x}_i)$ is an n-vector. Observe that the posterior mean (3.3) coincides with the smoothed kernel interpolant (2.22) if $m \equiv 0$.

As we are interested in approximating deterministic functions, evaluations of which can be performed exactly, in the following we consider only the *noise-free* setting where $\lambda = 0$ and consequently $\boldsymbol{y} = \boldsymbol{f}_X$.[1] The posterior equations (3.1) and (3.2) become

$$m_X(\boldsymbol{x}) = \mathbb{E}\left[f_{\mathrm{GP}}(\boldsymbol{x}) \mid \boldsymbol{f}_X\right] = m(\boldsymbol{x}) + \boldsymbol{k}_X(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}_X^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X), \tag{3.3}$$

$$K_X(\boldsymbol{x},\boldsymbol{x}') = \mathbb{C}\left[f_{\mathrm{GP}}(\boldsymbol{x}), f_{\mathrm{GP}}(\boldsymbol{x}') \mid \boldsymbol{f}_X\right] = K(\boldsymbol{x},\boldsymbol{x}') - \boldsymbol{k}_X(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x}'), \tag{3.4}$$

If $m \equiv 0$, the posterior mean matches the kernel interpolant (2.17) and the posterior variance $K_X(\boldsymbol{x},\boldsymbol{x})$ is equal to the square of the power function (2.20). This shows that the power function does not have to be merely an abstract error indicator but can be interpreted as measure of uncertainty associated to interpolation.

3.2 Bayesian Cubature

In (standard) *Bayesian cubature* (O'Hagan, 1991; Minka, 2000; Briol et al., 2019) the deterministic integrand $f : \Omega \to \mathbb{R}$ is modelled with a zero-mean Gaussian process $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$ and the data $\boldsymbol{y}$ consists of noise-free integrand evaluations, $\boldsymbol{y} = \boldsymbol{f}_X$. Because the integration operator $I(f) = \int_\Omega f \,\mathrm{d}\mu$ is a linear functional, the posterior $I(f_{\mathrm{GP}}) \mid \boldsymbol{f}_X$ is a Gaussian random variable. From the equivalence

[1] In this case, there are certain technicalities, related to the likelihood of the noise-free observations $\boldsymbol{y} = \boldsymbol{f}_X$ being degenerate, that need to be taken care of; see Cockayne et al. (2019a, Sec. 2.5) and Kanagawa et al. (2018, Sec. 3.1) for more details.

between Gaussian process regression and kernel interpolation as well as the interpretation of kernel cubature rules as integrated kernel interpolants it follows that the mean of this random variable is equal to the kernel cubature rule,

$$\mathbb{E}\left[I(f_{\mathrm{GP}}) \mid \boldsymbol{f}_X\right] = \int_\Omega m_X \,\mathrm{d}\mu = Q_K(f;X), \tag{3.5}$$

and its variance coincides with the squared worst-case error,

$$\mathbb{V}\left[I(f_{\mathrm{GP}}) \mid \boldsymbol{f}_X\right] = \int_\Omega K_X(\boldsymbol{x},\boldsymbol{x}')\,\mathrm{d}\mu(\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}') = e_K(X,\boldsymbol{w}_K)^2. \tag{3.6}$$

The use of (3.6) in probabilistic quantification of the numerical uncertainty associated to using $Q_K(f;X)$ to approximate $I(f)$ can be seen as the main difference between kernel and Bayesian cubature. Central challenge—absent in kernel cubature—of Bayesian cubature is then to make sure that this posterior variance actually is a meaningful indicator of the uncertainty; problems can arise if the Gaussian process, defined through its covariance kernel K, is not a good model of the true integrand or if the posterior variance is not scaled properly. Such issues of prior specification and uncertainty calibration are discussed in Section 3.3.

3.2.1 Probabilistic Numerics

Bayesian cubature is an example of a *probabilistic numerical method*. The central idea in *probabilistic numerics* (Hennig et al., 2015; Cockayne et al., 2019a) is that numerical approximation of an analytically intractable quantity can be viewed as a statistical inference problem. The function of interest can only be evaluated at a finite number of points and hence its value elsewhere is effectively unknown; it thus makes sense to exploit prior knowledge about the function by placing a prior on it, for example in the form of a Gaussian process. This allows for computing a full-fledged posterior distribution of the quantity of interest, such as an integral, that contains more information than is available in a single point estimate. The most famous exposition of this idea is perhaps due to Diaconis (1988, p. 163):

> Consider a given function $f : [0,1] \to \mathbf{R}$ such as
>
> $$f(x) = \exp\left\{\cosh\left(\frac{x + 2x^2 + \cos x}{3 + \sin x^3}\right)\right\}. \tag{1}$$
>
> If you require $\int_0^1 f(x)dx$, a formula such as (1) isn't of much use and leads to questions like "What does it mean to 'know' a function?" The formula says some things (e.g. f is smooth, positive, and bounded by 20 on $[0,1]$) but there are many other facts about f that we don't know (e.g., is f monotone, unimodal, or convex?).
>
> Once we allow that we don't know f, but do know some things, it becomes natural to take a Bayesian approach to the quadrature problem:
>
> - Put a prior on continuous functions $C[0,1]$

- Calculate f at $x_1, x_2, \ldots, x_n$
- Compute a posterior
- Estimate $\int_0^1 f$ by a Bayes rule

Although the origins of probabilistic numerics can be traced back to the work of Larkin (1972), it was not until a reintroduction by Diaconis (1988)[2] and O'Hagan (1991, 1992) that the field began slowly gaining traction. Important recent expository articles are the review and a "call to arms" of Hennig et al. (2015) and the historical account by Oates et al. (2019b). Much foundational work on the rigorous definition of a *Bayesian* probabilistic numerical method has been done by Cockayne et al. (2019a). Besides numerical integration, probabilistic methods have been developed for solving of ordinary (Skilling, 1992; Schober et al., 2018; Tronarp et al., 2019) and partial (Cockayne et al., 2017) differential equations and for numerical linear algebra (Hennig, 2015; Bartels et al., 2018; Cockayne et al., 2019b).

3.2.2 Literature Review

This section consists of a short historical account of Bayesian cubature and a comprehensive collection of references to all aspects and applications of Bayesian cubature. Included here are only references that explicitly adopt or contain the Gaussian process based probabilistic point of view (though not necessarily the term "Bayesian cubature"). Literature on the equivalent reproducing kernel Hilbert space characterisation has been reviewed in Chapter 2.

A probabilistic formulation for kernel cubature and the consequent ramifications for statistical quantification of uncertainty in numerical approximation make their first appearance in the seminal works of Larkin (1972, 1974),[3] but it is only in the independent work of O'Hagan (1988, 1991, 1992) where the Gaussian process formulation used here and the name "Bayesian quadrature" explicitly appear for the first time. The contributions by O'Hagan and Diaconis (1988) triggered further research during the next decade or so on in statistics (Kennedy and O'Hagan, 1996; Cook and Clayton, 1998; Kennedy, 1998, 2000) and, more influentially, machine learning (Minka, 2000; Rasmussen and Ghahramani, 2002) communities. For a relatively early application, see Kumar et al. (2008). In particular, Cook and Clayton (1998) develop methods for sequential selection of the evaluation points $\boldsymbol{x}_i$, Kennedy (1998) essentially deals with computation of kernel means, and Rasmussen and Ghahramani (2002) coin the

[2] Some earlier commentary, including appearance of the term "Bayesian numerical analysis", can be found already in Diaconis and Freedman (1983, p. 110).

[3] Suldin (1959, 1960) has done some earlier work in a more limited setting of Wiener measures. He does not view the Gaussian process as a prior for the integrand nor the output of the numerical approximation as a probability distribution, being rather a precursor to average-case analysis (Ritter, 2000). See Oates and Sullivan (2019, Sec. 2.2) for more details. A different early probabilistic idea appears in Ajne and Dalenius (1960).

term “Bayesian Monte Carlo” for Bayesian cubature based on Monte Carlo samples. Although he does not generally provide rigorous proofs, Minka (2000) has numerous interesting results and insights on the relationship between Bayesian quadrature and polynomial and spline based quadrature rules, discussed in detail in Chapter 5. A short contemporary review and some discussion can be found in Evans and Swartz (2000, Sec. 5.7). However, it was not until during the 2010s that a true outpouring of contributions began, and the rest of this section reviews work published during the past ten years in machine learning, statistics, and signal processing literature.

In Machine Learning. It seems that the beginning of renewed interest in Bayesian cubature can be dated to 2012, a year that witnessed the publication of a number of articles (Huszár and Duvenaud, 2012; Osborne et al., 2012b,a). The outpouring of work that soon followed contained contributions, to name a few, on selection of the integration points by numerical optimisation (Briol et al., 2015), convergence results in the misspecified setting when $f \notin \mathcal{H}_K$ (Kanagawa et al., 2016), selection of the sampling distribution in Bayesian Monte Carlo (Briol et al., 2017), relationship between Bayesian cubature and random feature expansions (Bach, 2017), and generalisations where also the measure μ is considered unknown (Oates et al., 2017b), derivative evaluations are employed (Prüher and Särkkä, 2016; Wu et al., 2018),[4] and the integrand is allowed to be vector-valued (Xi et al., 2018). Other recent works are Hamrick and Griffiths (2013); Ma et al. (2014); Kersting and Hennig (2016); Fitzsimons et al. (2017); Paul et al. (2018); Acerbi (2018b,a); Tronarp et al. (2018); Chai et al. (2019); Sonoda (2019); and Gessner et al. (2019).

A topic that has seen a fair amount of activity is modelling of positivity of the integrand (Osborne et al., 2012a; Gunter et al., 2014; Chai and Garnett, 2018; Wagstaff et al., 2018). Positivity cannot be encoded by the standard Gaussian process model and it can easily happen that the posterior mean function and some of the cubature weights become negative even if the true integrand is positive everywhere. However, one can introduce (for instance) the function $g(\boldsymbol{x}) = f(\boldsymbol{x})^{1/2}$ and place a GP prior with kernel K on g instead of directly on f. The posterior of f becomes a non-central Chi-squared process whose mean and covariance functions can be computed in closed form in some cases. This particular approach was introduced by Gunter et al. (2014) under the name *warped sequential active Bayesian integration* (WSABI) and has the potential advantage of having its posterior variance data-dependent so that selecting new integration points by minimising the variance provides an adaptive point selection scheme. The most general framework is due to Chai and Garnett (2018). It is worth noting that a square-root transform for enforcing positivity appears already in the early work of Larkin, see Larkin (1969, Sec. 7) and Larkin (1972, Sec. 2.7).

[4] Already O’Hagan (1992, Sec. 3.4) shortly discusses the use of derivative evaluations in Bayesian cubature.

In Statistics. Even though majority of work on Bayesian cubature since 2000 has been published in machine learning literature, there are some recent exceptions. In addition to some results on convergence for random and quasi-random point sets, Briol et al. (2019) provide the most comprehensive modern overview and discussion on Bayesian cubature. An expanded version appears in the the doctoral dissertation Briol (2018). The cubic computational cost in the number of integration points, due to the need to solve a linear system, is alleviated by Rathinavel and Hickernell (2018). Karvonen et al. (2019) discuss qualitative properties, such as positivity, of the Bayesian cubature weights and Ehler et al. (2019) are concerned with integration on general closed manifolds.

In Signal Processing. Bayesian cubature has been extensively applied to computation of Gaussian integrals arising in certain extensions to non-linear dynamical systems of the classical Kalman filter (Särkkä, 2013). Deisenroth et al. (2009, 2012) were first to propose the use of Gaussian process based numerical integration in this context. Since then, different versions and extensions of the idea have been studied in Särkkä et al. (2014); Prüher and Šimandl (2016); Särkkä et al. (2016); Prüher et al. (2017); Prüher and Straka (2018); and Prüher et al. (2018). An interesting connection is that a part of this approach, perhaps most clearly outlined by Prüher and Straka (2018, Sec. IV) (who use the term *Gaussian process moment transform*), closely resembles WSABI in that Gaussian integrals of f^2 are essentially approximated by placing a GP prior on f.

Other. Brouillat et al. (2009) and Marques et al. (2013, 2015) have applied Bayesian cubature to the global illumination problem, a numerical integration task arising in generation of computer graphics. Pronzato and Zhigljavsky (2018) study greedy selection of the integration points.

3.3 Uncertainty Quantification

In Bayesian cubature the prior model is specified via selection of the covariance kernel K. Its role was clearly articulated already by Larkin (1972, p. 406):

> Typically, we shall assume *general* properties, such as continuity or nonnegativity of the solution and/or its derivatives, and use the given *specific* properties in order to assist in making a selection from the class [...] of all functions possessing the assumed general properties. We shall choose [this class] either to be a Hilbert space or to be simply related to one.

In our setting, this class will obviously be the RKHS $\mathcal{H}_K$. However, one needs to proceed carefully, for, if the RKHS is infinite-dimensional, *the sample paths of the GP do not live in the RKHS* (Lukić and Beder, 2001; Steinwart, 2017; Kanagawa et al., 2018). This caution ought to be kept in mind when Bayesian cubature is used for the purpose of uncertainty quantification.

Stationary kernels are commonly used as a default choice, although more accurate estimates and more reliable uncertainty quantification could likely be often obtained with carefully designed non-stationary kernels (Stein and Hung, 2019). Stationary kernels are typically parametrised by a smoothness (or regularity) parameter and a collection of scale parameters. The *smoothness parameter* determines the degree of differentiability of the kernel and consequently how smooth the integrand is assumed to be. Because the number of continuous derivatives a function possesses cannot be accurately determined from a finite number of function evaluations, the smoothness parameter is usually fixed beforehand based on a vague notion of expected smoothness of the integrand.[5] An example is of course the parameter ν of the Matérn kernel (2.3) that for $\nu = \infty$ yields also the infinitely smooth Gaussian kernel (2.4). Stein (1999, Secs. 1.6 & 1.7) strongly advocates the use of Matérn kernels.

There are typically two *scale* parameters, the *magnitude* parameter $\sigma > 0$ and the *length-scale* parameter $\ell > 0$ that parametrise the kernel as

$$K_{\sigma,\ell}(\boldsymbol{x},\boldsymbol{x}') \coloneqq \sigma^2 K\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{\ell}\right),$$

where K is now a stationary "basis" kernel. The single length-scale parameter can also be replaced with a full length-scale matrix but as it is even more difficult to select this matrix than a single parameter, we do not consider this extension here. When computationally feasible, the scale parameters are usually determined from data—for meaningful uncertainty quantification this is crucial since the posterior variance as such does not depend on the function evaluations and a procedure that outputs the same uncertainty estimates for two wildly different functions is clearly of dubious utility. In Gaussian process regression these parameters are typically selected using either marginalisation or empirical Bayes (i.e., maximum likelihood). Cross-validation, which we do not discuss, is popular in scattered data approximation literature. Unfortunately, little is known on behaviour of the parameter estimates. The only results we are aware of concern empirical Bayes for the Gaussian kernel and points placed uniformly on the unit interval (Xu and Stein, 2017) and the Gaussian white noise model (Hadji and Szabó, 2019). As the convergence results reviewed in Section 2.3.4 apply to a *fixed* kernel (and, accordingly, fixed RKHS), little can be said about convergence of the Bayesian cubature posterior mean when the length-scale parameter is updated as new function evaluations are obtained unless it is assumed that the length-scale estimate converges to a finite non-zero value as $n \to \infty$.

3.3.1 Marginalisation of Kernel Parameters (Full Bayes)

A *fully Bayesian* approach is to place a prior on the scale parameters and marginalise them out. Unsurprisingly, for all but some special cases this ap-

[5] For maximum likelihood based selection of the smoothness parameter, see Szabó et al. (2015).

proach is intractable because it involves computation of a challenging integral. One of the most important special cases is that the magnitude parameter σ can be marginalised out in closed form if it is given the non-informative prior $p(\sigma^2) \propto 1/\sigma^2$ and ℓ is fixed. If this is done, the integral posterior $I(f_{\mathrm{GP}}) \mid \boldsymbol{f}_X$ becomes Student's t distribution with n degrees of freedom and the mean and variance

$$Q_{K_\ell}(f;X) \quad \text{and} \quad \frac{(\boldsymbol{f}_X - \boldsymbol{m}_X)^\top \boldsymbol{K}_{\ell,X}^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X)}{n-2} e_{K_\ell}(X, \boldsymbol{w}_{K_\ell})^2,$$

where $K_\ell(\boldsymbol{x},\boldsymbol{x}') = K([\boldsymbol{x}-\boldsymbol{x}']/\ell)$ and $\boldsymbol{K}_{\ell,X}$ is the kernel matrix of this kernel; see O'Hagan (1991, Sec. 2.2) or Santner et al. (2003, Sec. 4.1.3). The differences to the unmarginalised version in Section 3.2 are that there is an additional factor in the posterior variance and that the posterior has heavier tails. However, as n increases, the posterior converges to a Gaussian.

3.3.2 Maximum Likelihood for Kernel Parameters (Empirical Bayes)

In *empirical Bayes* one selects values of σ and ℓ that maximise the marginal likelihood of the "data" $\boldsymbol{f}_X$ that has been obtained so far. The marginal likelihood in the noise-free Gaussian process model described in Section 3.1 is (Rasmussen and Williams, 2006, Eq. (2.30))

$$l(\sigma,\ell) = \det(2\pi \boldsymbol{K}_{\sigma,\ell,X})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{f}_X - \boldsymbol{m}_X)^\top \boldsymbol{K}_{\sigma,\ell,X}^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X)\right),$$

maximisation of which as a function of σ and ℓ is equivalent to maximisation of

$$\log l(\sigma,\ell) = -\frac{1}{2}(\boldsymbol{f}_X - \boldsymbol{m}_X)^\top \boldsymbol{K}_{\sigma,\ell,X}^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X) - \frac{1}{2}\log\det(\boldsymbol{K}_{\sigma,\ell,X}) - \frac{n}{2}\log(2\pi). \quad (3.7)$$

The maximum likelihood estimate of σ is available in closed form. To derive an expression for it, note that

$$\begin{aligned}\log l(\sigma,\ell) = &-\frac{1}{2\sigma^2}(\boldsymbol{f}_X - \boldsymbol{m}_X)^\top \boldsymbol{K}_{\ell,X}^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X) - \frac{1}{2}\log\det(\boldsymbol{K}_{\ell,X}) \\ &- n\log\sigma - \frac{n}{2}\log(2\pi).\end{aligned}$$

Differentiation with respect to σ gives

$$\frac{\partial}{\partial\sigma}\log l(\sigma,\ell) = \frac{1}{\sigma^3}(\boldsymbol{f}_X - \boldsymbol{m}_X)^\top \boldsymbol{K}_{\ell,X}^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X) - \frac{n}{\sigma}.$$

By requiring the derivative to vanish we obtain the maximum likelihood estimate

$$\sigma_{\mathrm{ML}}^2 = \frac{(\boldsymbol{f}_X - \boldsymbol{m}_X)^\top \boldsymbol{K}_{\ell,X}^{-1}(\boldsymbol{f}_X - \boldsymbol{m}_X)}{n}. \quad (3.8)$$

Unfortunately, there is no closed-form solution for the maximum likelihood estimate of the length-scale parameter. Note that the integral posterior corresponding to the use of (3.8) is Gaussian with the mean $Q_{K_\ell}(f;X)$ and standard deviation $\sigma_{\mathrm{ML}} e_{K_\ell}(X, w_{K_\ell})$, and that, for large n, the posterior is consequently indistinguishable from the Student's t posterior obtained via marginalisation in Section 3.3.1.

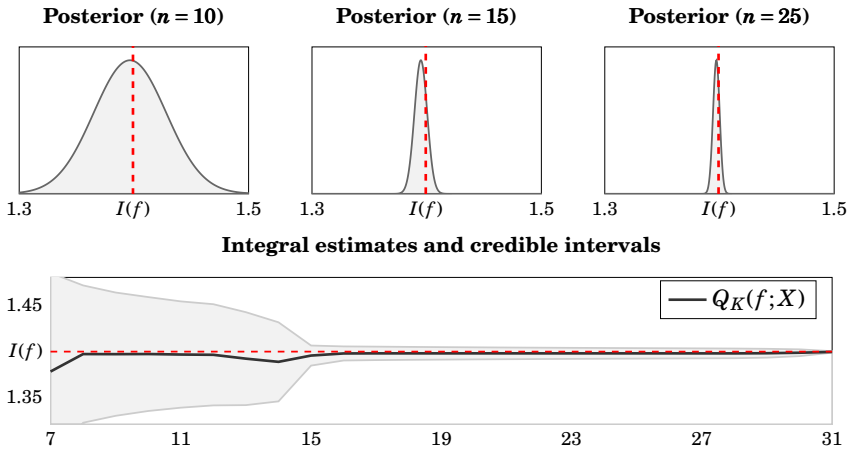

Figure 3.2. Illustration of uncertainty quantification for the toy integration problem (3.9). The upper figure shows how the Bayesian quadrature posterior contracts towards the true integral value $I(f) \approx 1.3992$ while the lower displays the 95% credible interval around the integral posterior mean $Q_K(f;X)$.

3.3.3 Example: Credible Intervals

Quality of the uncertainty quantification provided by Bayesian cubature for the unknown value of the integral is often assessed by examining Bayesian credible intervals. This has been suggested already by Larkin (1974, 1972, Sec. 5.4); see also Wahba (1983) for a relevant study on the more general regression context. For a recent use of credible intervals for this purpose, see Briol et al. (2019, Sec. 5) and Rathinavel and Hickernell (2018). The work of Rathinavel and Hickernell is particularly interesting as they use credible intervals to decide when to terminate the numerical integration procedure.

Figure 3.2 shows how the Bayesian quadrature posterior contracts towards the true integral value when the number of points increases. This toy problem consists of computation of

$$\int_0^1 f(x)\,\mathrm{d}x \qquad \text{for} \qquad f(x) = \exp\left(\sin(6x)^2 - \frac{x}{2}\right) \tag{3.9}$$

using the Matérn kernel (2.3) with $\nu = 5/2$. For each n, the magnitude and length-scale parameters of the kernel were selected using empirical Bayes.

4. Computational Methods (Publications I–III)

Recall from Section 2.3 that the integral approximation produced by a kernel cubature rule is

$$Q(f;X,\boldsymbol{w}_K) = \boldsymbol{f}_X^\top \boldsymbol{w}_K = \boldsymbol{f}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\mu,X}, \tag{4.1}$$

where $\boldsymbol{f}_X \in \mathbb{R}^n$ contains the integrand evaluations at X, $[\boldsymbol{K}]_{ij} = K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is the $n \times n$ kernel matrix and $\boldsymbol{k}_{\mu,X} \in \mathbb{R}^n$ consists of evaluations of the kernel mean. Computation of the weights $\boldsymbol{w}_K = \boldsymbol{K}_X^{-1}\boldsymbol{k}_{\mu,X}$ thus necessitates solving a linear system of n equations defined by the kernel matrix. Due to their cubic time and quadratic memory complexity, naive linear solvers cannot cope with more than some tens of thousands of points unless the kernel matrix is, for example, sparse. For integration problems requiring a large number of points due to high dimensionality of the integrand or a high level of accuracy desired, approximations or exploitation of potential structure present in $\boldsymbol{K}_X$ and $\boldsymbol{k}_{\mu,X}$ are required. A second problem that is often encountered when infinitely smooth kernels, such as Gaussians, are used is that the kernel matrix quickly becomes ill-conditioned (Schaback, 1995). This chapter deals with these two issues and consists of the following parts:

1. Section 4.1 contains a short literature review of approximate and non-approximate algorithms for evaluating (4.1).
2. Section 4.2 reviews fully symmetric sets and the results in Publications **I** and **II** on how these sets can be used to enable kernel cubature for up to tens of millions of points.
3. Section 4.3 is a summary of the results in Publication **III** on the use of the Mercer expansion for numerically stable and explicit approximation of the weights $\boldsymbol{w}_K$ when both the kernel and integration measure are Gaussian.

4.1 A Short Literature Review

In the literature there exist a myriad of methods for fast kernel interpolation and cubature and Gaussian process regression. Roughly speaking, these can be

divided into *approximate* and *exact* methods. Approximate methods make use of different approximations to, for example, the kernel to achieve a computational speed-up while exact methods are typically based on exploiting structurality of specifically designed point sets. Examples of approximate methods popular in machine learning include certain sparse methods (Snelson and Ghahramani, 2005), variational and spectral methods (Hensman et al., 2018), and reduced rank approximations (Solin and Särkkä, 2019), to name a few. For finitely smooth kernels there are exact methods based on a connection to Kalman filtering (Hartikainen and Särkkä, 2010; Särkkä et al., 2013). Little use of any of these methods has been made in kernel or Bayesian cubature.

Perhaps the simplest, but also exceedingly inflexible, exact method for kernel cubature is based on the use of full tensor grids and was suggested for Bayesian cubature already by O'Hagan (1991, Sec. 4). In short, if the kernel, domain and measure take the product forms

$$K(\boldsymbol{x},\boldsymbol{x}') = \prod_{i=1}^{d} K_i(x_i, x_i'), \qquad \Omega = \Omega_1 \times \cdots \times \Omega_d, \qquad \mu = \mu_1 \otimes \cdots \otimes \mu_d$$

for kernels $K_i : \Omega_i \times \Omega_i \to \mathbb{R}$, domains $\Omega_i \subset \mathbb{R}$, and measures μ_i on Ω_i and the point set is a Cartesian product of d one-dimensional sets,

$$X = X_1 \times \cdots \times X_d, \qquad \text{for} \qquad X_i \subset \Omega_i, \quad \#X = n_i,$$

then the kernel cubature weight for a point $\boldsymbol{x}_{\boldsymbol{\alpha}} = (x_{\alpha_1}, \ldots, x_{\alpha_d}) \in X$ defined by a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$ is

$$w_{K,\boldsymbol{\alpha}} = \prod_{i=1}^{d} w_{K_i,\alpha_i},$$

where $\boldsymbol{w}_{K_i} \in \mathbb{R}^{n_i}$ are the kernel cubature weights based on the kernel K_i, points X_i, and the measure μ_i on Ω_i. This reduces the computational complexity of kernel cubature from $\mathcal{O}((n_1 \times \cdots \times n_d)^3)$ to $\mathcal{O}(n_1^3 + \cdots + n_d^3)$, but imposes a severe restriction on the form of point sets that can be used. See Oettershagen (2017, Sec. 2.4) for a more comprehensive review that also covers sparse grids. Other exact methods for kernel cubature have been proposed by Fuselier et al. (2014) and Rathinavel and Hickernell (2018).

4.2 Fully Symmetric Kernel Cubature

This section contains a review of the results in Publications **I** and **II**. The main contribution of these publications is summarised in Theorem 4.1, which provides an efficient and exact algorithm for computing the kernel cubature weights when the point set is a union of fully symmetric sets.

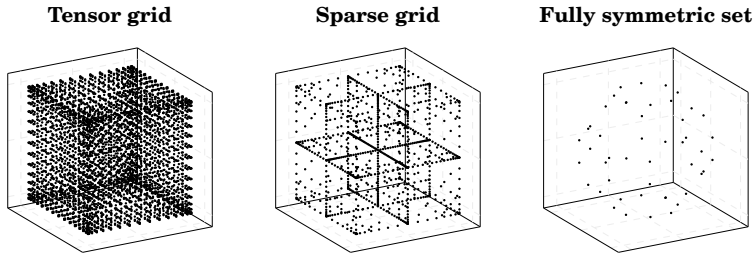

Figure 4.1. Three different point sets in $\mathbb{R}^3$ that are unions of fully symmetric sets.

4.2.1 Fully Symmetric Sets

The *fully symmetric set* generated by a *generator vector* $\boldsymbol{\lambda} \in \mathbb{R}^d$ is the set

$$[\boldsymbol{\lambda}] := \left\{ (s_1 \lambda_{q_1}, \ldots, s_d \lambda_{q_d}) : \boldsymbol{q} \in \Pi_d, \boldsymbol{s} \in S_d \right\} \subset \mathbb{R}^d,$$

where $\Pi_d \subset \mathbb{N}^d$ is the collection of all permutations of the first d positive integers and S_d contains every vector of the form $\boldsymbol{s} = (s_1, \ldots, s_d)$ for each s_i either 1 or -1. This means that $[\boldsymbol{\lambda}]$ is the collection of all points that can be obtained from $\boldsymbol{\lambda}$ by permutations and sign-changes of individual coordinates. Some fully symmetric sets are displayed in Figure 4.1. A generator vector $\boldsymbol{\lambda}$ that has m non-zero elements, l of which are distinct and have multiplicities $m_1, \ldots, m_l$, and m_0 elements that are zero generates a fully symmetric set of cardinality

$$\#[\boldsymbol{\lambda}] = \frac{2^m d!}{m_0! \cdots m_l!}. \tag{4.2}$$

This number grows very fast with the dimension d if $\boldsymbol{\lambda}$ contains more than a few distinct elements; see Table 4.1. Fully symmetric sets can be also defined using *permutation and sign-change matrices*. These are $d \times d$ matrices that have exactly one entry of 1 or -1 on each of their columns and rows. Note that these matrices are invertible. We denote their collection by $\mathscr{P}_d$, so that $[\boldsymbol{\lambda}] = \{\boldsymbol{P}\boldsymbol{\lambda} : \boldsymbol{P} \in \mathscr{P}_d\}$. The following four notions of different fully symmetric objects are necessary for stating our results:

- A **domain** $\Omega \subset \mathbb{R}^d$ is fully symmetric if it is closed under coordinate permutations and sign-changes:

  $$\Omega = \boldsymbol{P}\Omega := \{\boldsymbol{P}\boldsymbol{x} : \boldsymbol{x} \in \Omega\} \quad \text{for any} \quad \boldsymbol{P} \in \mathscr{P}_d.$$

  Standard domains such as $\mathbb{R}^d$, hypercubes of the form $[-a,a]^d$, and centered balls are fully symmetric.

- A **measure** μ on a fully symmetric domain Ω is fully symmetric if it is invariant under fully symmetric pushforwards:

  $$\mu = \boldsymbol{P}_* \mu \quad \text{for} \quad \text{for any} \quad \boldsymbol{P} \in \mathscr{P}_d,$$

Table 4.1. Sizes of fully symmetric sets, as computed from (4.2), generated by the d-dimensional ($d = 1,\ldots,9$) generator vector $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_l,0,\ldots,0)$ with $l \le m$ distinct non-zero elements $\lambda_1,\ldots,\lambda_l$.

Dimension (d)

	2	3	4	5	6	7	8	9
$l = 1$	4	6	8	10	12	14	16	18
$l = 2$	8	24	48	80	120	168	224	288
$l = 3$	-	48	192	480	960	1,680	2,688	4,032
$l = 4$	-	-	384	1,920	5,760	13,440	26,880	48,384
$l = 5$	-	-	-	3,840	23,040	80,640	215,040	483,840
$l = 6$	-	-	-	-	46,080	322,560	1,290,240	3,870,720
$l = 7$	-	-	-	-	-	645,120	5,160,960	23,224,320

where the pushforward measure is defined by $(\boldsymbol{P}_*\mu)(A) = \mu(\boldsymbol{P}^{-1}A)$ for measurable $A \subset \Omega$. The Lebesgue measure and isotropic Gaussian measures are standard examples of fully symmetric measures. As will be described in Section 4.2.2, the restriction that the measure needs to be fully symmetric can be circumvented by the use of a change of measure trick.

- A **kernel** $K\colon \Omega \times \Omega \to \mathbb{R}$ defined on a fully symmetric domain Ω is fully symmetric if

$$K(\boldsymbol{P}\boldsymbol{x},\boldsymbol{P}\boldsymbol{x}') = K(\boldsymbol{x},\boldsymbol{x}') \qquad \text{for any} \qquad \boldsymbol{P} \in \mathscr{P}_d.$$

  All isotropic kernels and kernels formed out of products or sums of dimension-wise one-dimensional isotropic kernels are fully symmetric.

- A **cubature rule** $Q(X,\boldsymbol{w})$ is fully symmetric if the point set is a union of fully symmetric sets and any two points in a particular fully symmetric set have the same weight:

$$X = \bigcup_{j=1}^{J} [\boldsymbol{\lambda}_j] \qquad \text{and} \qquad Q(f;X,\boldsymbol{w}) = \sum_{j=1}^{J} w_{\boldsymbol{\lambda}_j} \sum_{\boldsymbol{x}\in[\boldsymbol{\lambda}_j]} f(\boldsymbol{x})$$

  for distinct $\boldsymbol{\lambda}_j \in \Omega$ and some $w_{\boldsymbol{\lambda}_1},\ldots,w_{\boldsymbol{\lambda}_J} \in \mathbb{R}$. That is, given a natural ordering of the points the weight vector takes the block form

$$\boldsymbol{w} = \left(w_{\boldsymbol{\lambda}_1}\mathbb{1}_{\#[\boldsymbol{\lambda}_1]} \cdots w_{\boldsymbol{\lambda}_J}\mathbb{1}_{\#[\boldsymbol{\lambda}_J]}\right) \in \mathbb{R}^n,$$

  where $\mathbb{1}_m$ is an m-vector of ones. Fully symmetric cubature rules based on polynomial exactness conditions have been studied since the 1960s (Lyness, 1965; McNamee and Stenger, 1967). The most general constructions appear to be due to Genz (1986) and Genz and Keister (1996).

4.2.2 Exploiting Symmetry

When the point set X is a union of fully symmetric sets and the domain, measure, and kernel are all fully symmetric, it would seem natural that the corresponding kernel-based cubature rule should be fully symmetric. This is indeed the case and, moreover, the distinct weights, one for each fully symmetric set, can be computed very efficiently. The following theorem is the main result of Publication **I**.

Theorem 4.1. *Suppose that the domain Ω, measure μ, and kernel K are fully symmetric. Let X be a union of J distinct fully symmetric sets generated by $\{\boldsymbol{\lambda}_j\}_{j=1}^J$: $X = \bigcup_{j=1}^J [\boldsymbol{\lambda}_j]$. Then the resulting kernel cubature rule is fully symmetric with distinct weights $\boldsymbol{w}_K^{\boldsymbol{\lambda}} \in \mathbb{R}^J$:*

$$Q_K(f;X) = \sum_{j=1}^{J} w_{K,j}^{\boldsymbol{\lambda}} \sum_{\boldsymbol{x}\in[\boldsymbol{\lambda}_j]} f(\boldsymbol{x}).$$

The weights solve the linear system $\boldsymbol{S}\boldsymbol{w}_K^{\boldsymbol{\lambda}} = \boldsymbol{k}_{\mu,\boldsymbol{\lambda}}$ of J equations, where

$$[\boldsymbol{S}]_{ij} = \sum_{\boldsymbol{x}\in[\boldsymbol{\lambda}_j]} K(\boldsymbol{\lambda}_i, \boldsymbol{x}) \quad \text{and} \quad [\boldsymbol{k}_{\mu,\boldsymbol{\lambda}}]_j = K_\mu(\boldsymbol{\lambda}_j).$$

The worst-case error is

$$e_K(X, \boldsymbol{w}_K)^2 = I(K_\mu) - \sum_{j=1}^{J} w_{K,j}^{\boldsymbol{\lambda}} K_\mu(\boldsymbol{\lambda}_j) \#[\boldsymbol{\lambda}_j].$$

The naive method of computing the weights requires n^2 kernel evaluations and solving of a linear system of n equations; Theorem 4.1 reduces these to nJ and J, respectively. This makes kernel cubature rules feasible for up to millions of points (numerical examples in Publication **I** go up to 15,005,761 points). As cardinalities of fully symmetric sets tend to grow rapidly with the dimension, the nJ evaluations of the kernel needed to form $\boldsymbol{S}$ typically constitute the main computational bottleneck. Publication **II** extends Theorem 4.1 in various ways:

- Similar computational simplifications are possible for the Bayes–Sard cubature, to be introduced in Section 5.2. In particular, if the function space π in the definition of a Bayes–Sard rule consists of even monomials up to a given degree and is of dimension J_α, then the distinct Bayes–Sard weights can be solved from a linear system of $J + J_\alpha$ equations.

- A method for simultaneous computation of multiple related integrals (Xi et al., 2018) can also be made computationally competitive through the use of fully symmetric sets.

- Symmetric change of measure can be used to relax the requirement that μ be fully symmetric. Let μ_* be a fully symmetric measure on Ω such that μ

is absolutely continuous with respect to μ_*. Then the integral of interest can be written as

$$\int_\Omega f \,\mathrm{d}\mu = \int_\Omega f \frac{\mathrm{d}\mu}{\mathrm{d}\mu_*} \,\mathrm{d}\mu_*,$$

where $\mathrm{d}\mu/\mathrm{d}\mu_*$ is a Radon–Nikodym derivative, and a fully symmetric kernel-based cubature rule computed for μ_*. This method is related to importance sampling and seems to work well in similar settings as importance sampling.

It is also noteworthy that in the case $\#[\boldsymbol{\lambda}_1] = \cdots = \#[\boldsymbol{\lambda}_J]$ the set of eigenvalues of $\boldsymbol{S}$ is a subset of those of the full kernel matrix $\boldsymbol{K}_X$. Consequently, the condition number of $\boldsymbol{S}$ cannot exceed that of $\boldsymbol{K}_X$. However, it seems difficult to say anything about the relation of the condition numbers in the general case of fully symmetric sets of unequal cardinalities.

4.3 Mercer Expansions

This section contains a review of the results in Publication **III**, which were inspired by the work of Fasshauer and McCourt (2012) on numerically stable interpolation with increasingly flat kernels (see Section 5.4). Define the integral operator

$$L_K g(\boldsymbol{x}) \coloneqq \int_\Omega K(\boldsymbol{x},\boldsymbol{x}') g(\boldsymbol{x}') \,\mathrm{d}\mu(\boldsymbol{x}')$$

on $L^2(\Omega,\mu)$ and suppose that $\{\lambda_p\}_{p=0}^\infty$ and $\{\varphi_p\}_{p=0}^\infty$ are its positive eigenvalues and corresponding $L^2(\Omega,\mu)$-orthonormal eigenfunctions. That is,

$$L_K \varphi_p = \lambda_p \varphi_p \qquad \text{and} \qquad \int_\Omega \varphi_p \varphi_q \,\mathrm{d}\mu = \delta_{pq}.$$

Our results are based on Mercer's theorem, a version of which for general domains is given below. For a proof, see Sun (2005) and also Steinwart and Scovel (2012).

Theorem 4.2 (Mercer's theorem)**.** *Suppose that $\mu(A) > 0$ if $A \subset \Omega$ is open. If (i) $K_{\boldsymbol{x}} \in L^2(\Omega,\mu)$ for every $\boldsymbol{x} \in \Omega$ and (ii) the integral operator L_K is bounded and positive on $L^2(\Omega,\mu)$ and $L_K g$ is continuous for every $g \in L^2(\Omega,\mu)$, then $\{\lambda_p^{1/2}\varphi_p\}_{p=0}^\infty$ form an orthonormal basis of $\mathcal{H}_K$ and*

$$K(\boldsymbol{x},\boldsymbol{x}') = \sum_{p=0}^\infty \lambda_p \varphi_p(\boldsymbol{x}) \varphi_p(\boldsymbol{x}'), \tag{4.3}$$

with the convergence being absolute and uniform on any compact subset of $\Omega \times \Omega$.

Let $d = 1$ and $\Omega = \mathbb{R}$ and consider the standard Gaussian measure

$$\mathrm{d}\mu(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \,\mathrm{d}x$$

and the Gaussian kernel

$$K(x,x') = \exp\left(-\frac{(x-x')^2}{2\ell^2}\right).$$

The assumptions of Theorem 4.2 are satisfied and the eigensystem is available in closed form (Fasshauer and McCourt, 2012):

$$\lambda_p = \sqrt{\frac{1/2}{1/2+\delta^2+\varepsilon^2}}\left(\frac{\varepsilon^2}{1/2+\delta^2+\varepsilon^2}\right)^p \tag{4.4}$$

and

$$\varphi_p(x) = \sqrt{\frac{\beta}{p!}}\,\mathrm{e}^{-\delta^2x^2}\,\mathrm{H}_p(\beta x), \tag{4.5}$$

where H_p are the (unnormalised) probabilists' Hermite polynomials (i.e., $\langle \mathrm{H}_p, \mathrm{H}_q\rangle_{L^2(\Omega,\mu)} = p!\delta_{pq}$) and

$$\varepsilon = \frac{1}{\sqrt{2}\ell}, \qquad \beta = (1+8\varepsilon^2)^{1/4}, \qquad \text{and} \qquad \delta^2 = \frac{1}{4}(\beta^2-1). \tag{4.6}$$

For explicit computations verifying that this indeed is an orthonormal eigensystem of L_K, see Fasshauer and McCourt (2015, Sec. 12.2.1).

We can now truncate the expansion (4.3) after n terms.[1] This gives

$$\boldsymbol{K}_X \approx \boldsymbol{\Phi}_X \boldsymbol{\Lambda} \boldsymbol{\Phi}_X^{\mathsf{T}} \qquad \text{and} \qquad \boldsymbol{k}_{\mu,X} \approx \boldsymbol{\Phi}_X \boldsymbol{\Lambda} \boldsymbol{\varphi}_\mu,$$

where $[\boldsymbol{\Phi}_X]_{ij} = \varphi_{j-1}(x_i)$ is an $n \times n$ matrix, $\boldsymbol{\Lambda}$ is a diagonal matrix containing the first n eigenvalues, and $[\boldsymbol{\varphi}_\mu]_i = I(\varphi_i)$. Invertibility of $\boldsymbol{\Phi}_X$ follows from invertibility of the classical Vandermonde matrix. The kernel quadrature weights are then approximated as

$$\boldsymbol{w}_K \approx \widetilde{\boldsymbol{w}}_K := \left(\boldsymbol{\Phi}_X \boldsymbol{\Lambda} \boldsymbol{\Phi}_X^{\mathsf{T}}\right)^{-1} \boldsymbol{\Phi}_X \boldsymbol{\Lambda} \boldsymbol{\varphi}_\mu = \boldsymbol{\Phi}_X^{-\mathsf{T}} \boldsymbol{\varphi}_\mu. \tag{4.7}$$

Since the *approximate weights* $\widetilde{\boldsymbol{w}}_K$ solve the linear system $\boldsymbol{\Phi}_X^{\mathsf{T}} \widetilde{\boldsymbol{w}}_K = \boldsymbol{\varphi}_\mu$, rows of which are

$$\sum_{i=1}^{n} \widetilde{w}_{K,i} \varphi_{j-1}(x_i) = I(\varphi_{j-1}) \qquad \text{for} \qquad j = 1,\dots,n,$$

we observe that $\widetilde{\boldsymbol{w}}_K$ define the unique quadrature rule for points X that is exact for the first n eigenfunctions. The approximation (4.7) is further simplified if the points are selected by scaling the points x_i^{GH} of the classical *Gauss–Hermite quadrature rule*. The Gauss–Hermite rule is the unique n-point quadrature rule that satisfies

$$\sum_{i=1}^{n} w_i^{\mathrm{GH}} (x_i^{\mathrm{GH}})^m = I(x^m) \qquad \text{for each} \qquad m = 0,\dots,2n-1.$$

[1] Truncation lengths other than the number of points are possible, but do not result in an attractive closed-form expression for the approximate weights.

Its points are the roots of the nth degree Hermite polynomial H_n. By using the scaling $x_i = x_i^{\mathrm{GH}}/\beta$ the integration points thus become the roots of the nth eigenfunction φ_n. A result of Mysovskikh (1968), for which see also Cools (1997, Sec. 7), that relates $\boldsymbol{\Phi}_X$ to $\boldsymbol{w}^{\mathrm{GH}}$ and an explicit formula for the integrals $I(\varphi_{j-1})$ in (4.7) yield the following result, which is the main contribution of Publication **III**. This result can be extended to higher dimensions by using tensor-product grids.

Theorem 4.3. *Let $\{x_i^{GH}\}_{i=1}^n$ and $\boldsymbol{w}^{GH}$ be the points and weights of the n-point Gauss–Hermite quadrature rule. If the integration points $x_i = x_i^{GH}/\beta$ are used, then the approximate weights in* (4.7) *are*

$$\widetilde{w}_{K,i} = \left(\frac{1}{1+2\delta^2}\right)^{1/2} w_i^{GH} \mathrm{e}^{\delta^2 x_i^2} \sum_{p=0}^{\lfloor (n-1)/2 \rfloor} \frac{1}{2^p p!} \left(\frac{\beta^2}{1+2\delta^2} - 1\right)^p \mathrm{H}_{2p}(x_i^{GH}) \tag{4.8}$$

for $i = 1,\ldots,n$, where the constants β and δ are defined in (4.6) *and H_{2p} are the even probabilists' Hermite polynomials.*

As numerically demonstrated in Publication **III**, the combination of these points and weights has a number of advantageous properties:

- Equation (4.8) provides a numerically stable and accurate approximation to the kernel quadrature weights. The error $\|\boldsymbol{w}_K - \widetilde{\boldsymbol{w}}_K\|$ of the weight approximation decreases as n and ℓ are increased.

- That both $\widetilde{\boldsymbol{w}}_K$ and $\boldsymbol{w}_K$ are positive is verified by numerical experiments up to $n = 100$, but we have not been able to furnish a rigorous proof. This supports the conclusion that the scaled Gauss–Hermite points $x_i = x_i^{\mathrm{GH}}/\beta$ are in some sense "good" for kernel quadrature based on the Gaussian kernel.

- An exponential rate of convergence of both $Q(X,\boldsymbol{w}_K)$ and $Q(X,\widetilde{\boldsymbol{w}}_K)$ for functions in the RKHS $\mathscr{H}_K$ can be proved under the assumption that $\widetilde{\boldsymbol{w}}_K$ are positive. The only other convergence results that hold in this non-compact setting (recall that results in Section 2.3.4 are only for compact domains) we are aware of appear in Kuo and Woźniakowski (2012) and Kuo et al. (2017).

- As $\ell \to \infty$, $x_i \to x_i^{\mathrm{GH}}$ and $\widetilde{\boldsymbol{w}}_K \to \boldsymbol{w}^{\mathrm{GH}}$, which is to be expected based on the results on flat limits cited in Section 5.4.

The main limitation of Theorem 4.3 is that its point sequences are not nested and, as the points depend on the length-scale via the scaling by $1/\beta$, the length-scale cannot be easily selected using one of the data-dependent methods outlined in Section 3.3.

5. Connections to Classical Cubature (Publications IV & V)

Properties of splines, Gaussian quadrature rules, and other classical methods of numerical analysis have been studied extensively, and such methods form a reliable toolkit for approximation. It would therefore be desirable to transform their output into a full non-degenerate probability distribution because a useful extension that retains all positive aspects of a widely used method tends to be more appealing, particularly when it comes to implementation, than a wholly new, untested, and exotic method. Accordingly, this chapter studies different modelling choices that yield Bayesian cubature methods whose posterior means coincide with classical numerical integration methods. In the context of cubature, some aspects of the topic have been discussed by Larkin (1970, Sec. 3); Diaconis (1988, Sec. 1); O'Hagan (1991, Sec. 3.3); Minka (2000); Särkkä et al. (2016, Sec. IV); and Prüher and Särkkä (2016, Thm. 5.2), but no proper overview appears to have been published. Similar questions have also been of great importance in development of probabilistic methods for differential equations (Schober et al., 2014; Teymur et al., 2016; Schober et al., 2018).

The chapter consists of the following parts:

1. Section 5.1 defines what we mean by a polynomial approximation method or cubature rule.

2. Sections 5.2 to 5.4 provide a review of three different methods for reproducing polynomial methods and cubature rules via Gaussian process regression and Bayesian cubature. In particular, Sections 5.2 and 5.3 review the relevant contents of Publications **IV** and **V**, while Section 5.4 discusses an approach that has been extensively studied in kernel interpolation literature.

3. Finally, Section 5.5 reviews some fairly well-known results on the equivalence of spline interpolation and Gaussian process regression and their implications to corresponding quadrature methods, most importantly the trapezoidal rule.

When more convenient, we employ kernel cubature terminology. It is therefore essential to keep the fundamental equivalences (3.5) and (3.6) and the corre-

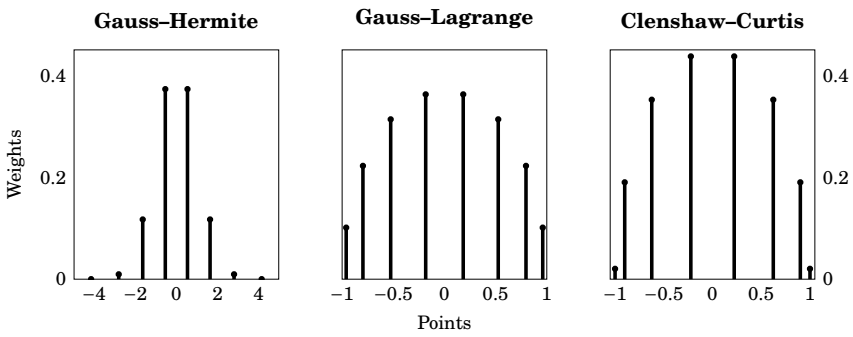

Figure 5.1. Three classical polynomial quadrature rules for $n = 8$: the Gauss–Hermite rule for integration over the real line against the standard Gaussian measure and the Gauss–Lagrange and Clenshaw–Curtis rules for uniform integration over $[-1, 1]$.

spondence between kernel interpolation and Gaussian process regression in mind throughout this chapter.

5.1 Polynomial Approximation Methods

It is a classical result that, for any n distinct points on the real line and any univariate function f defined at these points, there exists a unique *polynomial interpolant* $p_{f,X}$ of degree $n-1$ (this follows from, for example, invertibility of the Vandermonde matrix). Because, by its very definition, this interpolant satisfies the interpolation conditions

$$p_{f,X}(x_i) = \sum_{q=0}^{n-1} a_q x_i^q = f(x_i) \qquad \text{for each} \qquad i = 1, \ldots, n,$$

the coefficient vector $\boldsymbol{a} \in \mathbb{R}^n$ has to solve the linear system

$$\boldsymbol{V}_X \boldsymbol{a} = \boldsymbol{f}_X,$$

where $[\boldsymbol{V}_X]_{ij} = x_i^{j-1}$ is the $n \times n$ invertible *Vandermonde matrix*. In particular, $p_{f,X} = f$ if f is a polynomial of degree at most $n-1$. A natural extension is to replace the polynomials with a different class of functions. Collections of functions for which the natural generalisation of the Vandermonde matrix remains invertible are called *Chebyshev systems* (Karlin and Studden, 1966).

Unfortunately, in higher dimensions it can happen that the natural extension of the Vandermonde matrix for multivariate polynomials is no longer invertible. To guarantee that this does not happen, a unisolvency assumption unnecessary in one dimension is required: a point set on which it is possible to construct unique interpolants is said to be unisolvent.

Definition 5.1 (Unisolvency)**.** Let π be a linear space of real-valued functions on Ω. An n-point set $X \subset \Omega$ is *π-unisolvent* if the zero function is the only function in any subspace $\tilde{\pi} \subset \pi$ of dimension at most n that vanishes at X.

If $\dim(\pi) \geq n$, unisolvency guarantees the existence of a unique interpolant in any $\tilde{\pi} \subset \pi$ such that $\dim(\tilde{\pi}) = n$ to any function defined on X. This is because the requirement that the zero function be the only function in $\tilde{\pi}$ to vanish at X implies that

$$\underbrace{\begin{bmatrix} \varphi_1(\boldsymbol{x}_1) & \cdots & \varphi_n(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\boldsymbol{x}_n) & \cdots & \varphi_n(\boldsymbol{x}_n) \end{bmatrix}}_{=:\boldsymbol{\Phi}_X} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $\{\varphi_i\}_{i=1}^n$ is any basis of $\tilde{\pi}$, which holds if and only if $\boldsymbol{a} = \boldsymbol{0}$. This in turn is equivalent to $\boldsymbol{\Phi}_X \in \mathbb{R}^{n \times n}$ being invertible. This matrix, a generalised version of the Vandermonde matrix above, is called an *alternant matrix*. Hence the function

$$\varphi_{f,X}(\boldsymbol{x}) = \sum_{i=1}^{n} a_i \varphi_i(\boldsymbol{x}), \qquad \boldsymbol{a} = \boldsymbol{\Phi}_X^{-1} \boldsymbol{f}_X,$$

is well-defined, in $\operatorname{span}\{\varphi_1, \ldots, \varphi_n\}$, and satisfies the interpolation condition $\varphi_{f,X}|_X = f|_X$.

By integrating a univariate polynomial interpolant one obtains a unique (for the given points) *polynomial quadrature rule* that is exact for all polynomials up to at least degree $n-1$. That is, the quadrature rule $Q(X, \boldsymbol{w}_P)$ with the weights $\boldsymbol{w}_P = \boldsymbol{V}_X^{-\mathsf{T}} \boldsymbol{p}_\mu$, $[\boldsymbol{p}_\mu]_i = I(x^{i-1})$, satisfies $Q(p; X, \boldsymbol{w}_P) = I(p)$ for every polynomial p of degree at most $n-1$. It is possible that $Q(X, \boldsymbol{w}_P)$ is exact for a larger space of polynomials.

Definition 5.2 (Degree of a quadrature rule)**.** A quadrature rule $Q(X, \boldsymbol{w})$ is *of degree m* if $Q(p; X, \boldsymbol{w}) = I(p)$ for every polynomial p of degree at most m and $Q(p; X, \boldsymbol{w}) \neq I(p)$ for some polynomial p of degree $m+1$.

By judicious choice of the points one can ensure that a quadrature rule is of degree $2n-1$ (which is maximal degree possible with n points). Such a rule is unique for the given measure[1] and called a *Gaussian quadrature rule*. For example, the Gauss–Hermite rule, already discussed in Section 4.3, is the Gaussian quadrature rule for μ the standard Gaussian measure on $\Omega = \mathbb{R}$ and *Gauss–Legendre rule* the Gaussian rule for μ uniform on $\Omega = [-1, 1]$. The *Clenshaw–Curtis rule* (Clenshaw and Curtis, 1960) is one of the most widely used non-Gaussian polynomial rules of degree $n-1$. *Gauss–Patterson rules* are examples of rules of intermediate degree (Patterson, 1968; Genz and Keister, 1996). Some important polynomial quadrature rules are displayed in Figure 5.1. Note that the polynomial weights are rarely solved directly from $\boldsymbol{V}_X^{\mathsf{T}} \boldsymbol{w}_P = \boldsymbol{p}_\mu$. For instance,

[1] See Gautschi (2004, p. 3) for the extremely weak assumptions that μ needs to satisfy.

weights of Gaussian quadrature rules are related to the eigendecomposition of a certain tridiagonal matrix constructed out of recursion coefficient of the three-term recurrence formula of the relevant orthogonal polynomials (Gautschi, 2004, Sec. 3.1.1). Construction of polynomial cubature rules is much more involved in higher dimensions (Cools, 1997), partially because not every $n \in \mathbb{N}$ is a dimension of a full space of multivariate polynomials up to a given degree.

The principal question that Sections 5.2 and 5.3 attempt to answer is whether or not it is possible to construct Bayesian quadrature rules that coincide with polynomial quadrature rules (or other classical cubature rules).

5.2 Bayes–Sard Cubature

This section reviews the main results in Publication **IV** on replicating classical cubature rules by the use of a parametric prior mean model in a way that results in a non-degenerate posterior.

5.2.1 Flat Prior Limit

In Publication **IV**, the prior mean function m in Gaussian process regression is given the parametric form

$$m(\boldsymbol{x}) = \sum_{q=1}^{Q} \theta_q \varphi_q(\boldsymbol{x}) =: \boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x}), \qquad \boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{5.1}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{Q \times Q}$ for $Q \leq n$ is a positive-definite covariance matrix, and the deterministic functions $\varphi_q : \Omega \to \mathbb{R}$ span a Q-dimensional linear function space π. When the coefficients $\boldsymbol{\theta}$ are marginalised out, the GP posterior mean and covariance take the forms

$$m^{\pi}_{X,\boldsymbol{\Sigma}}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{k}_X(\boldsymbol{x}) + \boldsymbol{\beta}^\top \boldsymbol{\varphi}(\boldsymbol{x}), \tag{5.2}$$

$$\begin{aligned} K^{\pi}_{X,\boldsymbol{\Sigma}}(\boldsymbol{x},\boldsymbol{x}') = {} & K(\boldsymbol{x},\boldsymbol{x}') + \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{\varphi}(\boldsymbol{x}) \\ & - \left[\boldsymbol{k}_X(\boldsymbol{x}) - \boldsymbol{\Phi}_X \boldsymbol{\Sigma} \boldsymbol{\varphi}(\boldsymbol{x})\right]^\top \left[\boldsymbol{K}_X + \boldsymbol{\Phi}_X \boldsymbol{\Sigma} \boldsymbol{\Phi}_X^\top\right]^{-1} \left[\boldsymbol{k}_X(\boldsymbol{x}) - \boldsymbol{\Phi}_X \boldsymbol{\Sigma} \boldsymbol{\varphi}(\boldsymbol{x})\right], \end{aligned} \tag{5.3}$$

where $[\boldsymbol{\Phi}_X]_{ij} = \varphi_j(\boldsymbol{x}_i)$ is an $n \times Q$ alternant matrix and the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^Q$ solve the linear system

$$\begin{bmatrix} \boldsymbol{K}_X & \boldsymbol{\Phi}_X \\ \boldsymbol{\Phi}_X^\top & -\boldsymbol{\Sigma}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_X \\ \boldsymbol{0} \end{bmatrix}.$$

Suppose then that the point set X is π-unisolvent. As this implies that the zero function is the only function in π to vanish at X, the matrix $\boldsymbol{\Phi}_X$ has full rank. As $\boldsymbol{\Sigma}^{-1} \to \boldsymbol{0}$, the prior on $\boldsymbol{\theta}$ becomes "weakly informative" and the posterior

mean (5.2) and covariance (5.3) convergence to

$$m_X^\pi(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{k}_X(\boldsymbol{x}) + \boldsymbol{\beta}^\top \boldsymbol{\varphi}(\boldsymbol{x}), \tag{5.4}$$

$$\begin{aligned} K_X^\pi(\boldsymbol{x},\boldsymbol{x}') = K(\boldsymbol{x},\boldsymbol{x}') &- \boldsymbol{k}_X(\boldsymbol{x})^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x}) \\ &+ \left[\boldsymbol{\Phi}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x}) - \boldsymbol{\varphi}(\boldsymbol{x})\right]^\top \left[\boldsymbol{\Phi}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{\Phi}_X\right]^{-1} \left[\boldsymbol{\Phi}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x}) - \boldsymbol{\varphi}(\boldsymbol{x})\right], \end{aligned} \tag{5.5}$$

where the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ now solve

$$\begin{bmatrix} \boldsymbol{K}_X & \boldsymbol{\Phi}_X \\ \boldsymbol{\Phi}_X^\top & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_X \\ \boldsymbol{0} \end{bmatrix}.$$

All relevant matrices above are invertible because $\boldsymbol{\Phi}_X$ has full rank by the unisolvency assumption. For earlier appearances of essentially the same Gaussian process model, see O'Hagan (1978); Wahba (1978); and Santner et al. (2003, Sec. 4.1.2). The posterior mean is also related to interpolation with conditionally positive-definite kernels (Wendland, 2005, Ch. 8).

5.2.2 Construction and Properties of Bayes–Sard Cubature

Integration of the Gaussian process posterior with the mean and covariance functions (5.4) and (5.5) produces the integral mean and variance

$$\mathbb{E}\left[I(f_{\mathrm{GP}}) \mid \boldsymbol{f}_X\right] = \sum_{i=1}^n w_{K,i}^{\mathrm{BSC}} f(\boldsymbol{x}_i) = (\boldsymbol{w}_K^{\mathrm{BSC}})^\top \boldsymbol{f}_X, \tag{5.6}$$

$$\mathbb{V}\left[I(f_{\mathrm{GP}}) \mid \boldsymbol{f}_X\right] = I(K_\mu) - \boldsymbol{k}_{\mu,X}^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\mu,X} + (\boldsymbol{w}_\pi^{\mathrm{BSC}})^\top \left[\boldsymbol{\Phi}_X^\top \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\mu,X} - \boldsymbol{\varphi}_\mu\right], \tag{5.7}$$

where $[\boldsymbol{\varphi}_\mu]_i = I(\varphi_i)$ is a Q-vector and the weights $\boldsymbol{w}_K^{\mathrm{BSC}} \in \mathbb{R}^n$ and $\boldsymbol{w}_\pi^{\mathrm{BSC}} \in \mathbb{R}^Q$ are solved from the linear system

$$\begin{bmatrix} \boldsymbol{K}_X & \boldsymbol{\Phi}_X \\ \boldsymbol{\Phi}_X^\top & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_K^{\mathrm{BSC}} \\ \boldsymbol{w}_\pi^{\mathrm{BSC}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{k}_{\mu,X} \\ \boldsymbol{\varphi}_\mu \end{bmatrix}.$$

We call this probabilistic integration method the *Bayes–Sard cubature* due to its resemblance to a numerical integration method by Sard (1949). An important property of the posterior mean is that it integrates every function in π exactly.

Proposition 5.3. *Suppose that $Q \leq n$ and that X is π-unisolvent. Then the Bayes–Sard posterior mean* (5.6) *coincides with the true integral for every $f \in \pi$:*

$$\mathbb{E}\left[I(f_{GP}) \mid \boldsymbol{f}_X\right] = I(f) \quad \text{if} \quad f \in \pi.$$

A version of Bayesian cubature almost identical to Bayes–Sard cubature has been considered in full generality by Larkin (1974) and O'Hagan (1991). DeVore et al. (2018) work with the equivalent formulation of finding the weights that minimise the worst-case error under the restriction that all functions in π

are integrated exactly. The special case $Q = 1$ and $\varphi_1 \equiv 1$ has been considered by Kennedy (1998); Pronzato and Zhigljavsky (2018); and Rathinavel and Hickernell (2018). In this case the cubature weights are

$$\boldsymbol{w}_K^{\mathrm{BSC}} = \left(\boldsymbol{I} - \frac{\boldsymbol{K}_X^{-1} \mathbb{1}_n \mathbb{1}_n^\mathsf{T}}{\mathbb{1}_n^\mathsf{T} \boldsymbol{K}_X^{-1} \mathbb{1}} \right) \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\mu,X} + \frac{\boldsymbol{K}_X^{-1} \mathbb{1}_n}{\mathbb{1}_n^\mathsf{T} \boldsymbol{K}_X^{-1} \mathbb{1}}$$

and they sum up to one. The central difference between the flat prior formulation we use and previous work is that the coefficients $\boldsymbol{\theta}$ are typically *directly* assigned an improper prior. For example, O'Hagan (1991) equips the kernel with a magnitude parameter σ and uses the prior $p(\sigma^2, \boldsymbol{\theta}) \propto 1/\sigma^2$ (recall Section 3.3.1). This transforms the posterior into Student's t with $n - Q$ degrees of freedom and with variance that involves the factor $(n - Q - 2)^{-1}$. A number of such results for different priors on the parameters are collected in Santner et al. (2003, Thm. 4.1.2). As O'Hagan (1991, Sec. 2.3) already points out, a consequence of the use of this model is that the case $Q = n$ yields an improper posterior distribution. The Bayes–Sard cubature does not suffer from this limitation and, as shown in Publication **IV**, can be thus used in endowing classical cubature rules with a non-zero posterior variance.

Theorem 5.4. *Suppose that $Q = n$ and that X is π-unisolvent. Then*

$$\boldsymbol{w}_K^{BSC} = \boldsymbol{\Phi}_X^{-\mathsf{T}} \boldsymbol{\varphi}_\mu \qquad \text{and} \qquad \mathbb{V}\left[I(f_{GP}) \mid \boldsymbol{f}_X \right] = e_K(X, \boldsymbol{w}_K^{BSC})^2.$$

This theorem says that the squared worst-case error in *any* RKHS of *any* cubature rule whose weights have been (uniquely) selected so as to integrate every function in π exactly can be interpreted as a Bayesian posterior variance. In fact, the same conclusions holds for any cubature rule with non-zero weights. To highlight the somewhat artificial nature of this construction, we also reproduce the corresponding proof from Publication **IV**.

Proposition 5.5. *Suppose that $\Omega \subset \mathbb{R}^d$ has a non-empty interior, $\mu(A) > 0$ whenever $A \subset \Omega$ is open, and $\mu(\{\boldsymbol{x}\}) = 0$ for every $\boldsymbol{x} \in \Omega$. Consider an n-point cubature rule $Q(X, \boldsymbol{w})$ with non-zero weights. Then there exists an n-dimensional function space π such that $\boldsymbol{w} = \boldsymbol{\Phi}_X^{-\mathsf{T}} \boldsymbol{\varphi}_\mu$.*

Proof. By the assumptions there exist disjoint measurable subsets $D_i \subset \Omega$ such that $\boldsymbol{x}_i \in D_i$ and $\mu(D_i) > 0$ for each $i = 1, \ldots, n$. Select the functions

$$\varphi_i = \chi_{D_i \setminus \{\boldsymbol{x}_i\}} + \frac{\mu(D_i)}{w_i} \chi_{\{\boldsymbol{x}_i\}},$$

where χ_A is the characteristic function of the set A. Then $I(\varphi_i) = \mu(D_i)$ and $\boldsymbol{\Phi}_X$ is diagonal with the elements $[\boldsymbol{\Phi}_X]_{ii} = \mu(D_i)/w_i$. It follows that $\boldsymbol{w} = \boldsymbol{\Phi}_X^{-\mathsf{T}} \boldsymbol{\varphi}_\mu$. □

The construction is much more appealing if the points and weights are in some sense "sensible":

- One-dimensional classical polynomial quadrature rules, such as Gaussian rules or the Clenshaw–Curtis rule, can be assigned a probabilistic interpretation by selecting the functions $\varphi_i(x) = x^{i-1}$.

- If μ is a probability measure, uniformly weighted (quasi) Monte Carlo rules are reproduced by selecting $\varphi_i = \chi_{D_i}$ for disjoint sets D_i such that $\boldsymbol{x}_i \in D_i$ and $\mu(D_i) = 1/n$.

In the next section we study if similar reproduction of classical rules is possible by selecting the *kernel* suitably.

5.3 Polynomial Kernels

This sections reviews the results in Publication **V** on using polynomial kernels to construct polynomial quadrature rules. Let $\{p_q\}_{q=0}^{m-1}$ be a collection of polynomials that span the space of univariate polynomials of degree at most $m-1$ and consider the *polynomial kernel*

$$K_m^{\mathrm{pol}}(x,x') \coloneqq \sum_{q=0}^{m-1} p_q(x)p_q(x'). \tag{5.8}$$

Note that this kernel is not strictly speaking positive-definite in that (2.1) holds only for $n \leq m$. The RKHS induced by K_m^{pol} is finite-dimensional, consisting of polynomials of degree at most $m-1$. Observe also that this kernel is essentially equivalent to the one usually called polynomial kernel (e.g., Steinwart and Christmann, 2008, Lem. 4.7),

$$\widetilde{K}_m^{\mathrm{pol}}(x,x') \coloneqq (xx'+1)^{m-1}.$$

Indeed, the multinomial theorem yields

$$\widetilde{K}_m^{\mathrm{pol}}(x,x') = \sum_{q+r=m-1} \frac{(m-1)!}{q!r!}(xx')^q = \sum_{q=0}^{m-1}\left(\sum_{r=0}^{m-1}\frac{(m-1)!}{q!(m-1-r)!}\right)x^q(x')^q,$$

which means that the selection

$$p_q(x) = \left(\sum_{r=0}^{m-1}\frac{(m-1)!}{q!(m-1-r)!}\right)^{1/2} x^q$$

makes K_m^{pol} and $\widetilde{K}_m^{\mathrm{pol}}$ equivalent. The following is the main result of Publication **V**.

Proposition 5.6. *Let $Q(X,\boldsymbol{w}_P)$ be a polynomial quadrature rule of degree $r-1$. Then the kernel quadrature rule $Q(X,\boldsymbol{w}_K)$ based on the kernel* (5.8) *coincides with this rule if $n \leq m \leq r$. In this case, $e_K(X,\boldsymbol{w}_K) = 0$.*

From the identification of the worst-case error as Bayesian quadrature posterior variance we observe the unfortunate fact that classical polynomial quadrature rules are reproduced with zero posterior variance. This renders the kernel (5.8) virtually useless in endowing classical rules with a useful posterior probability distribution. The conclusion is not particularly novel, though a source of explicit analysis seems elusive. For example, Diaconis (1988, p. 164) clearly had something similar in mind: “Is Simpson’s rule Bayes? (Only for priors concentrated on cubic polynomials.)”

5.4 Increasingly Flat Stationary Kernels

A different approach to obtain polynomial methods from kernel-based ones is to consider *increasingly flat kernels*. Recall that stationary kernels can be parametrised by the length-scale parameter $\ell > 0$ such that

$$K_\ell(\boldsymbol{x},\boldsymbol{x}') \coloneqq K_0\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{\ell}\right) \tag{5.9}$$

for some stationary kernel K_0. As $\ell \to \infty$, the kernel becomes increasingly flat and the kernel matrix increasingly ill-conditioned because all its elements converge to $K_0(\mathbf{0})$. However, as the corresponding kernel interpolant, say $s^\ell_{f,X}$, has to interpolate f for each ℓ, it is not entirely clear how it should behave and if its limit exists. Research on this question was initiated by Driscoll and Fornberg (2002).[2] The perhaps surprising conclusions are that

- if the kernel K_0 is isotropic and infinitely smooth and the point set X is unisolvent for polynomial interpolation, the kernel interpolant converges to a polynomial interpolant as $\ell \to \infty$;
- if the kernel K_0 is instead only finitely smooth, convergence is to a polyharmonic spline interpolant.

Infinitely smooth isotropic kernels were initially considered by Driscoll and Fornberg (2002); Fornberg et al. (2004); and Larsson and Fornberg (2005). The results we present in detail below appear in Schaback (2005) and Lee et al. (2007); see also (Schaback, 2008). Finitely smooth kernels are considered in Song et al. (2012) and Lee et al. (2014), with further generalisations appearing in Lee et al. (2015).

For the following theorem recall that the dimension of the space

$$\Pi^d_m \coloneqq \left\{\boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{N}^d_0, \quad |\boldsymbol{\alpha}| \leq m\right\}$$

of d-variate polynomials of degree at most m is

$$M^d_m \coloneqq \dim \Pi^d_m = \frac{(m+d)!}{m!\,d!}.$$

We also need to define the Fourier transform $\widehat{K}_0$ of K_0:

$$\widehat{K}_0(\boldsymbol{\xi}) \coloneqq \int_{\mathbb{R}^d} K_0(\boldsymbol{x})\,\mathrm{e}^{-2\pi i \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\xi}}\,\mathrm{d}\boldsymbol{x}.$$

The kernels K_ℓ are positive-definite on $\mathbb{R}^d$ if $\widehat{K}_0$ is non-negative and not identically zero (Wendland, 2005, Thm. 6.11).

[2] It is interesting to note that the question appears to have been first posed in the context of Bayesian cubature already by O'Hagan (1991, Sec. 3.3). Also Minka (2000) and Särkkä et al. (2016) discuss this conjecture.

Theorem 5.7. *Let $m \in \mathbb{N}$ be such that $M_{m-1}^d < n \leq M_m^d$. Suppose that the kernel is of the form*

$$K_\ell(\boldsymbol{x}, \boldsymbol{x}') = \sum_{q=0}^{\infty} a_q \ell^{-2q} \|\boldsymbol{x} - \boldsymbol{x}'\|^{2q}$$

for some coefficients $a_p \in \mathbb{R}$ and that the Fourier transform $\widehat{K}_0$ is positive on an open subset of $\mathbb{R}^d$. If X is Π_m^d-unisolvent and

(i) $n = M_m^d$, then $\lim_{\ell \to \infty} s_{f,X}^\ell = p$ for the unique *polynomial $p \in \Pi_m^d$ such that $p|_X = f|_X$;*

(ii) $n < M_m^d$, then $\lim_{\ell \to \infty} s_{f,X}^\ell = p$ for some *polynomial $p \in \Pi_m^d$ such that $p|_X = f|_X$.*

If the kernel is Gaussian, $\lim_{\ell \to \infty} s_{f,X}^\ell$ exists and is a polynomial interpolant for any *X.*

The statements for general kernels appear in Theorem 3.5 of Lee et al. (2007) while the special case of a Gaussian kernel can be found in Theorem 2 of Schaback (2005). In the latter case the limiting polynomial interpolant coincides with the *de Boor and Ron interpolant* (de Boor and Ron, 1990, 1992). Because kernel cubature rules are just integrated kernel interpolants, Theorem 5.7 ought to imply, at least for compact domains, that the increasingly flat limit in the case $n = M_m^d$ is a polynomial cubature rule of degree (at least) m. Unfortunately, as with the construction based on polynomial kernels in Section 5.3, the posterior variance (i.e., the squared power function) vanishes at the limit $\ell \to \infty$. We have not found this straightforward result anywhere in the literature and accordingly supply its proof.

Proposition 5.8. *Let $m \in \mathbb{N}$ be such that $M_{m-1}^d < n \leq M_m^d$. Suppose that there exists a unique polynomial interpolant $p \in \Pi_m^d$ to f at X and that $s_{f,X}^\ell$ converges to this interpolant as $\ell \to \infty$. Then $\lim_{\ell \to \infty} P_X^\ell = 0$ for the corresponding power function* (2.20)*.*

Proof. The squared power function is

$$P_X^\ell(\boldsymbol{x})^2 = K_\ell(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_X^\ell(\boldsymbol{x})^\mathsf{T} (\boldsymbol{K}_X^\ell)^{-1} \boldsymbol{k}_X^\ell(\boldsymbol{x}),$$

with dependence on ℓ made explicit. Since $\boldsymbol{u}_X^\ell = (\boldsymbol{K}_X^\ell)^{-1} \boldsymbol{k}_X^\ell(\boldsymbol{x})$ are the Lagrange cardinal functions, which satisfy $u_{X,i}^\ell(\boldsymbol{x}_j) = \delta_{ij}$, and the assumptions imply that their limits are unique polynomials having the same property, we have

$$\begin{aligned}
\lim_{\ell \to \infty} P_X^\ell(\boldsymbol{x})^2 &= \lim_{\ell \to \infty} \left[K_\ell(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}_X^\ell(\boldsymbol{x})^\mathsf{T} (\boldsymbol{K}_X^\ell)^{-1} \boldsymbol{k}_X^\ell(\boldsymbol{x}) \right] \\
&= K_0(\boldsymbol{0}) - \sum_{i=1}^{n} \lim_{\ell \to \infty} K_\ell(\boldsymbol{x}_i, \boldsymbol{x}) u_{X,i}^\ell(\boldsymbol{x}) \\
&= K_0(\boldsymbol{0}) - K_0(\boldsymbol{0}) \sum_{i=1}^{n} \lim_{\ell \to \infty} u_{X,i}^\ell(\boldsymbol{x}).
\end{aligned}$$

Because $\sum_{i=1}^{n} u_{X,i}^{\ell}(\boldsymbol{x})$ interpolates 1 at X for any $\ell > 0$, so does its limit. As the constant polynomial shares this property and polynomial interpolation at X has been assumed unique, it follows that $\sum_{i=1}^{n} u_{X,i}^{\ell}(\boldsymbol{x}) \to 1$ and consequently $P_X^{\ell} \to 0$ as $\ell \to \infty$. □

It is possible that the posterior variance can be prevented from vanishing at the limit by introducing an appropriate length-scale dependent scaling of the kernel.

5.5 Spline-Based Methods

Splines are a classical method of numerical analysis, going back to the work of Schoenberg (1946, 1964). Let $0 < x_1 < \cdots < x_n < 1$. The *natural spline interpolant of degree* $2m+1$, s_{2m+1}, is a sufficiently smooth piecewise polynomial interpolant to f having the following properties:

(i) restricted on $[0, x_1]$ or $[x_n, 1]$, s_{2m+1} is a polynomial of degree m;

(ii) restricted on any of the $n-1$ intervals $[x_i, x_{i+1}]$, s_{2m+1} is a polynomial of degree $2m+1$;

(iii) $s_{2m+1} \in C^{2m}([0,1])$.

A total of $2(m+1)n$ coefficients are needed to define s_{2m+1}: $2(m+1)$ come from s_{2m+1} being of degree m on $[0, x_1]$ and $[x_n, 1]$ and $(n-1)(2m+2)$ from it being of degree $2m+1$ on each of the $n-1$ inner intervals. The interpolation condition $s_{2m+1}|_X = f|_X$ fixes n of these coefficients while the smoothness condition (iii) imposes $(2m+1)n$ additional constraints; s_{2m+1} is thus uniquely defined by properties (i)–(iii).

It can be shown that the posterior mean (5.4) constructed out of the m times integrated Brownian motion kernel (2.8) and with appropriately selected function space π is the natural spline interpolant of degree $2m+1$. The origins of this result appear to be somewhat obscure; see the discussion by Diaconis and Freedman (1983, p. 110); Diaconis (1988, Ex. 2); and Lee and Wasilkowski (1986, Rmk. 5.1). The most accessible explicit proof we have found is by Wahba (1990, Sec. 1.3).

Theorem 5.9. *Let* $0 < x_1 < \cdots < x_n < 1$ *and consider the* $m \leq n-1$ *times integrated Brownian motion kernel* K_m *in* (2.8) *and the flat prior limit model of Section 5.2.1. If* $Q = m+1$ *and* $\varphi_q(x) = x^{q-1}$ *for* $q = 1, \ldots, m+1$*, then the posterior mean* m_X^{π} *in* (5.4) *is the natural spline interpolant of degree* $2m+1$.

A similar result is true when the parametric prior mean (5.1) is not used if property (i) of the natural spline is modified by requiring that $s_{2m+1}^{(q)}(0) = 0$ for $q = 0, \ldots, m$ (Lee and Wasilkowski, 1986, Sec. 5.3). Thus we see that natural splines can be conveniently interpreted as posterior means of Gaussian processes

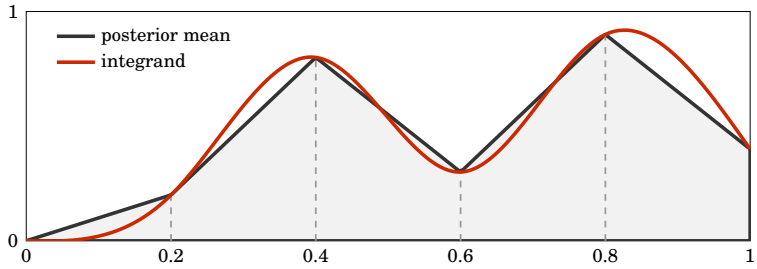

Figure 5.2. The trapezoidal rule is recovered as the integral of the Gaussian process posterior mean with the Brownian motion kernel K_0 if $f(0) = 0$ and $x_n = 1$.

with integrated Brownian motion kernels. Further results exist also for other types of splines (Kimeldorf and Wahba, 1970a,b, 1971).

In quadrature, $m = 0$ is probably one of the most interesting cases because the resulting kernel quadrature rule coincides with the *trapezoidal rule* if the points are selected appropriately. The trapezoidal rule for approximation of $\int_0^1 f(x)\mathrm{d}x$ using points $0 =: x_0 < x_1 < \cdots < x_n \leq 1$ is

$$\begin{aligned} Q_{\mathrm{tra}}(f) &:= \sum_{i=1}^{n} \frac{f(x_{i-1}) + f(x_i)}{2}(x_i - x_{i-1}) \\ &= \frac{x_1}{2} f(0) + \sum_{i=1}^{n-1} \frac{x_{i+1} - x_{i-1}}{2} f(x_i) + \frac{x_n - x_{n-1}}{2} f(x_n), \end{aligned} \tag{5.10}$$

which is just the integral of the piecewise linear interpolant to f, as depicted in Figure 5.2. The following result can be found in Ritter (2000, Ch. II, Sec. 3), though its origins are in the work of Suldin (1959, 1960).[3]

Proposition 5.10. *Consider the Brownian motion kernel* $K_0(x, x') = \min\{x, x'\}$. *If* $f\colon [0,1] \to \mathbb{R}$ *satisfies* $f(0) = 0$, *then kernel quadrature rule with points* $X = \{x_1, \ldots, x_n\}$, $x_n = 1$, *for approximation of* $\int_0^1 f(x)\mathrm{d}x$ *coincides with the trapezoidal rule* (5.10).

Because the corresponding worst-case error is (for this formula, see Ritter, 2000, Ch. II, Sec. 3.3)

$$e_{K_0}(X, \boldsymbol{w}_K)^2 = \frac{1}{12} \sum_{i=1}^{n} (x_i - x_{i-1})^3,$$

and hence non-zero, Proposition 5.10 shows that the trapezoidal rule can be seen as a Bayesian quadrature rule with non-degenerate posterior variance. Of particular interest is the uniform point selection $x_i = 2i/(2n+1)$ that produces a uniformly weighted rule,

$$Q_{K_0}(f; X) = \frac{2}{2n+1} \sum_{i=1}^{n} f\left(\frac{2i}{2n+1}\right).$$

[3] Although these references are concerned with the average-case setting, the result we cite is valid also in the worst-case setting.

6. Summary and Discussion

We conclude this overview with a brief summary and assessment of the significance of the five publications and a discussion on some of the central challenges that, we believe, are currently holding up more widespread adoption of kernel and Bayesian cubature methods.

6.1 Summary and Assessment of Publications

This section briefly summarises the main contributions of the publication that make up this thesis and evaluates their significance.

Publication I (Section 4.2). This publication develops a method for efficient computation of kernel cubature rules that use fully symmetric point sets. The method computes the cubature weights exactly and remains computationally feasible even for up to millions of points. Its main advantage over alternatives exploiting tensor and sparse grids is the relative flexibility with which the points can be selected because fully symmetric constituent sets can be added and removed at will. Although the methodology is not fully mature yet, this approach appears very promising in enabling kernel cubature for hitherto infeasible large point sets that often arise in, for example, financial applications. The main issues, that we are currently working on, are related to efficient computation of the kernel hyperparameters, automatic selection of the fully symmetric sets, and reliable uncertainty quantification in high dimensions.

Publication II (Section 4.2). This publication contains a number of extensions of the fully symmetric computational methods developed in Publication **I**. The main extensions are for (i) the Bayes–Sard cubature proposed in Publication **IV**, (ii) the multi-output Bayesian cubature method proposed by Xi et al. (2018), and (iii) non-symmetric measures via a change of measure trick. The extension for multi-output Bayesian cubature, naive implementation of which does not scale well with the number of points and integrands, seems particularly promising as it enables simultaneous computation of a huge number of related integrals.

Publication III (Section 4.3). This publication shows how the weights of one-dimensional kernel quadrature for the Gaussian kernel and measure can be approximated in an accurate, explicit, and computationally stable manner if the points are selected via a scaling of the points of the classical Gauss–Hermite quadrature rule. An exponential rate of convergence is proved for integrands in the Gaussian RKHS under the assumption that the sum of absolute values of the approximate weights does not grow too fast. For the most part, we consider the weight approximation an interesting closed-form curiosity. The fact the scaled Gauss–Hermite points are, to our knowledge, the "best" ones proposed for this kernel quadrature problem suggests that the results of the publication may have some practical implications in the future.

Publication IV (Section 5.2). This publication proposes the use of a finite-dimensional parametric mean model that is marginalised out such that exactness conditions akin to those appearing in polynomial cubature are incorporated into the resulting Bayesian cubature rule. Even though some of its parts have been proposed before (O'Hagan, 1978; Wahba, 1978), the model has not seen sufficient use in Bayesian cubature. In general, the parametric prior mean model makes the integral estimates of Bayesian cubature more robust against improper selection of the kernel, which is useful especially in high dimensions. The main novel contribution of the publication is perhaps the interesting probabilistic interpretation (Theorem 5.4 and Proposition 5.5) of *any* cubature rule as a Bayesian cubature rule such that the worst-case error corresponds to the posterior standard deviation. Even though selection of the prior via the kernel remains a challenge, it has been already demonstrated that this interpretation can be useful (Prüher et al., 2018).

Publication V (Section 5.3). The main contribution of this publication is an explicit discussion of how polynomial kernels can be used to interpret classical polynomial quadrature rules as Bayesian quadrature rules. The gist of the technical contributions of the article appears to be part of the folklore of the field (see e.g. Diaconis, 1988, p. 164), though we have been unable to locate explicit derivations. As the resulting Bayesian quadrature rules are of zero posterior variance, the main result (Proposition 5.6), being incapable of providing useful uncertainty quantification, is largely a curiosity.

6.2 Challenges

Finally, it seems proper to discuss some partially unresolved challenges in kernel and Bayesian cubature.

6.2.1 Computational Cost

As discussed already Chapter 4, the cubic computational cost in the number of points of kernel cubature places it at a significant disadvantage compared to standard methods that introduce either negligible computational overhead (Monte Carlo and quasi Monte Carlo) or can mostly rely on pre-computed weights (Gaussian quadratures). The problem is further exacerbated by the frequent need to fit the kernel length-scale parameter using maximum likelihood or some other methods. In the case of maximum likelihood, costly optimisation of (3.7) is required.

Fortunately, there already exist several methods for (partially) efficient computation (Publications **I** & **II**; Rathinavel and Hickernell 2018) and many of the fast Gaussian process methods developed in machine learning literature have not been adequately tested in the cubature setting. However, it appears to us that there are fundamental limits to what can be achieved. Namely, computationally competitive *non-approximate* algorithms for weight computation are feasible only if the point sets admit some structurality (e.g., symmetricity or being quasi Monte Carlo sets) that can be exploited. When the point sets are unstructured, *approximate* algorithms are certainly possible, but it is not clear if their computational overhead and weight approximation error can be balanced in a satisfactory way.

6.2.2 Kernel Means

To compute the kernel cubature weights and the worst-case error one needs to be able to evaluate the integrals

$$K_\mu(\boldsymbol{x}_i) = \int_\Omega K(\boldsymbol{x}_i, \boldsymbol{x}) \,\mathrm{d}\mu(\boldsymbol{x}) \qquad \text{and} \qquad I(K_\mu) = \int_\Omega \int_\Omega K(\boldsymbol{x}', \boldsymbol{x}) \,\mathrm{d}\mu(\boldsymbol{x}) \,\mathrm{d}\mu(\boldsymbol{x}'),$$

former of which is the kernel mean. That these integrals may not be available in closed form has been long recognised as a fundamental practical problem (O'Hagan, 1992, Sec. 3.2). There are a number of kernel-measure pairs for which the integrals are analytically tractable (e.g., Briol et al., 2019, Sec. 4.2), but the measure at hand is often not a member of this class and may be accessible only via sampling. Numerical approximation of the integrals, though possible and suggested by Sommariva and Vianello (2006b, Sec. 2); Tronarp et al. (2018); and Briol et al. (2019, Appx. B), can be as difficult a problem as the original integration problem and introduces an additional level of numerical approximation that may need to be modelled if one is interested in principled uncertainty quantification by the means of Bayesian cubature.

Although there are approaches to tackle or circumvent the need to compute kernel means using stick breaking (Oates et al., 2017b) and Stein kernels (Barp et al., 2018), these tend to be of limited utility or muddle the interpretability of the prior model. It seems to us that, if they are to be computationally competitive and statistically interpretable, kernel and Bayesian cubature need to

remain restricted to "traditional" integration problems, such as computation of $\int_{[0,1]^d} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$, instead of, say, involved posterior integration problems arising in statistics.

6.2.3 Uncertainty Calibration

Perhaps the most fundamental question in probabilistic numerics is whether or not the uncertainty quantification a probabilistic numerical method provides for the true value of the quantity of interest is in some sense meaningful. Unfortunately, there exists little theory; only evidence that the uncertainty quantification in terms of, say, coverage properties of Bayesian credible intervals is sensible comes from a limited number of numerical experiments. Because, by necessity, the form of the kernel is typically fixed beforehand, the question of uncertainty calibration is intricately linked to the behaviour of kernel hyperparameter estimates discussed in Section 3.3. The only relevant result we are aware of is due to Xu and Stein (2017) who conjecture that

$$\sigma^2_{\mathrm{ML}} \sim \frac{\ell^{2m}}{\sqrt{2\pi} 2^m (m + 1/2)} n^{m-1/2}$$

when the kernel is Gaussian with length-scale $\ell > 0$, the n points are placed uniformly on $[0, 1]$, and $f(x) = x^m$ for $m \in \mathbb{N}_0$ (they prove this fully for $m = 0$ and partially for $m = 1$).

The relatively large body of literature on the behaviour of maximum likelihood estimates for related Gaussian process regression problems (van der Vaart and van Zanten, 2011; Szabó et al., 2013, 2015; Hadji and Szabó, 2019) leads us to believe that it is only a matter of time before results on consistency and calibration of Gaussian process estimators with noise-free evaluations begin to appear. Based on the regression results we expect, quite reasonably, that uncertainty calibration will be, at least asymptotically, very sensitive to model misspecification. Probabilistic parameter fitting methods such as maximum likelihood cannot be expected to fare well unless the true function can be seen as a realisation of the underlying Gaussian process; this class is small and it is practically impossible to confirm whether or not an integrand, which may be an output of a complicated computer simulation lacking closed-form expression, is a member of this class.

References

Acerbi, L. (2018a). An exploration of acquisition and mean functions in variational Bayesian Monte Carlo. In *1st Symposium on Advances in Approximate Bayesian Inference*.

Acerbi, L. (2018b). Variational Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 31, pages 8213–8223.

Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer.

Ajne, B. and Dalenius, T. (1960). Några tillämpningar av statistiska idéer på numerisk integration. *Nordisk Matematisk Tidskrift*, 8(4):145–152.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(19):1–38.

Barp, A., Oates, C. J., Porcu, E., and Girolami, M. (2018). A Riemannian–Stein kernel method. Preprint, arXiv:1810.04946v1.

Barrar, R. B. and Loeb, H. L. (1975). Multiple zeros and applications to optimal linear functionals. *Numerische Mathematik*, 25(3):251–262.

Barrar, R. B., Loeb, H. L., and Werner, H. (1974). On the existence of optimal integration formulas for analytic functions. *Numerische Mathematik*, 23(2):105–117.

Bartels, S., Cockayne, J., Ipsen, I. C. F., Girolami, M., and Hennig, P. (2018). Probabilistic linear solvers: A unifying view. Preprint, arXiv:1810.03398.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

Bezhaev, A. Yu. (1991). Cubature formulae on scattered meshes. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 6(2):95–106.

Bogachev, V. I. (1998). *Gaussian Measures*. Number 62 in Mathematical Surveys ad Monographs. American Mathematical Society.

Bojanov, B. D. (1979). On the existence of optimal quadrature formulae for smooth functions. *Calcolo*, 16(1):61–70.

Bojanov, B. D. (1994). Optimal recovery of functions and integrals. In *First European Congress of Mathematics*, pages 371–390. Birkhäuser Basel.

Brauchart, J. S. and Dick, J. (2012). Quasi–Monte Carlo rules for numerical integration over the unit sphere $\mathbb{S}^2$. *Numerische Mathematik*, 121(3):473–502.

Brauchart, J. S., Saff, E. B., Sloan, I. H., and Womersley, R. S. (2014). QMC designs: Optimal order quasi Monte Carlo integration schemes on the sphere. *Mathematics of Computation*, 83(290):2821–2851.

Briol, F.-X. (2018). *Statistical Computation with Kernels*. PhD thesis, Department of Statistics, University of Warwick.

Briol, F.-X., Oates, C. J., Cockayne, J., Chen, W. Y., and Girolami, M. (2017). On the sampling problem for kernel quadrature. In *34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 586–595.

Briol, F.-X., Oates, C. J., Girolami, M., and Osborne, M. A. (2015). Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, volume 25, pages 1162–1170.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019). Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22.

Brouillat, J., Bouville, C., Loos, B., Hansen, C., and Bouatouch, K. (2009). A Bayesian Monte Carlo approach to global illumination. *Computer Graphics Forum*, 28(8):2315–2329.

Chai, H. and Garnett, R. (2018). An improved Bayesian framework for quadrature of constrained integrands. Preprint, arXiv:1802.04782v2.

Chai, H., Ton, J.-F., Garnett, R., and Osborne, M. A. (2019). Automated model selection with Bayesian quadrature. Preprint, arXiv:1902.09724v2.

Chernih, A., Sloan, I. H., and Womersley, R. S. (2014). Wendland functions with increasing smoothness converge to a Gaussian. *Advances in Computational Mathematics*, 40(1):185–200.

Clenshaw, C. W. and Curtis, A. R. (1960). A method for numerical integration on an automatic computer. *Numerische Mathematik*, 2(1):197–205.

Cockayne, J., Oates, C., Sullivan, T., and Girolami, M. (2017). Probabilistic numerical methods for partial differential equations and Bayesian inverse problems. Preprint, arXiv:1605.07811v3.

Cockayne, J., Oates, C., Sullivan, T., and Girolami, M. (2019a). Bayesian probabilistic numerical methods. *SIAM Review*. To appear.

Cockayne, J., Oates, C. J., Ipsen, I. C. F., and Girolami, M. (2019b). A Bayesian conjugate gradient method. *Bayesian Analysis*. To appear.

Cook, T. D. and Clayton, M. K. (1998). Sequential Bayesian quadrature. Technical report, Department of Statistics, University of Wisconsin.

Cools, R. (1997). Constructing cubature formulae: the science behind the art. *Acta Numerica*, 6:1–54.

de Boor, C. and Ron, A. (1990). On multivariate polynomial interpolation. *Constructive Approximation*, 6(3):287–302.

de Boor, C. and Ron, A. (1992). The least solution for the polynomial interpolation problem. *Mathematische Zeitschrift*, 210(1):347–378.

De Marchi, S. and Schaback, R. (2009). Nonstandard kernels and their applications. *Dolomites Research Notes on Approximation*, 2(1):16–43.

De Marchi, S. and Schaback, R. (2010). Stability of kernel-based interpolation. *Advances in Computational Mathematics*, 32(2):155–161.

Deisenroth, M. P., Huber, M. F., and Hanebeck, U. D. (2009). Analytic moment-based Gaussian process filtering. In *26th International Conference on Machine Learning*, pages 225–232.

Deisenroth, M. P., Turner, R. D., Huber, M. F., Hanebeck, U. D., and Rasmussen, C. E. (2012). Robust filtering and smoothing with Gaussian processes. *IEEE Transactions on Automatic Control*, 57(7):1865–1871.

DeVore, R., Foucart, S., Petrova, G., and Wojtaszczyk, P. (2018). Computing a quantity of interest from observational data. *Constructive Approximation*.

Diaconis, P. (1988). Bayesian numerical analysis. In *Statistical decision theory and related topics IV*, volume 1, pages 163–175. Springer-Verlag New York.

Diaconis, P. and Freedman, D. (1983). Frequency properties of Bayes rules. In *Scientific Inference, Data Analysis, and Robustness*, pages 105–115. Academic Press.

Dick, J. (2006). A Taylor space for multivariate integration. *Monte Carlo Methods and Applications*, 12(2):99–112.

Dick, J., Kuo, F. Y., and Sloan, I. H. (2013). High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288.

Dick, J. and Pillichshammer, F. (2010). *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press.

Driscoll, T. A. and Fornberg, B. (2002). Interpolation in the limit of increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 43(3–5):413–422.

Ehler, M., Gräf, M., and Oates, C. J. (2019). Optimal Monte Carlo integration on closed manifolds. *Statistics and Computing*. To appear.

Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Number 20 in Oxford Statistical Science Series. Oxford University Press.

Fasshauer, G. and McCourt, M. (2015). *Kernel-based Approximation Methods Using MATLAB*. Number 19 in Interdisciplinary Mathematical Sciences. World Scientific Publishing.

Fasshauer, G. E. (2007). *Meshfree Approximation Methods with MATLAB*. Number 6 in Interdisciplinary Mathematical Sciences. World Scientific Publishing.

Fasshauer, G. E. (2011). Positive definite kernels: past, present and future. *Dolomite Research Notes on Approximation*, 4:21–63.

Fasshauer, G. E. and McCourt, M. J. (2012). Stable evaluation of Gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762.

Fitzsimons, J., Cutajar, K., Osborne, M., Roberts, S., and Filippone, M. (2017). Bayesian inference of log determinants. In *32nd Conference on Uncertainty in Artificial Intelligence*.

Fornberg, B., Wright, G., and Larsson, E. (2004). Some observations regarding interpolants in the limit of flat radial basis functions. *Computers & Mathematics with Applications*, 47(1):37–55.

Fuselier, E., Hangelbroek, T., Narcowich, F. J., Ward, J. D., and Wright, G. B. (2014). Kernel based quadrature on spheres and other homogeneous spaces. *Numerische Mathematik*, 127(1):57–92.

Gautschi, W. (2004). *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press.

Gavrilov, A. V. (1998). On best quadrature formulas in the reproducing kernel Hilbert space. *Sibirsky Zhurnal Vychislitelnoy Matematiki*, 1(4):313–320. In Russian.

Gavrilov, A. V. (2007). On optimal quadrature formulas. *Journal of Applied and Industrial Mathematics*, 1(2):190–192.

Genz, A. (1986). Fully symmetric interpolatory rules for multiple integrals. *SIAM Journal on Numerical Analysis*, 23(6):1273–1283.

Genz, A. and Keister, B. D. (1996). Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. *Journal of Computational and Applied Mathematics*, 71(2):299–309.

Gessner, A., Gonzalez, J., and Mahsereci, M. (2019). Active multi-information source Bayesian quadrature. Preprint, arXiv:1903.11331v1.

Gunter, T., Osborne, M. A., Garnett, R., Hennig, P., and Roberts, S. J. (2014). Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, volume 24, pages 2789–2797.

Hadji, A. and Szabó, B. (2019). Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? Preprint, arXiv:1904.01383v1.

Hamrick, J. B. and Griffiths, T. L. (2013). Mental rotation as Bayesian quadrature. In *NIPS 2013 Workshop on Bayesian Optimization in Theory and Practice*.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *20th IEEE International Workshop on Machine Learning for Signal Processing*, pages 379–384.

Hennig, P. (2015). Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25(1):234–260.

Hennig, P., Osborne, M. A., and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2179).

Hensman, J., Durrande, N., and Solin, A. (2018). Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52.

Hickernell, F. J. (1998). A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322.

Huszár, F. and Duvenaud, D. (2012). Optimally-weighted herding is Bayesian quadrature. In *28th Conference on Uncertainty in Artificial Intelligence*, pages 377–385.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. Preprint, arXiv:1807.02582v1.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2016). Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, volume 29, pages 3288–3296.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2019). Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*.

Karlin, S. (1968). *Total Positivity*, volume 1. Stanford University Press.

Karlin, S. and Studden, W. J. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics*. Interscience Publishers.

Karvonen, T., Kanagawa, M., and Särkkä, S. (2019). On the positivity and magnitudes of Bayesian quadrature weights. *Statistics and Computing*. To appear.

Kennedy, M. (1998). Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8(4):365–375.

Kennedy, M. (2000). A Bayesian approach to computing posterior distribution and quantile functions. *Journal of Statistical Planning and Inference*, 83(1):183–201.

Kennedy, M. C. and O'Hagan, A. (1996). Iterative rescaling for Bayesian quadrature. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5–9, 1994*, pages 639–646. Clarendon Press.

Kersting, H. and Hennig, P. (2016). Active uncertainty calibration in Bayesian ODE solvers. In *Uncertainty in Artificial Intelligence*, volume 32, pages 309–318.

Kimeldorf, G. S. and Wahba, G. (1970a). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Kimeldorf, G. S. and Wahba, G. (1970b). Spline functions and stochastic processes. *Sankhyā: The Indian Journal of Statistics, Series A*, 32(2):173–180.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

Kumar, A., Nair, P. B., Keane, A. J., and Shahpar, S. (2008). Robust design using Bayesian Monte Carlo. *International Journal for Numerical Methods in Engineering*, 73(11):1497–1517.

Kuo, F. Y., Sloan, I. H., and Woźniakowski, H. (2017). Multivariate integration for analytic functions with Gaussian kernels. *Mathematics of Computation*, 86:829–853.

Kuo, F. Y. and Woźniakowski, H. (2012). Gauss–Hermite quadratures for functions from Hilbert spaces with Gaussian reproducing kernels. *BIT Numerical Mathematics*, 52(2):425–436.

Larkin, F. M. (1969). Estimation of a non-negative function. *BIT Numerical Mathematics*, 9(1):30–52.

Larkin, F. M. (1970). Optimal approximation in Hilbert spaces with reproducing kernel functions. *Mathematics of Computation*, 24(112):911–921.

Larkin, F. M. (1972). Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2(3):379–422.

Larkin, F. M. (1974). Probabilistic error estimates in spline interpolation and quadrature. In *Information Processing 74: Proceedings of IFIP Congress 74*, pages 605–609. North-Holland.

Larsson, E. and Fornberg, B. (2005). Theoretical and computational aspects of multivariate interpolation with increasingly flat radial basis functions. *Computers & Mathematics with Applications*, 49(1):103–130.

Lee, D. and Wasilkowski, G. W. (1986). Approximation of linear functionals on a Banach space with a Gaussian measure. *Journal of Complexity*, 2(1):12–43.

Lee, Y. J., Micchelli, C. A., and Yoon, J. (2014). On convergence of flat multivariate interpolation by translation kernels with finite smoothness. *Constructive Approximation*, 40(1):37–60.

Lee, Y. J., Micchelli, C. A., and Yoon, J. (2015). A study on multivariate interpolation by increasingly flat kernel functions. *Journal of Mathematical Analysis and Applications*, 427(1):74–87.

Lee, Y. J., Yoon, G. J., and Yoon, J. (2007). Convergence of increasingly flat radial basis interpolants to polynomial interpolants. *SIAM Journal on Mathematical Analysis*, 39(2):537–553.

Lukić, M. N. and Beder, J. H. (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969.

Lyness, J. N. (1965). Symmetric integration rules for hypercubes I. Error coefficients. *Mathematics of Computation*, 19(90):260–276.

Ma, Y., Garnett, R., and Schneider, J. (2014). Active area search via Bayesian quadrature. In *Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 595–603.

Marques, R., Bouville, C., Ribardière, M., Santos, L. P., and Bouatouch, K. (2013). A spherical Gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1619–1932.

Marques, R., Bouville, C., Santos, L. P., and Bouatouch, K. (2015). *Efficient quadrature rules for illumination integrals: from Quasi Monte Carlo to Bayesian Monte Carlo*. Synthesis Lectures on Computer Graphics and Animation. Morgan & Claypool Publishers.

Matérn, B. (1960). Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden från statens skogforskninginstitut*, 49(5).

McNamee, J. and Stenger, F. (1967). Construction of fully symmetric numerical integration formulas. *Numerische Mathematik*, 10(4):327–344.

Minh, H. Q. (2010). Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338.

Minka, T. (2000). Deriving quadrature rules from Gaussian processes. Technical report, Statistics Department, Carnegie Mellon University.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141.

Mysovskikh, I. P. (1968). On the construction of cubature formulas with fewest nodes. *Soviet Mathematics Doklady*, 9:277–280.

Narcowich, F. J. and Ward, J. D. (2004). Scattered-data interpolation on $\mathbb{R}^n$: Error estimates for radial basis and band-limited functions. *SIAM Journal on Mathematical Analysis*, 36(1):284–300.

Narcowich, F. J., Ward, J. D., and Wendland, H. (2006). Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24(2):175–186.

Novak, E. (1988). *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Number 1349 in Lecture Notes in Mathematics. Springer-Verlag.

Novak, E. and Ritter, K. (1996). High dimensional integration of smooth functions over cubes. *Numerische Mathematik*, 75(1):79–97.

Novak, E. and Ritter, K. (1997). *The Curse of Dimension and a Universal Method for Numerical Integration*, pages 177–187. Springer.

Novak, E. and Woźniakowski, H. (2008). *Tractability of Multivariate Problems, Volume I: Linear Information*. Number 6 in EMS Tracts in Mathematics. European Mathematical Society.

Novak, E. and Woźniakowski, H. (2010). *Tractability of Multivariate Problems, Volume II: Standard Information for Functionals*. Number 12 in EMS Tracts in Mathematics. European Mathematical Society.

Oates, C. J., Cockayne, J., Aykroyd, R. G., and Girolami, M. (2019a). Bayesian probabilistic numerical methods in time-dependent state estimation for industrial hydrocyclone equipment. *Journal of the American Statistical Association*.

Oates, C. J., Cockayne, J., Prangle, D., Sullivan, T. J., and Girolami, M. (2019b). Optimality criteria for probabilistic numerical methods. In *Multivariate Algorithms and Information-Based Complexity*. De Gruyter. To appear.

Oates, C. J., Girolami, M., and Chopin, N. (2017a). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718.

Oates, C. J., Niederer, S., Lee, A., Briol, F.-X., and Girolami, M. (2017b). Probabilistic models for integration error in the assessment of functional cardiac models. In *Advances in Neural Information Processing Systems*, volume 30, pages 110–118.

Oates, C. J. and Sullivan, T. J. (2019). A modern retrospective on probabilistic numerics. *Statistics and Computing*. To appear.

Oettershagen, J. (2017). *Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification*. PhD thesis, Institut für Numerische Simulation, Universität Bonn.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42.

O'Hagan, A. (1988). Bayesian quadrature. Warwick statistics research report 159, Department of Statistics, University of Warwick.

O'Hagan, A. (1991). Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260.

O'Hagan, A. (1992). Some Bayesian numerical analysis. *Bayesian Statistics*, 4:345–363.

Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. J., and Ghahramani, Z. (2012a). Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems*, volume 25, pages 46–54.

Osborne, M. A., Garnett, R., Roberts, S. J., Hart, C., Aigrain, S., and Gibson, N. P. (2012b). Bayesian quadrature for ratios. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR Workshop and Conference Proceedings*, pages 832–840.

Patterson, T. N. L. (1968). The optimum addition of points to quadrature formulae. *Mathematics of Computation*, 22:847–856.

Paul, S., Chatzilygeroudis, K., Ciosek, K., Mouret, J.-B., Osborne, M. A., and Whiteson, S. (2018). Alternating optimisation and quadrature for robust control. In *The Thirty-Second AAAI Conference on Artificial Intelligence*.

Platte, R. B. and Driscoll, T. A. (2005). Polynomials and potential theory for Gaussian radial basis function interpolation. *SIAM Journal on Numerical Analysis*, 43(2):750–766.

Platte, R. B., Trefethen, L. N., and Kuijlaars, A. B. J. (2011). Impossibility of fast stable approximation of analytic functions from equispaced samples. *SIAM Review*, 53(2):308–318.

Pronzato, L. and Zhigljavsky, A. (2018). Bayesian quadrature and energy minimization for space-filling design. Preprint, arXiv:1808.10722v1.

Prüher, J., Karvonen, T., Oates, C. J., Straka, O., and Särkkä, S. (2018). Improved calibration of numerical integration error in sigma-point filters. Preprint, arXiv:1811.11474v1.

Prüher, J. and Šimandl, M. (2016). Bayesian quadrature variance in sigma-point filtering. In *International Conference on Informatics in Control, Automation and Robotics (revised selected papers)*, volume 12, pages 355–370. Springer International Publishing.

Prüher, J. and Straka, O. (2018). Gaussian process quadrature moment transform. *IEEE Transactions on Automatic Control*, 63(9):2844–2854.

Prüher, J. and Särkkä, S. (2016). On the use of gradient information in Gaussian process quadratures. In *26th IEEE International Workshop on Machine Learning for Signal Processing*.

Prüher, J., Tronarp, F., Karvonen, T., Särkkä, S., and Straka, O. (2017). Student-t process quadratures for filtering of non-linear systems with heavy-tailed noise. In *20th International Conference on Information Fusion*.

Punzi, A., Sommariva, A., and Vianello, M. (2008). Meshless cubature over the disk using thin-plate splines. *Journal of Computational and Applied Mathematics*, 221(2):430–436.

Rasmussen, C. E. and Ghahramani, Z. (2002). Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 15, pages 505–512.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press.

Rathinavel, J. and Hickernell, F. (2018). Fast automatic Bayesian cubature using lattice sampling. Preprint, arXiv:1809.09803v1.

Richter, N. (1970). Properties of minimal integration rules. *SIAM Journal on Numerical Analysis*, 7(1):67–79.

Richter-Dyn, N. (1971a). Minimal interpolation and approximation in Hilbert spaces. *SIAM Journal on Numerical Analysis*, 8(3):583–597.

Richter-Dyn, N. (1971b). Properties of minimal integration rules. II. *SIAM Journal on Numerical Analysis*, 8(3):497–508.

Rieger, C. and Zwicknagl, B. (2010). Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32:103–129.

Rieger, C. and Zwicknagl, B. (2014). Improved exponential convergence rates by oversampling near the boundary. *Constructive Approximation*, 39(2):323–341.

Ritter, K. (2000). *Average-Case Analysis of Numerical Problems*. Number 1733 in Lecture Notes in Mathematics. Springer.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer.

Sard, A. (1949). Best approximate integration formulas; best approximation formulas. *American Journal of Mathematics*, 71(1):80–91.

Schaback, R. (1993). Comparison of radial basis function interpolants. In *Multivariate Approximation: From CAGD to Wavelets*, pages 293–305. World Scientific.

Schaback, R. (1995). Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264.

Schaback, R. (2005). Multivariate interpolation by polynomials and radial basis functions. *Constructive Approximation*, 21(3):293–317.

Schaback, R. (2008). Limit problems for interpolation by analytical radial basis functions. *Journal of Computational and Applied Mathematics*, 212(2):127–149.

Schaback, R. and Wendland, H. (2006). Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15:543–639.

Scheuerer, M., Schaback, R., and Schlather, M. (2013). Interpolation of spatial data – a stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24(4):601–629.

Schober, M., Duvenaud, D. K., and Hennig, P. (2014). Probabilistic ODE solvers with Runge-Kutta means. In *Advances in Neural Information Processing Systems*, volume 27, pages 739–747.

Schober, M., Särkkä, S., and Hennig, P. (2018). A probabilistic model for the numerical solution of initial value problems. *Statistics and Computing*.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions: Part A.—on the problem of smoothing or graduation. a first class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):45–99.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences*, 52(4):947–950.

Skilling, J. (1992). Bayesian solution of ordinary differential equations. In *Maximum Entropy and Bayesian Methods*, pages 23–37. Springer.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, pages 1257–1264.

Solin, A. and Särkkä, S. (2019). Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*.

Sommariva, A. and Vianello, M. (2006a). Meshless cubature by Green's formula. *Applied Mathematics and Computation*, 183(2):1098–1107.

Sommariva, A. and Vianello, M. (2006b). Numerical cubature on scattered data by radial basis functions. *Computing*, 76(3–4):295–310.

Song, G., Riddle, J., Fasshauer, G. E., and Hickernell, F. J. (2012). Multivariate interpolation with increasingly flat radial basis functions of finite smoothness. *Advances in Computational Mathematics*, 36(3):485–501.

Sonoda, S. (2019). Unitary kernel quadrature for training parameter distributions. Preprint, arXiv:1902.00648v2.

Stein, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Number 30 in Princeton Mathematical Series. Princeton University Press.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer.

Stein, M. L. and Hung, Y. (2019). Comment on "Probabilistic integration: A role in statistical computation?". *Statistical Science*, 34(1):34–37.

Steinwart, I. (2017). Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. Preprint, arXiv:1403.1040v3.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Information Science and Statistics. Springer.

Steinwart, I., Hush, D., and Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643.

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417.

Suldin, A. V. (1959). Wiener measure and its applications to approximation methods. I. *Izvestiya vysshikh uchebnykh zavedeniy. Matematika*, (6):145–158. In Russian.

Suldin, A. V. (1960). Wiener measure and its applications to approximation methods. II. *Izvestiya vysshikh uchebnykh zavedeniy. Matematika*, (5):165–179. In Russian.

Sun, H. (2005). Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21(3):337–349.

Sun, H.-W. and Zhou, D.-X. (2008). Reproducing kernel Hilbert spaces associated with analytic translation-invariant Mercer kernels. *Journal of Fourier Analysis and Applications*, 14(1):89–101.

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electronic Journal of Statistics*, 7:991–1018.

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Number 3 in Institute of Mathematical Statistics Textbooks. Cambridge University Press.

Särkkä, S., Hartikainen, J., Svensson, L., and Sandblom, F. (2014). Gaussian process quadratures in nonlinear sigma-point filtering and smoothing. In *17th International Conference on Information Fusion*.

Särkkä, S., Hartikainen, J., Svensson, L., and Sandblom, F. (2016). On the relation between Gaussian process quadratures and sigma-point methods. *Journal of Advances in Information Fusion*, 11(1):31–46.

Särkkä, S., Solin, A., and Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61.

Teymur, O., Zygalakis, K., and Calderhead, B. (2016). Probabilistic linear multistep methods. In *Advances in Neural Information Processing Systems*, volume 29, pages 4321–4328.

Traub, J. F., Woźniakowski, H., and Wasilkowski, G. W. (1988). *Information-Based Complexity*. Academic Press.

Tronarp, F., Karvonen, T., and Särkkä, S. (2018). Mixture representation of the Matérn class with applications in state space approximations and Bayesian quadrature. In *28th IEEE International Workshop on Machine Learning for Signal Processing*.

Tronarp, F., Kersting, H., Särkkä, S., and Hennig, P. (2019). Probabilistic solutions to ordinary differential equations as non-linear Bayesian filtering: A new perspective. Preprint, arXiv:1810.03440v4.

van der Vaart, A. W. and van Zanten, J. H. (2008). *Reproducing Kernel Hilbert spaces of Gaussian Priors*, volume 3 of *IMS Collections*, pages 200–222. Institute of Mathematical Statistics.

van der Vaart, A. W. and van Zanten, J. H. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119.

Wagstaff, E., Hamid, S., and Osborne, M. (2018). Batch selection for parallelisation of Bayesian quadrature. Preprint, arXiv:1812.01553.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):364–372.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):133–150.

Wahba, G. (1990). *Spline Models for Observational Data*. Number 59 in CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wendland, H. (2005). *Scattered Data Approximation*. Number 17 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Wendland, H. and Rieger, C. (2005). Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748.

Wu, A., Aoi, M. C., and Pillow, J. W. (2018). Exploiting gradients and Hessians in Bayesian optimization and Bayesian quadrature. Preprint, arXiv:1704.00060v1.

Xi, X., Briol, F.-X., and Girolami, M. (2018). Bayesian quadrature for multiple related integrals. In *35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5373–5382.

Xu, W. and Stein, M. L. (2017). Maximum likelihood estimation for a smooth Gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):138–175.

Zwicknagl, B. and Schaback, R. (2013). Interpolation and approximation in Taylor spaces. *Journal of Approximation Theory*, 171:65–83.

Aalto-DD 160/2019

ISBN 978-952-60-8703-0 (printed)
ISBN 978-952-60-8704-7 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Electrical Engineering and Automation
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS