FAST AUTOMATIC BAYESIAN CUBATURE USING MATCHING KERNELS AND DESIGNS

BY

JAGADEESWARAN RATHINAVEL

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Applied Mathematics
in the Graduate College of the
Illinois Institute of Technology

Approved _____
Advisor

Chicago, Illinois
December 2019

ACKNOWLEDGMENT

I want to thank my advisor Prof. Fred J Hickernell for his support and guidance in my completion of this thesis and throughout my studies here at IIT. His support and motivation have given me the confidence to endure through the research.

I would like to also thank the GAIL project collaborators with whom I have worked to add my new algorithms to the GAIL MATLAB toolbox: Prof. Sou-Cheng Choi, Yuhan Ding, Lan Jiang, Xin Tong, and Kan Zhang. Especially, Prof. Sou-Cheng Choi's support and guidance as the project leader helped me to focus on my cubature algorithms.

My special gratitude also goes to my thesis committee members, Prof. Jinqiao Duan, Prof. Fred J Hickernell, Prof. Shuwang Li, and Prof. Geoffrey Williamson. Above all, I want to thank them because they were flexible and willing to dedicate time to review my work and attend my comprehensive and defense examinations.

I would like to thank Prof. Dirk Nuyens for suggestions, valuable tips and notes when we were researching higher order nets and kernels.

I would like to thank the organizers of the SAMSI-Lloyds-Turing Workshop on Probabilistic Numerical Methods, where a part of preliminary version of this work was discussed. I also thank Prof. Chris Oates and Prof. Sou-Cheng Choi for valuable comments.

I would like to specifically thank my friend Samuel Davidson for reviewing and suggesting the improvements on the text.

Last but not least, I would not be able to make it without the support of my family. I would like to thank my wife for her continuous support and sacrifice. I also would like to thank my parents for their endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Automatic cubatures approximate multidimensional integrals to user-specified error tolerances. In many real-world integration problems, the analytical solution is either unavailable or difficult to compute. To overcome this, one can use numerical algorithms that approximately estimate the value of the integral.

For high dimensional integrals, quasi-Monte Carlo (QMC) methods are very popular. QMC methods are equal-weight quadrature rules where the quadrature points are chosen deterministically, unlike Monte Carlo (MC) methods where the points are chosen randomly. The families of integration lattice nodes and digital nets are the most popular quadrature points used. These methods consider the integrand to be a deterministic function. An alternate approach, called Bayesian cubature, postulates the integrand to be an instance of a Gaussian stochastic process.

For high dimensional problems, it is difficult to adaptively change the sampling pattern. But one can automatically determine the sample size, n, given a fixed and reasonable sampling pattern. We take this approach using a Bayesian perspective. We assume a Gaussian process parameterized by a constant mean and a covariance function defined by a scale parameter and a function specifying how the integrand values at two different points in the domain are related. These parameters are estimated from integrand values or are given non-informative priors. This leads to a credible interval for the integral. The sample size, n, is chosen to make the credible interval for the Bayesian posterior error no greater than the desired error tolerance.

However, the process just outlined typically requires vector-matrix operations with a computational cost of $O(n^3)$. Our innovation is to pair low discrepancy nodes with matching kernels, which lowers the computational cost to $O(n \log n)$. We begin the thesis by introducing the Bayesian approach to calculate the posterior cubature error and define our automatic Bayesian cubature (Chapter 2). Although much of this

material is known, it is used to develop the necessary foundations. The contributions of this thesis are as follows:

- The fast Bayesian transform is introduced. This generalizes the techniques that speedup Bayesian cubature when the kernel matches low discrepancy nodes (Chapter 3).
- The fast Bayesian transform approach is demonstrated using two methods: 1) rank-1 lattice sequences and shift-invariant kernels (Chapter 4), and 2) Sobol' sequences and Walsh kernels (Chapter 5). These two methods are implemented as fast automatic Bayesian cubature algorithms in the Guaranteed Automatic Integration Library (GAIL).
- We develop additional numerical implementation techniques (Chapter 6): 1) rewriting the covariance kernel to avoid cancellation error, 2) gradient descent for hyperparameter search, and 3) non-integer kernel order selection.

The thesis concludes by applying our fast automatic Bayesian cubature algorithms to three sample integration problems (Chapter 7). We show that our algorithms are faster than the basic Bayesian cubature and that they provide answers within the error tolerance in most cases. A significant portion of this thesis comprising an automatic Bayesian cubature algorithm using lattice sequences and shift-invariant kernels was published and discussed in [1, 2].

The Bayesian cubatures that we develop are guaranteed for integrands belonging to cone of functions which reside in the middle of the sample space. The concept of a cone of functions is explained in Section 2.6.

CHAPTER 1

INTRODUCTION

1.1 Cubature

Cubature is the problem of inferring a numerical value for a definite integral, $\mu := \int_{\mathbb{R}^d} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, where μ has no closed form analytic expression. Typically, g is accessible through a black-box function routine. Cubature means numerical multivariate integration and is a key component of many problems in scientific computing, finance [3], statistical modeling, imaging [4], uncertainty quantification, machine learning, etc. [5].

The integral may often be expressed as

$$\mu := \mu(f) := \mathbb{E}[f(\boldsymbol{X})] = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{1.1}$$

where $f : [0,1]^d \to \mathbb{R}$ is the integrand, and $\boldsymbol{X} \sim \mathcal{U}[0,1]^d$. The process of transforming the original integral into the form of (1.1) is not addressed here. The cubature may be an affine function of integrand values:

$$\widehat{\mu} := \widehat{\mu}(f) := w_0 + \sum_{i=1}^{n} f(\boldsymbol{x}_i) w_i, \tag{1.2}$$

where the weights, w_0, and $\boldsymbol{w} = (w_i)_{i=1}^n \in \mathbb{R}^n$, and the nodes, $\{\boldsymbol{x}_i\}_{i=1}^n \subset [0,1]^d$, are chosen to make the error, $|\mu - \widehat{\mu}|$, small. The integration domain $[0,1]^d$ is convenient for the low discrepancy node sets that we use. The nodes are assumed to be deterministic. The integral of function f is the same over $[0,1]^d$ or $(0,1)^d$ or $[0,1)^d$. So we use $[0,1]^d$ or $[0,1)^d$ depending on the application. Most often $[0,1)^d$ is preferred especially for extensible node-sets because it partitions easily into congruent subhypercubes. This research focuses on multivariate numerical integrals where the computational cost is a bottleneck.

1.2 Stopping Criterion

We construct a reliable stopping criterion that determines the number of integrand values required, n, to ensure that the error is no greater than a user-defined error tolerance denoted by ε, i.e.,

$$|\mu - \widehat{\mu}| \leq \varepsilon. \tag{1.3}$$

Rather than relying on strong assumptions about the integrand, such as an upper bound on its variance or total variation, we construct a stopping criterion that is based on a credible interval arising from a Bayesian approach to the problem. We build upon the work of Briol et al. [6], Diaconis [7], O'Hagan [8], Ritter [9], Rasmussen and Ghahramani [10], and others. Our algorithm is an example of *probabilistic numerics*. To study numerical algorithms from a statistical point of view, where uncertainty is formally due to the presence of an unknown numerical error, is the goal of probabilistic numerics.

Our primary contribution in this research is to demonstrate how the choice of a family of covariance kernels that match the low discrepancy sampling nodes facilitates fast computation of the cubature and the data-driven stopping criterion. Our Bayesian cubature requires a computational cost of

$$\mathcal{O}\big(n\$(f) + N_{\text{opt}}[n\$(C) + n\log(n)]\big), \tag{1.4}$$

where $\$(f)$ is the cost of one integrand value, $\$(C)$ is the cost of a single covariance kernel value, $\mathcal{O}(n\log(n))$ is the cost of a fast Bayesian transform, and N_{opt} is an upper bound on the number of optimization steps required to choose the hyperparameters. If function evaluation is expensive, e.g., the output of a computationally intensive simulation, or if $\$(f) = \mathcal{O}(d)$ for large d, then $\$(f)$ might be similar in magnitude to $N_{\text{opt}}\log(n)$ in practice. Typically, $\$(C) = \mathcal{O}(d)$. Note that the $\mathcal{O}(n\log(n))$ contribution is d independent.

In contrast to our fast algorithm, the typical computational cost for Bayesian

cubature is

$$\mathcal{O}\big(n\$(f) + N_{\text{opt}}[n^2\$(C) + n^3]\big), \tag{1.5}$$

which is explained in Section 2.7. Note that apart from evaluating the integrand, the computational cost in (1.5) is much larger than that in (1.4).

1.3 Low Discrepancy Points

Low discrepancy points are characterized by how uniformly the points are distributed, which is measured by the *discrepancy*. The goal is to have maximum uniform space filling. The discrepancy is defined as below. Let $\mathcal{M}$ be the set of all intervals of the form $\prod_{\ell=1}^{d}[a_\ell, b_\ell) = \{\boldsymbol{x} \in \mathbb{R}^d : a_\ell \leq x_\ell \leq b_\ell, 0 \leq a_\ell \leq b_\ell \leq 1\}$, where $|\mathcal{P}|$ is the cardinality of the set $\mathcal{P}$, and λ_L is the Lebesgue measure. Then, the discrepancy of a point set $\mathcal{P}$ is,

$$D(\mathcal{P}) := \sup_{M \in \mathcal{M}} \left| \frac{|M \cap \mathcal{P}|}{|\mathcal{P}|} - \lambda_L(M) \right|.$$

The *low discrepancy points* satisfy $D(\mathcal{P}) = \mathcal{O}((\log n)^d/n)$. In this work we experiment with two most popular low discrepancy point sets, 1) lattice points, and 2) Sobol' points.

1.4 Prior Work

Hickernell [11] compares different approaches to cubature error analysis depending on whether the rule is deterministic or random and whether the integrand is assumed to be deterministic or random. Error analysis that assumes a deterministic integrand lying in a Banach space leads to an error bound that is typically impractical for deciding how large n must be to satisfy (1.3). The deterministic error bound includes a (semi-)norm of the integrand, which is often more complex to compute than the original integral.

Hickernell and Jiménez-Rugama [12, 13] have developed stopping criteria for

cubature rules based on low discrepancy nodes by tracking the decay of the discrete Fourier coefficients of the integrand. The algorithms proposed here also rely on discrete Fourier coefficients, but in a different way. We only discuss automatic Bayesian cubature for absolute error tolerances in this thesis. The recent work by Hickernell, Jiménez-Rugama, and Li [14] suggests how one might accommodate more general error criteria, such as relative error tolerances which has been adapted in the MATLAB implementation of our algorithms.

Chapter 2 explains the Bayesian approach to calculate the posterior cubature error and defines our automatic Bayesian cubature. Although much of this material is known, it is included for completeness. We end Chapter 2 by demonstrating why Bayesian cubature is typically computationally expensive. Chapter 3 introduces the concept of covariance kernels that match the nodes and expedite the computations required by our automatic Bayesian cubature. Chapter 4 implements this concept for shift invariant kernels and rank-1 lattice nodes. It also develops approaches to build shift-invariant kernels of continuous valued kernel order rather than fixing the kernel order to integer values. Chapter 5 demonstrates another implementation of matching nodes and kernel using Sobol' points and Walsh kernels. It also shows that the fast Walsh Hadamard as the fast Bayesian transform for this case. Chapter 6 describes how to avoid cancellation error for kernels of product form. It also covers some of the additional techniques used in the implementation of our Bayesian Cubature algorithms. Numerical examples are provided in Chapter 7 to demonstrate the performance and advantages of our new algorithms. We conclude with a brief discussion and potential future work.

We use the terms integrand or function interchangeably to denote the function f being considered for the numerical integration. Also, we use the terms, nodes, points, node-sets, designs, and data-sites interchangeably to denote the points $\mathcal{P} :=$

$\{\boldsymbol{x}_i\}_{i=1}^n \subset [0,1)^d$ used in the cubature.

CHAPTER 2

BAYESIAN CUBATURE

The Bayesian approach for numerical analysis was popularized by Diaconis [7]. The earliest reference for such kind of approach dates back to Poincaré, where, the theory of interpolation was discussed. Diaconis motivates the reader by interpreting the most well known numerical methods, 1) trapezoidal rule and 2) splines, from the statistical point of view with whatever is known about the integrand as prior information. For example, the trapezoidal rule can be interpreted as a Bayesian method with prior information being modeled as a Brownian motion in the sample space $\mathcal{C}[0,1)$, the space of continuous functions.

This research is focused on the Bayesian approach for numerical integration that is known as Bayesian cubature as introduced by O'Hagan [15]. Bayesian cubature returns a probability distribution, that expresses belief about the true value of integral, $\mu(f)$. This posterior probability distribution is based on a prior that depends on f, which is computed via Bayes' rule using the *data* contained in the function evaluations [6]. The distribution in general captures numerical uncertainty due to the fact that we have only used a finite number of function values to evaluate the integral.

2.1 Bayesian Posterior Error

We assume the integrand, f, is an instance of a stochastic Gaussian process, i.e., $f \sim \mathcal{GP}(m, s^2 C_{\boldsymbol{\theta}})$. Specifically, f is a real-valued random function with constant mean m and covariance function $s^2 C_{\boldsymbol{\theta}}$, where s is a positive scale factor, and $C_{\boldsymbol{\theta}} : [0,1]^d \times [0,1]^d \to \mathbb{R}$ is a symmetric, positive-definite function and parameterized by $\boldsymbol{\theta}$:

$$\mathsf{C}^T = \mathsf{C}, \quad \boldsymbol{a}^T \mathsf{C} \boldsymbol{a} > 0, \quad \text{where } \mathsf{C} = \left(C_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)_{i,j=1}^{n},$$

$$\text{for all } \boldsymbol{a} \neq 0,\ n \in \mathbb{N},\ \text{distinct } \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in [0,1]^d. \quad (2.1)$$

The covariance function, C, and the Gram matrix, C, depend implicitly on $\boldsymbol{\theta}$, but the notation may omit this for simplicity's sake. Procedures for estimating or integrating out the hyperparameters m, s, and $\boldsymbol{\theta}$ are explained later in this section.

For a Gaussian process, all vectors of linear functionals of f have a multivariate Gaussian distribution. For any deterministic sampling scheme with distinct nodes, $\{\boldsymbol{x}_i\}_{i=1}^n$, and defining $\boldsymbol{f} := (f(\boldsymbol{x}_i))_{i=1}^n$ as the multivariate Gaussian vector of function values, it follows from the definition of a Gaussian process that

$$\boldsymbol{f} \sim \mathcal{N}(m\mathbf{1}, s^2\mathsf{C}), \tag{2.2a}$$

$$\mu \sim \mathcal{N}(m, s^2 c_0), \tag{2.2b}$$

$$\text{where } c_0 := \int_{[0,1]^d \times [0,1]^d} C_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{t}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{t}, \tag{2.2c}$$

$$\operatorname{cov}(\boldsymbol{f}, \mu) = \left(\int_{[0,1]^d} C(\boldsymbol{t}, \boldsymbol{x}_i) \, \mathrm{d}\boldsymbol{t} \right)_{i=1}^n =: \boldsymbol{c}. \tag{2.2d}$$

Here, c_0 and $\boldsymbol{c}$ depend implicitly on $\boldsymbol{\theta}$. We assume the covariance function C is simple enough that the integrals in these definitions can be computed analytically. We need the following lemma to derive the posterior error of our cubature.

Lemma 2.1.1. *[16, (A.6), (A.11–13)] If* $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)^T \sim \mathcal{N}(\boldsymbol{m}, \mathsf{C})$*, where* $\boldsymbol{Y}_1$ *and* $\boldsymbol{Y}_2$ *are random vectors of arbitrary length, and*

$$\boldsymbol{m} = \begin{pmatrix} \boldsymbol{m}_1 \\ \\ \boldsymbol{m}_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\boldsymbol{Y}_1) \\ \\ \mathbb{E}(\boldsymbol{Y}_2) \end{pmatrix},$$

$$\mathsf{C} = \begin{pmatrix} \mathsf{C}_{11} & \mathsf{C}_{21}^T \\ \\ \mathsf{C}_{21} & \mathsf{C}_{22} \end{pmatrix} = \begin{pmatrix} \operatorname{var}(\boldsymbol{Y}_1) & \operatorname{cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) \\ \\ \operatorname{cov}(\boldsymbol{Y}_2, \boldsymbol{Y}_1) & \operatorname{var}(\boldsymbol{Y}_2) \end{pmatrix}$$

then

$$\boldsymbol{Y}_1 | \boldsymbol{Y}_2 \ \sim \ \mathcal{N}\big(\boldsymbol{m}_1 + \mathsf{C}_{21}^T \mathsf{C}_{22}^{-1} (\boldsymbol{Y}_2 - \boldsymbol{m}_2), \quad \mathsf{C}_{11} - \mathsf{C}_{21}^T \mathsf{C}_{22}^{-1} \mathsf{C}_{21}\big).$$

Moreover, the inverse of the matrix C *may be partitioned as*

$$\mathsf{C}^{-1} = \begin{pmatrix} \mathsf{A}_{11} & \mathsf{A}_{21}^T \\ \mathsf{A}_{21} & \mathsf{A}_{22} \end{pmatrix},$$

$$\mathsf{A}_{11} = (\mathsf{C}_{11} - \mathsf{C}_{12}\mathsf{C}_{22}^{-1}\mathsf{C}_{21})^{-1}, \qquad \mathsf{A}_{21} = -\mathsf{C}_{22}^{-1}\mathsf{C}_{21}\mathsf{A}_{11},$$

$$\mathsf{A}_{22} = \mathsf{C}_{22}^{-1} + \mathsf{C}_{22}^{-1}\mathsf{C}_{21}\mathsf{A}_{11}\mathsf{C}_{21}^T\mathsf{C}_{22}^{-1}.$$

It follows from Lemma 2.1.1 that the *conditional* distribution of the integral given observed function values, $\boldsymbol{f} = \boldsymbol{y}$ is also Gaussian:

$$\mu|(\boldsymbol{f} = \boldsymbol{y}) \sim \mathcal{N}\big(m(1 - \boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1}) + \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}, \;\; s^2(c_0 - \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})\big). \tag{2.3}$$

The natural choice for the cubature is the posterior mean of the integral, namely,

$$\widehat{\mu}|(\boldsymbol{f} = \boldsymbol{y}) = m(1 - \mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{c}) + \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}, \tag{2.4}$$

which takes the form of (1.2). Under this definition, the cubature error has zero mean and a variance depending on the choice of nodes:

$$(\mu - \widehat{\mu})|(\boldsymbol{f} = \boldsymbol{y}) \sim \mathcal{N}\left(0, \quad s^2(c_0 - \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})\right).$$

A credible interval for the integral is given by

$$\mathbb{P}_f\left[|\mu - \widehat{\mu}| \le \mathrm{err}_{\mathrm{CI}}\right] = 99\%, \tag{2.5a}$$

$$\mathrm{err}_{\mathrm{CI}} = 2.58s\sqrt{c_0 - \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}}. \tag{2.5b}$$

Naturally, 2.58 and 99% can be replaced by other quantiles and credible levels.

2.2 Hyperparameter Estimation

The credible interval in (2.5) suggests how our automatic Bayesian cubature proceeds. Integrand data is accumulated until the width of the credible interval, $\mathrm{err}_{\mathrm{CI}}$, is no greater than the error tolerance. As n increases, one expects $c_0 - \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}$ to

decrease for well-chosen nodes, $\{\boldsymbol{x}_i\}_{i=1}^n$. Please note that the credible interval depends on the parameters m, s, and $\boldsymbol{\theta}$

Note that err_{CI} has no explicit dependence on the integrand values, even though one would intuitively expect that a larger integrand should imply a larger err_{CI}. This is because the hyperparameters, m, s, and $\boldsymbol{\theta}$, have not yet been inferred from integrand data. After inferring the hyperparameters, err_{CI} *does reflect the size* of the integrand values. This section describes three approaches to hyperparameter estimation.

2.3 Empirical Bayes

The first and a very straight forward approach is to estimate the parameters via maximum likelihood estimation. The log-likelihood function of the parameters given the function data $\boldsymbol{y}$ is:

$$\begin{aligned} l(s, m, \boldsymbol{\theta}|\boldsymbol{y}) = &-\frac{1}{2}s^{-2}(\boldsymbol{y} - m\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y} - m\mathbf{1}) \\ &- \frac{1}{2}\log(\det\,\mathsf{C}) - \frac{n}{2}\log(s^2) + \text{constants}. \end{aligned}$$

Maximizing the log-likelihood first with respect to m, then with respect to s, and finally with respect to $\boldsymbol{\theta}$ yields

$$\begin{aligned} m_{\text{EB}} &= \frac{\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}, \\ s^2_{\text{EB}} &= \frac{1}{n}(\boldsymbol{y} - m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y} - m_{\text{EB}}\mathbf{1}) \\ &= \frac{1}{n}\boldsymbol{y}^T\left[\mathsf{C}^{-1} - \frac{\mathsf{C}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}^{-1}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}\right]\boldsymbol{y}, \\ \boldsymbol{\theta}_{\text{EB}} &= \underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\log\left(\boldsymbol{y}^T\left[\mathsf{C}^{-1} - \frac{\mathsf{C}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}^{-1}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}\right]\boldsymbol{y}\right) + \frac{1}{n}\log(\det(\mathsf{C}))\right\}. \end{aligned}$$

The empirical Bayes estimate of $\boldsymbol{\theta}$ balances minimizing the covariance scale factor, s^2_{EB}, against minimizing $\det(\mathsf{C})$.

Under these estimates of the parameters, the cubature (2.4) and the credible

interval (2.5) simplify to

$$\widehat{\mu}_{\text{EB}} = \left(\frac{(1 - \mathbf{1}^T \mathsf{C}^{-1}\boldsymbol{c})\mathbf{1}}{\mathbf{1}^T \mathsf{C}^{-1}\mathbf{1}} + \boldsymbol{c}\right)^T \mathsf{C}^{-1}\boldsymbol{y},$$
$$\text{err}^2_{\text{EB}} := \frac{2.58^2}{n}\boldsymbol{y}^T \left[\mathsf{C}^{-1} - \frac{\mathsf{C}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}^{-1}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}\right]\boldsymbol{y}(c_0 - \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}),$$

$$\mathbb{P}_f\left[|\mu - \widehat{\mu}_{\text{EB}}| \leq \text{err}_{\text{EB}}\right] = 99\%. \tag{2.6}$$

Here c_0, $\boldsymbol{c}$, and C are assumed implicitly to be based on $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{EB}}$.

2.3.1 Gradient descent to find optimal shape parameter. The equation specifying $\boldsymbol{\theta}_{\text{EB}}$ as defined in (2.16) does not say how the parameter search can be done. There exist empirical algorithms [17, 18] that one could use to accomplish the same. Since the objective function is known we could compute the gradient. Using the gradient of $l(s, m, \boldsymbol{\theta}|\boldsymbol{y})$, one can apply optimization techniques such as gradient descent to find the optimal value faster. Let us define the objective function for the same purpose by excluding the negative sign, which modifies the problem to become a minimization of

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) := \frac{1}{n}\log(\det\,\mathsf{C}) + \log\left((\boldsymbol{y} - m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y} - m_{\text{EB}}\mathbf{1})\right) + \text{constants}.$$

Taking derivative with respect to θ_ℓ, for $\ell = 1, \cdots, d$

$$\begin{aligned}\frac{\partial}{\partial\theta_\ell}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{1}{n}\frac{\partial}{\partial\theta_\ell}\log(\det\,\mathsf{C}) + \frac{\partial}{\partial\theta_\ell}\log\left((\boldsymbol{y} - m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y} - m_{\text{EB}}\mathbf{1})\right)\\ &= \frac{1}{n}\text{trace}\left(\mathsf{C}^{-1}\frac{\partial\mathsf{C}}{\partial\theta_\ell}\right) - \frac{\left((\boldsymbol{y} - m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}\right)^T\left(\frac{\partial\mathsf{C}}{\partial\theta_\ell}\right)\left((\boldsymbol{y} - m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}\right)}{(\boldsymbol{y} - m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y} - m_{\text{EB}}\mathbf{1})}\end{aligned}$$

where we used some of the results from [19]. This can be used with gradient descent as follows,

$$\theta_\ell^{(j+1)} = \theta_\ell^{(j)} - \nu\frac{\partial}{\partial\theta_\ell}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}), \quad j = 0, 1, \cdots \tag{2.7}$$

where ν is the step size for the gradient descent.

2.4 Full Bayes

Rather than use maximum likelihood to determine m and s, one can treat them as hyper-parameters with a non-informative, conjugate prior, namely $\boldsymbol{\rho}_{m,s^2}(\xi,\lambda) \propto 1/\lambda$. Then the posterior density for the integral given the data using Bayes theorem is,

$$
\begin{aligned}
&\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y}) \\
&\quad\propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\boldsymbol{f}=\boldsymbol{y}, m=\xi, s^2=\lambda)\rho_{\boldsymbol{f}}(\boldsymbol{y}|\xi,\lambda)\rho_{m,s^2}(\xi,\lambda)\,\mathrm{d}\xi\mathrm{d}\lambda \\
&\qquad\qquad \text{by the properties of conditional probability} \\
&\quad\propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\boldsymbol{f}=\boldsymbol{y}, m=\xi, s^2=\lambda)\rho_{\boldsymbol{f}}(\boldsymbol{y}|\xi,\lambda)\rho_{m,s^2}(\xi,\lambda)\,\mathrm{d}\xi\mathrm{d}\lambda \\
&\qquad\qquad \text{by Bayes' Theorem} \\
&\quad\propto \int_0^\infty \frac{1}{\lambda^{(n+3)/2}} \int_{-\infty}^\infty \exp\Bigg(-\frac{1}{2\lambda}\Bigg\{\frac{[z-\xi(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}]^2}{c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}} \\
&\qquad\qquad + (\boldsymbol{y}-\xi\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y}-\xi\mathbf{1})\Bigg\}\Bigg)\,\mathrm{d}\xi\mathrm{d}\lambda \\
&\qquad\qquad \text{by (2.2), (2.3) and } \rho_{m,s^2}(\xi,\lambda)\propto 1/\lambda \\
&\quad\propto \int_0^\infty \frac{1}{\lambda^{(n+3)/2}} \int_{-\infty}^\infty \exp\left(-\frac{\alpha\xi^2-2\beta\xi+\gamma}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\,\mathrm{d}\xi\mathrm{d}\lambda,
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha &= (1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})^2 + \mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}), \\
\beta &= (1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})(z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}) + \mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}), \\
\gamma &= (z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y})^2 + \boldsymbol{y}^T\mathsf{C}^{-1}\boldsymbol{y}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}).
\end{aligned}
$$

In the derivation above and below, factors that are independent of ξ, λ, or z can be discarded since we only need to preserve the proportion. But, factors that depend on ξ, λ, or z must be kept. Completing the square $\alpha\xi^2-2\beta\xi+\gamma = \alpha(\xi-\beta/\alpha)^2 - (\beta^2/\alpha)+\gamma$, allows us to evaluate the integrals with respect to ξ and λ:

$$\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y})$$

$$
\begin{aligned}
&\propto \int_0^\infty \frac{1}{\lambda^{(n+3)/2}} \exp\left(-\frac{\gamma-\beta^2/\alpha}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\cdots \\
&\qquad \cdots\int_{-\infty}^\infty \exp\left(-\frac{\alpha(\xi-\beta/\alpha)^2}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\,\mathrm{d}\xi\mathrm{d}\lambda \\
&\propto \int_0^\infty \frac{1}{\lambda^{(n+2)/2}} \exp\left(-\frac{\gamma-\beta^2/\alpha}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\mathrm{d}\lambda \\
&\propto \left(\gamma-\frac{\beta^2}{\alpha}\right)^{-n/2} \propto \left(\alpha\gamma-\beta^2\right)^{-n/2}.
\end{aligned}
$$

Finally, we simplify the key term:

$$
\begin{aligned}
\alpha\gamma-\beta^2 &= \mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})(z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y})^2 \\
&\quad -2\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})(z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}) \\
&\quad +(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})^2\boldsymbol{y}^T\mathsf{C}^{-1}\boldsymbol{y}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}) \\
&\quad +[\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}\boldsymbol{y}^T\mathsf{C}^{-1}\boldsymbol{y}-(\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y})^2](c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})^2 \\
&\propto \mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}\left(z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}-\frac{(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}\right)^2 \\
&\quad -\frac{[(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y}]^2}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}+(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})^2\boldsymbol{y}^T\mathsf{C}^{-1}\boldsymbol{y} \\
&\quad (c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})[\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}\boldsymbol{y}^T\mathsf{C}^{-1}\boldsymbol{y}-(\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y})^2] \\
&\propto \left(z-\left[\frac{(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})\mathbf{1}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}+\boldsymbol{c}\right]^T\mathsf{C}^{-1}\boldsymbol{y}\right)^2 \\
&\quad +\left[\frac{(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})^2}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}+(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})\right]\times\boldsymbol{y}^T\left[\mathsf{C}^{-1}-\frac{\mathsf{C}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}^{-1}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}\right]\boldsymbol{y} \\
&\propto (z-\widehat{\mu}_{\text{full}})^2+(n-1)\sigma^2_{\text{full}} \\
&\propto \left(1+\frac{1}{n-1}\frac{(z-\mu_{\text{full}})^2}{\widehat{\sigma}^2_{\text{full}}}\right),
\end{aligned}
$$

i.e.,

$$
\alpha\gamma-\beta^2 \propto \left(1+\frac{(z-\widehat{\mu}_{\text{full}})^2}{(n-1)\widehat{\sigma}^2_{\text{full}}}\right), \tag{2.8}
$$

where $\widehat{\mu}_{\text{full}}=\widehat{\mu}_{\text{EB}}$ and

$$
\widehat{\sigma}^2_{\text{full}} \;:=\; \frac{1}{n-1}\boldsymbol{y}^T\left[\mathsf{C}^{-1}-\frac{\mathsf{C}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}^{-1}}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}\right]\boldsymbol{y}\;\times\;\left[\frac{(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})^2}{\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}}+(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})\right].
$$

The confidence interval is:

$$\mathbb{P}_f\left[|\mu - \widehat{\mu}_{\text{EB}}| \leq \text{err}_{\text{full}}\right] = 99\%, \tag{2.9}$$

where

$$\text{err}_{\text{full}} := t_{n_j-1,0.995}\widehat{\sigma}_{\text{full}} > \text{err}_{\text{EB}}.$$

Here $t_{n-1,0.995}$ denotes the 99.5 percentile of a standard Student's t-distribution with $n-1$ degrees of freedom. This means that $\mu|(\boldsymbol{f} = \boldsymbol{y})$, properly centered and scaled, has a Student's t-distribution with $n-1$ degrees of freedom. The estimated integral is the same as in the empirical Bayes case, $\widehat{\mu}_{\text{full}} = \widehat{\mu}_{\text{EB}}$, but the credible interval is wider. In other words, the stopping criterion for the full Bayes case is more conservative than that in the empirical Bayes case, (2.6).

Because the shape parameter, $\boldsymbol{\theta}$, enters the definition of the covariance kernel in a non-trivial way, the only way to treat it as a hyperparameter and assign a tractable prior would be for the prior to be discrete. We believe in practice that choosing such a prior involves more guesswork than using the empirical Bayes estimate of $\boldsymbol{\theta}$ in (2.16) or the cross-validation approach described next.

2.4.1 Full Bayes with general prior. Rather than use non-informative, conjugate prior one can use general prior, namely $\boldsymbol{\rho}_{m,s^2}(\xi,\lambda) \propto g(\lambda)$, which can generalize to any general function. One would be curious if the posterior function can be obtained from the data, i.e, the integrand values. The posterior density for the integral given the data using Bayes theorem is,

$$\begin{aligned}
&\rho_\mu(z|\boldsymbol{f} = \boldsymbol{y}) \\
&\quad \propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\boldsymbol{f} = \boldsymbol{y}, m = \xi, s^2 = \lambda)\rho_{\boldsymbol{f}}(\boldsymbol{y}|\xi,\lambda)\rho_{m,s^2}(\xi,\lambda)\,\mathrm{d}\xi\mathrm{d}\lambda \\
&\qquad\qquad \text{by the properties of conditional probability} \\
&\quad \propto \int_0^\infty \int_{-\infty}^\infty \rho_\mu(z|\boldsymbol{f} = \boldsymbol{y}, m = \xi, s^2 = \lambda)\rho_{\boldsymbol{f}}(\boldsymbol{y}|\xi,\lambda)\rho_{m,s^2}(\xi,\lambda)\,\mathrm{d}\xi\mathrm{d}\lambda
\end{aligned}$$

$$\text{by Bayes' Theorem}$$

$$\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{(n+1)/2}} \int_{-\infty}^\infty \exp\Bigg(-\frac{1}{2\lambda}\Bigg\{\frac{[z-\xi(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y}]^2}{c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}} + (\boldsymbol{y}-\xi\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y}-\xi\mathbf{1})\Bigg\}\Bigg)\,\mathrm{d}\xi\mathrm{d}\lambda$$

$$\text{by (2.2), (2.3) and } \rho_{m,s^2}(\xi,\lambda)\propto g(\lambda)$$

$$\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{(n+1)/2}} \int_{-\infty}^\infty \exp\left(-\frac{\alpha\xi^2-2\beta\xi+\gamma}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\mathrm{d}\xi\mathrm{d}\lambda,$$

where

$$\alpha=(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})^2+\mathbf{1}^T\mathsf{C}^{-1}\mathbf{1}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}),$$
$$\beta=(1-\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1})(z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y})+\mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{y}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}),$$
$$\gamma=(z-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y})^2+\boldsymbol{y}^T\mathsf{C}^{-1}\boldsymbol{y}(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}).$$

In the derivation above and below, factors that are independent of ξ, λ, or z can be discarded since we only need to preserve the proportion. But, factors that depend on ξ, λ, or z must be kept. Completing the square $\alpha\xi^2-2\beta\xi+\gamma=\alpha(\xi-\beta/\alpha)^2-(\beta^2/\alpha)+\gamma$, allows us to evaluate the integrals with respect to ξ and λ:

$$\begin{aligned}
&\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y})\\
&\quad\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{(n+1)/2}}\exp\left(-\frac{\gamma-\beta^2/\alpha}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\cdots\\
&\qquad\qquad\cdots\int_{-\infty}^\infty \exp\left(-\frac{\alpha(\xi-\beta/\alpha)^2}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\mathrm{d}\xi\mathrm{d}\lambda\\
&\quad\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}}\exp\left(-\frac{\gamma-\beta^2/\alpha}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\mathrm{d}\lambda.
\end{aligned}$$

This can be interpreted as Laplace transform of $g(\lambda)$,

$$\begin{aligned}
\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y})&\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}}\exp\left(-\frac{\gamma-\beta^2/\alpha}{2\lambda(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\right)\mathrm{d}\lambda\\
&\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}}\exp\left(-\frac{1}{\lambda}\chi\right)\mathrm{d}\lambda,\\
&\qquad\text{where}\quad \chi=\frac{\gamma-\beta^2/\alpha}{2(c_0-\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c})}\propto 1+\frac{(z-\widehat{\mu}_{\text{full}})^2}{(n-1)\widehat{\sigma}^2_{\text{full}}}.
\end{aligned}$$

Let $\lambda = \dfrac{1}{w}, \quad \mathrm{d}\lambda = -w^{-2}\mathrm{d}w$ then,

$$\begin{aligned}
\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y}) &\propto \int_0^\infty \frac{g(\lambda)}{\lambda^{n/2}} \exp\left(-\frac{1}{\lambda}\chi\right) \mathrm{d}\lambda \\
&= \int_0^\infty \frac{g(1/w)}{w^{-n/2}} \exp\left(-w\chi\right) (-w^{-2})\mathrm{d}w \\
&= \int_\infty^0 -g(1/w) w^{\frac{n}{2}-2} \exp\left(-w\chi\right) \mathrm{d}w \\
&= \int_0^\infty g(1/w) w^{\frac{n-4}{2}} \exp\left(-w\chi\right) \mathrm{d}w \\
&= \mathcal{LT}\{g(1/\cdot)\}^{(\frac{n-4}{2})}(\chi),
\end{aligned}$$

where $\mathcal{LT}(\cdot)$ denotes the Laplace transform and $(\frac{n-4}{2})$ indicates the $\frac{n-4}{2}$th derivative taken after the transform. Here we used frequency domain derivative property of the Laplace transform. The above result can be further simplified by replacing $\gamma - \beta^2/\alpha$ from (2.8),

$$\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y}) \propto \mathcal{LT}\{g(1/\cdot)\}^{(\frac{n-4}{2})}(\chi) \propto \mathcal{LT}\left\{g(1/\cdot)\right\}^{(\frac{n-4}{2})}\left(1 + \frac{(z-\widehat{\mu}_{\text{full}})^2}{(n-1)\widehat{\sigma}^2_{\text{full}}}\right) \quad \text{by} \quad (2.8).$$

Thus, $\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y})$ is proportional to $(\frac{n-4}{2})$th derivative of the Laplace transform of $g(1/\cdot)$ evaluated at χ, where $\chi \propto 1 + \frac{(z-\widehat{\mu}_{\text{full}})^2}{(n-1)\widehat{\sigma}^2_{\text{full}}}$.

We demonstrate the general prior with the non-informative conjugate that we used above, i.e., if $g(1/\lambda) = \lambda$ then,

$$\begin{aligned}
\rho_\mu(z|\boldsymbol{f}=\boldsymbol{y}) &= \int_0^\infty g(1/w) w^{\frac{n}{2}-2} \exp\left(-w\chi\right) \mathrm{d}w \\
&= \left(\mathcal{LT}(g(1/t))\right)^{(\frac{n}{2}-2)}|_{t=\chi} = \left(\mathcal{LT}(t)\right)^{(\frac{n}{2}-2)}|_{t=\chi} \\
&= \left(1/u^2\right)^{(\frac{n}{2}-2)}|_{u=\chi} \\
&\propto \chi^{-n/2} = \left(\frac{\gamma - \beta^2/\alpha}{2(c_0 - \boldsymbol{c}^T \mathsf{C}^{-1}\boldsymbol{c})}\right)^{-n/2} \\
&\propto \left(\gamma - \frac{\beta^2}{\alpha}\right)^{-n/2} \\
&\propto \left(\alpha\gamma - \beta^2\right)^{-n/2},
\end{aligned}$$

where we used the fact that the Laplace transform of $g(1/t) = t$ is $1/u^2$. After the transform, taking $(\frac{n}{2} - 2)$th derivative gives us the result. This shows when using a generic prior, it leads to a posterior of the form $\rho_\mu(z|\boldsymbol{f} = \boldsymbol{y}) \propto \mathcal{LT}\{g(1/\cdot)\}^{(\frac{n-4}{2})}(\chi)$ with full Bayes approach, i.e, the posterior $\rho_\mu(z|\boldsymbol{f} = \boldsymbol{y})$ is a function of $1 + \frac{(z-\widehat{\mu}_{\text{full}})^2}{(n-1)\widehat{\sigma}^2_{\text{full}}}$.

Our motivation to experiment with the general prior was to show that it may be possible to infer the prior from the integrand samples. We demonstrated it with the non-informative prior, which shows the possibility to compute the prior from function values. Obtaining an arbitrary prior from the integrand samples is the topic of future work.

2.5 Generalized Cross-Validation

A third parameter optimization technique is *leave-one-out cross-validation* (CV). Let $\widetilde{y}_i = \mathbb{E}[f(\boldsymbol{x}_i)|\boldsymbol{f}_{-i} = \boldsymbol{y}_{-i}]$, where the subscript $-i$ denotes the vector excluding the i^{th} component. This is the conditional expectation of $f(\boldsymbol{x}_i)$ given all data but the function value at $\boldsymbol{x}_i$. The cross-validation criterion, which is to be minimized, is sum of squares of the difference between these conditional expectations and the observed values:

$$\text{CV} = \sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2. \tag{2.10}$$

Let $\mathsf{A} = \mathsf{C}^{-1}$, let $\boldsymbol{\zeta} = \mathsf{A}(\boldsymbol{y} - m\mathbf{1})$, and partition C, A, and $\boldsymbol{\zeta}$ as

$$\mathsf{C} = \begin{pmatrix} c_{ii} & \boldsymbol{C}^T_{-i,i} \\ \boldsymbol{C}_{-i,i} & \mathsf{C}_{-i,-i} \end{pmatrix}, \qquad \mathsf{A} = \begin{pmatrix} a_{ii} & \boldsymbol{A}^T_{-i,i} \\ \boldsymbol{A}_{-i,i} & \mathsf{A}_{-i,-i} \end{pmatrix}, \qquad \boldsymbol{\zeta} = \begin{pmatrix} \zeta_i \\ \boldsymbol{\zeta}_{-i} \end{pmatrix},$$

where the subscript i denotes the i^{th} row or column, and the subscript $-i$ denotes all rows or columns except the i^{th}. Following this notation, Lemma 2.1.1 implies that

$$\widetilde{y}_i = m + \boldsymbol{C}^T_{-i,i}\mathsf{C}^{-1}_{-i,-i}(\boldsymbol{y}_{-i} - m\mathbf{1})$$

$$\zeta_i = a_{ii}(y_i - m) + \boldsymbol{A}^T_{-i,i}(\boldsymbol{y}_{-i} - m\mathbf{1})$$

$$= a_{ii}[(y_i - m) - \boldsymbol{C}_{-i,i}^T \mathsf{C}_{-i,-i}^{-1}(\boldsymbol{y}_{-i} - m\mathbf{1})]$$
$$= a_{ii}(y_i - \widetilde{y}_i).$$

Thus, (2.10) may be re-written as

$$\text{CV} = \sum_{i=1}^{n} \left(\frac{\zeta_i}{a_{ii}} \right)^2, \quad \text{where} \quad \boldsymbol{\zeta} = \mathsf{C}^{-1}(\boldsymbol{y} - m\mathbf{1}).$$

The *generalized cross-validation* criterion (GCV) replaces the i^{th} diagonal element of A in the denominator by the average diagonal element of A [20, 21, 22]:

$$\text{GCV} = \frac{\sum_{i=1}^{n} \zeta_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n} a_{ii}\right)^2} = \frac{(\boldsymbol{y} - m\mathbf{1})^T \mathsf{C}^{-2}(\boldsymbol{y} - m\mathbf{1})}{\left(\frac{1}{n}\text{trace}(\mathsf{C}^{-1})\right)^2}.$$

The loss function GCV depends on m and $\boldsymbol{\theta}$, but not on s. Minimizing the GCV yields

$$m_{\text{GCV}} = \frac{\mathbf{1}^T \mathsf{C}^{-2} \boldsymbol{y}}{\mathbf{1}^T \mathsf{C}^{-2} \mathbf{1}},$$

$$\boldsymbol{\theta}_{\text{GCV}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left\{ \log\left(\boldsymbol{y}^T \left[\mathsf{C}^{-2} - \frac{\mathsf{C}^{-2}\mathbf{1}\mathbf{1}^T\mathsf{C}^{-2}}{\mathbf{1}^T\mathsf{C}^{-2}\mathbf{1}} \right] \boldsymbol{y} \right) - 2\log\left(\text{trace}(\mathsf{C}^{-1})\right) \right\}.$$

Plugging this value of m into (2.4) yields

$$\widehat{\mu}_{\text{GCV}} = \left(\frac{(1 - \mathbf{1}^T\mathsf{C}^{-1}\boldsymbol{c})\mathsf{C}^{-1}\mathbf{1}}{\mathbf{1}^T\mathsf{C}^{-2}\mathbf{1}} + \boldsymbol{c} \right)^T \mathsf{C}^{-1}\boldsymbol{y}.$$

An estimate for s may be obtained by noting that by Lemma 2.1.1,

$$\text{var}[f(\boldsymbol{x}_i) | \boldsymbol{f}_{-i} = \boldsymbol{y}_{-i}] = s^2 a_{ii}^{-1}.$$

Thus, we may estimate s_{GCV} using an argument similar to that used in deriving the GCV and then substituting m_{GCV} for m:

$$s^2 = \text{var}[f(\boldsymbol{x}_i) | \boldsymbol{f}_{-i} = \boldsymbol{y}_{-i}] a_{ii}$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2 a_{ii} = \frac{1}{n}\sum_{i=1}^{n} \frac{\zeta_i^2}{a_{ii}}$$

$$\approx \frac{\frac{1}{n}\sum_{i=1}^{n}\zeta_i^2}{\frac{1}{n}\sum_{i=1}^{n}a_{ii}} = \frac{(\boldsymbol{y}-m\mathbf{1})^T\mathsf{C}^{-2}(\boldsymbol{y}-m\mathbf{1})}{\operatorname{trace}(\mathsf{C}^{-1})} =: s^2_{\mathrm{GCV}}.$$

The confidence interval based on generalized cross-validation corresponds to (2.5) with the GCV estimates for m, s, and $\boldsymbol{\theta}$:

$$\mathrm{err}_{\mathrm{GCV}} = 2.58 s_{\mathrm{GCV}}\sqrt{c_0 - \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c}}, \tag{2.11}$$

$$\mathbb{P}_f\left[|\mu - \widehat{\mu}_{\mathrm{GCV}}| \le \mathrm{err}_{\mathrm{GCV}}\right] = 99\%. \tag{2.12}$$

The methods developed for hyperparameter estimation from the previous sections are summarized as a theorem below:

Theorem 2.5.1. *There are at least three approaches to estimating or integrating out the hyperparameters defining the Gaussian process from which the integrand is drawn: empirical Bayes, full Bayes, and generalized cross-validation. Under these three approaches, we have the following:*

$$m_{\mathrm{EB}} = \frac{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}}, \qquad m_{\mathrm{GCV}} = \frac{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-2}\boldsymbol{y}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-2}\mathbf{1}}, \tag{2.13}$$

$$s^2_{\mathrm{EB}} = \frac{1}{n}\boldsymbol{y}^T\left[\mathsf{C}_{\boldsymbol{\theta}}^{-1} - \frac{\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}}\right]\boldsymbol{y}, \tag{2.14}$$

$$\begin{aligned}\widehat{\sigma}^2_{\mathrm{full}} = {} & \frac{1}{n-1}\boldsymbol{y}^T\left[\mathsf{C}_{\boldsymbol{\theta}}^{-1} - \frac{\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}}\right]\boldsymbol{y} \\ & \times\left[\frac{(1-\boldsymbol{c}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1})^2}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}} + (c_0 - \boldsymbol{c}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{c})\right],\end{aligned} \tag{2.15}$$

$$s^2_{\mathrm{GCV}} = \boldsymbol{y}^T\left[\mathsf{C}_{\boldsymbol{\theta}}^{-2} - \frac{\mathsf{C}_{\boldsymbol{\theta}}^{-2}\mathbf{1}\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-2}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-2}\mathbf{1}}\right]\boldsymbol{y}\left[\operatorname{trace}(\mathsf{C}_{\boldsymbol{\theta}}^{-1})\right]^{-1},$$

$$\boldsymbol{\theta}_{\mathrm{EB}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left\{\log\left(\boldsymbol{y}^T\left[\mathsf{C}_{\boldsymbol{\theta}}^{-1} - \frac{\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}}\right]\boldsymbol{y}\right) + \frac{1}{n}\log(\det(\mathsf{C}_{\boldsymbol{\theta}}))\right\}, \tag{2.16}$$

$$\boldsymbol{\theta}_{\mathrm{GCV}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left\{\log\left(\boldsymbol{y}^T\left[\mathsf{C}_{\boldsymbol{\theta}}^{-2} - \frac{\mathsf{C}_{\boldsymbol{\theta}}^{-2}\mathbf{1}\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-2}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-2}\mathbf{1}}\right]\boldsymbol{y}\right) - \log\left(\operatorname{trace}(\mathsf{C}_{\boldsymbol{\theta}}^{-2})\right)\right\}, \tag{2.17}$$

$$\widehat{\mu}_{\mathrm{EB}} = \widehat{\mu}_{\mathrm{full}} = \left(\frac{(1-\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{c})\mathbf{1}}{\mathbf{1}^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{1}} + \boldsymbol{c}\right)^T\mathsf{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}, \tag{2.18}$$

$$\widehat{\mu}_{\mathrm{GCV}} = \left(\frac{(1 - \mathbf{1}^T \mathsf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{c}) \mathsf{C}_{\boldsymbol{\theta}}^{-1} \mathbf{1}}{\mathbf{1}^T \mathsf{C}_{\boldsymbol{\theta}}^{-2} \mathbf{1}} + \boldsymbol{c} \right)^T \mathsf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}. \tag{2.19}$$

The credible intervals widths, $\mathrm{err}_{\mathrm{CI}}$*, are given by*

$$\mathrm{err}_{\mathsf{x}} = 2.58 s_{\mathsf{x}} \sqrt{c_0 - \boldsymbol{c}^T \mathsf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{c}}, \qquad \mathsf{x} \in \{\mathrm{EB}, \mathrm{GCV}\}, \tag{2.20}$$

$$\mathrm{err}_{\mathrm{full}} = t_{n-1,0.995} \widehat{\sigma}_{\mathrm{full}} > \mathrm{err}_{\mathrm{EB}}. \tag{2.21}$$

The resulting credible intervals are then

$$\mathbb{P}_f \left[|\mu - \widehat{\mu}_{\mathsf{x}}| \le \mathrm{err}_{\mathsf{x}} \right] = 99\%, \qquad \mathsf{x} \in \{\mathrm{EB}, \mathrm{full}, \mathrm{GCV}\}. \tag{2.22}$$

Here $t_{n-1,0.995}$ *denotes the* 99.5 *percentile of a standard Student's t-distribution with* $n - 1$ *degrees of freedom. In the formulas above,* $\boldsymbol{\theta}$ *is assumed to take on the values* $\boldsymbol{\theta}_{\mathrm{EB}}$ *or* $\boldsymbol{\theta}_{\mathrm{GCV}}$ *as appropriate.*

In the theorem above, note that if the original covariance kernel, C, is replaced by bC for some positive constant b, the cubature, $\widehat{\mu}$, the estimates of $\boldsymbol{\theta}$, and the credible interval half-widths, $\mathrm{err}_{\mathsf{x}}$ for $\mathsf{x} \in \{\mathrm{EB}, \mathrm{full}, \mathrm{GCV}\}$, all remain unchanged. The estimates of s^2 are multiplied by b^{-1}, as would be expected.

2.6 Cone of Functions and the Credible interval

In this research we assume that the integrand belongs to a cone of well-behaved functions, $\mathscr{C}$, to make the computations bounded in terms of function data. The concept of cone in general for cubature error analysis can be stated using the error bound definition. Suppose that

$$|\mu(f) - \widehat{\mu}_n(f)| \le \mathrm{err}_{\mathrm{CI}}(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)) \tag{2.23}$$

for some f, which it is 99% of the time under our hypothesis. Also note that our $\mathrm{err}_{\mathrm{CI}}$ (2.20) (2.21) are positively homogeneous functions, meaning,

$$\mathrm{err}_{\mathrm{CI}}(ay_1, \cdots, ay_n) = |a| \, \mathrm{err}_{\mathrm{CI}}(y_1, \cdots, y_n).$$

One can verify the homogeneity of (2.20) and (2.21) easily. Thus if f satisfies (2.23), then

$$\begin{aligned}|\mu(af) - \widehat{\mu}_n(af)| &= |a|\,|\mu(f) - \widehat{\mu}_n(f)| \\ &\le |a|\,\mathrm{err}_{\mathrm{CI}}(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)) \\ &= \mathrm{err}_{\mathrm{CI}}(af(\boldsymbol{x}_1), \cdots, af(\boldsymbol{x}_n))\end{aligned}$$

for all real a. Thus the set of all f satisfying (2.23) is a *cone*, $\mathscr{C}$. Cones of functions satisfy the property that if $f \in \mathscr{C}$ then $af \in \mathscr{C}$.

In the context of Bayesian cubature, one can explain the cone concept beginning with the definition of credible interval (2.5). Let $f \sim \mathcal{GP}$, be an instance of a Gaussian stochastic process:

$$\mathbb{P}_f\left[|\mu(f) - \widehat{\mu}_n(f)| \le \mathrm{err}_{\mathrm{CI}}(f)\right] \ge 99\%.$$

This can be interpreted as $|\mu(f) - \widehat{\mu}_n(f)| \le \mathrm{err}_{\mathrm{CI}}(f)$ with 99% confidence. If f is in the 99% middle of the sample space with $f(\boldsymbol{x}_i) = y_i$ then af is also in the middle 99% of the sample space with $af(\boldsymbol{x}_i) = ay_i$.

We demonstrate the credible interval using the following example. For this purpose, choose a smooth and periodic integrand $f_{\mathrm{smooth}}(\boldsymbol{x}) = \exp(\sum_{\ell=1}^{d} \cos(2\pi x_\ell))$ and another integrand $f_{\mathrm{peaky}}(\boldsymbol{x}) = f_{\mathrm{smooth}} + a_{\mathrm{peaky}} f_{\mathrm{noise}}$ where $a_{\mathrm{peaky}} \in \mathbb{R}$. Here $f_{\mathrm{noise}}(\boldsymbol{x}) = (1 - \exp(2\pi\sqrt{-1}\boldsymbol{x}^T\boldsymbol{\zeta}))$, $\boldsymbol{\zeta} \in \mathbb{R}^d$ is some d-dimensional vector belonging to the dual space of the lattice nodes for some $\{\boldsymbol{x}_i\}_{i=1}^n$. The $\boldsymbol{\zeta}$ in the dual space of lattice nodes implies that $f_{\mathrm{noise}}(\boldsymbol{x}_i) = 0$ at the sampling nodes $\{\boldsymbol{x}_i\}_{i=1}^n$. The f_{nice} is obtained by kernel interpolation of the n samples of f_{smooth} at $\{\boldsymbol{x}_i\}_{i=1}^n$. We chose the Matérn kernel (2.24) for the interpolation. Please note that $f_{\mathrm{peaky}}(\boldsymbol{x}_i) = f_{\mathrm{nice}}(\boldsymbol{x}_i) = f_{\mathrm{smooth}}(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$.

In Figure 2.1, the sampled function values are shown as dots. One can imagine

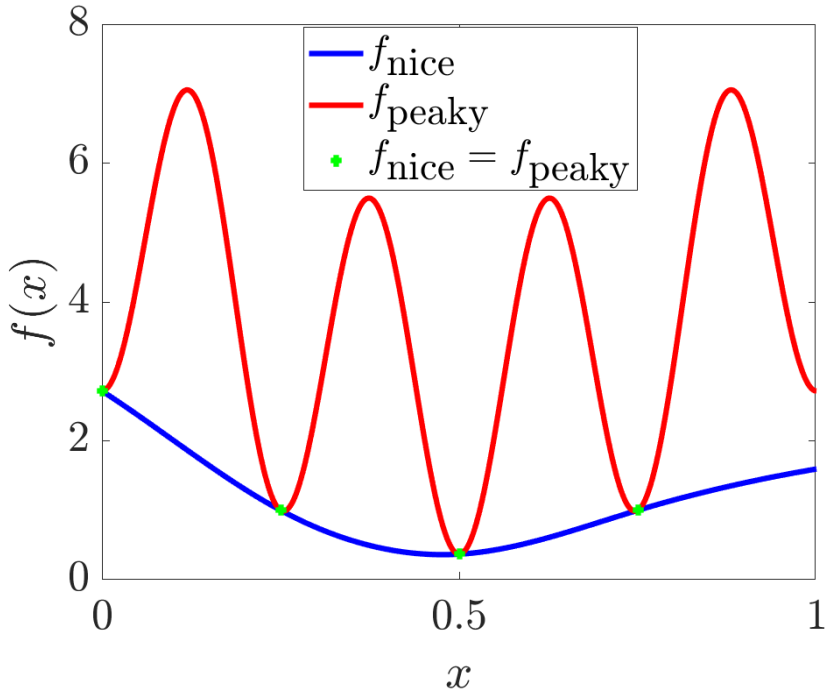

Figure 2.1. Example integrands 1) f_{nice}, a smooth function, 2) f_{peaky}, a peaky function. The function values $f_{\text{peaky}}(\boldsymbol{x}_i) = f_{\text{nice}}(\boldsymbol{x}_i) = f_{\text{smooth}}(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$.

these samples were obtained from f_{nice}, a moderately smoother function or from f_{peaky}, a highly oscillating function. In this example, we used $a_{\text{peaky}} = 2$.

When using $n = 16$ rank-1 lattice points, and $r = 1$ shift-invariant kernel, we get the posterior distribution of μ as shown in Figure 2.2. The true integral value is shown as μ_{smooth} which is at the center of the plot. The integral of the peaky function f_{peaky} lies outside of the 99% of the credible interval given by (2.6), whereas the μ_{nice} falls within.

Our Bayesian cubature algorithms compute the approximate integral using only the samples of the integrand. Estimated integral value of our algorithm closely matches the integral of a smooth function that falls within the middle of the confidence interval. If the true integrand were to resemble the smooth approximate function then the estimated integral will be accurate.

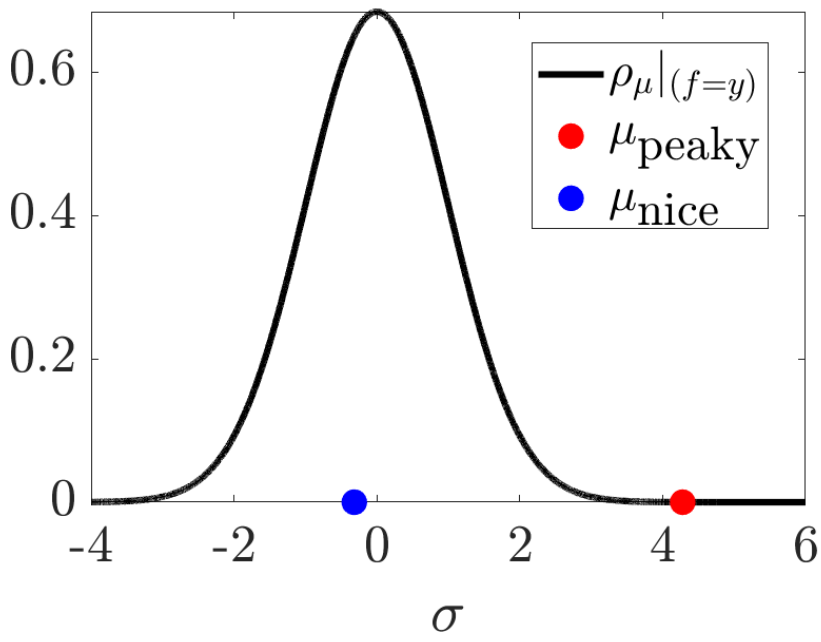

Figure 2.2. Probablity distributions showing the relative position integral of a smooth and a peaky function. f_{nice} lies within the center 99% of the confidence interval, and f_{peaky} lies on the outside of 99% of the confidence interval.

2.7 The Automatic Bayesian Cubature Algorithm

The previous section presents three credible intervals, (2.6), (2.9), and (2.12), for the μ, the desired integral. Each credible interval is based on different assumptions about the hyperparameters m, s, and $\boldsymbol{\theta}$. We stress that one must estimate these hyperparameters or assume a prior distribution on them because the credible intervals are used as stopping criteria for our cubature rule. Since a credible interval makes a statement about a typical function—not an outlier—one must try to ensure that the integrand is a typical draw from the assumed Gaussian process.

Our Bayesian cubature algorithm increases the sample size until the width of the credible interval is small enough. This is accomplished through successively doubling the sample size. The steps are detailed in Algorithm 1.

We recognize that multiple applications of our credible intervals in one run

of the algorithm is not strictly justified. However, if our integrand comes from the middle of the sample space and not the extremes, we expect our automatic Bayesian cubature to approximate the integral within the desired error tolerance with high probability. The example in the next section and the examples in Chapter 7 support that expectation. We also believe that an important factor contributing to the occasional failure of our algorithm is unreasonable parameterizations of the stochastic process from which the integrand is hypothesized to be drawn. Overcoming this latter challenge is a topic for future research.

Algorithm 1 Automatic Bayesian Cubature

Require: a generator for the sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$; a black-box function, f; an absolute error tolerance, $\varepsilon > 0$; the positive initial sample size, n_0; the maximum sample size $n_{\max}$

1: $n \leftarrow n_0,\ n' \leftarrow 0,\ \text{err} \leftarrow \infty$
2: **while** err $> \varepsilon$ and $n \le n_{\max}$ **do**
3: Generate $\{\boldsymbol{x}_i\}_{i=n'+1}^{n}$ and sample $\{f(\boldsymbol{x}_i)\}_{i=n'+1}^{n}$
4: Compute $\boldsymbol{\theta}$ by (2.16) or (2.17)
5: Compute err according to (2.20), (2.21), or (2.11)
6: $n' \leftarrow n,\ n \leftarrow 2n'$
7: **end while**
8: Sample size to compute $\widehat{\mu}$, $n \leftarrow n'$
9: Compute $\widehat{\mu}$, the approximate integral, according to (2.18) or (2.19)
10: **return** $\widehat{\mu}$, n and err

As described above, the computational cost of Algorithm 1 is the sum of the following:

- $\mathcal{O}\big(n\$(f)\big)$ for the integrand data, where $\$(f)$ is the computational cost of a single $f(\boldsymbol{x})$; $\$(f)$ may be large if it is the result of an expensive simulation; $\$(f)$

is typically proportional to d;

- $\mathcal{O}\big(N_{\text{opt}} n^2 \$(C_{\boldsymbol{\theta}})\big)$ for the evaluation of the Gram matrix $\mathsf{C}_{\boldsymbol{\theta}}$, N_{opt} is the number of optimization steps required, and $\$(C_{\boldsymbol{\theta}})$ is the computational cost of a single $C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x})$; $\$(C_{\boldsymbol{\theta}})$ is typically proportional to d; and
- $\mathcal{O}\big(N_{\text{opt}} n^3\big)$ for the matrix inversions and determinant calculations; this cost is independent of d.

As we see in the example in the next section, the cost increases quickly as the n required to meet the error tolerance increases. This motivates the fast Bayesian cubature algorithm presented in Chapter 3.

2.8 Example with the Matérn Kernel

To demonstrate automatic Bayesian cubature consider a Matérn covariance kernel:

$$C_\theta(\boldsymbol{x}, \boldsymbol{t}) = \prod_{\ell=1}^{d} \exp(-\theta|\boldsymbol{x}_\ell - \boldsymbol{t}_\ell|)(1 + \theta|\boldsymbol{x}_\ell - \boldsymbol{t}_\ell|). \tag{2.24}$$

Also, consider the integration problem of evaluating *multivariate Gaussian probabilities*:

$$\mu = \int_{(\boldsymbol{a},\boldsymbol{b})} \frac{\exp\big(-\frac{1}{2}\boldsymbol{t}^T \Sigma^{-1} \boldsymbol{t}\big)}{\sqrt{(2\pi)^d \det(\Sigma)}} \, \mathrm{d}\boldsymbol{t}, \tag{2.25}$$

where $(\boldsymbol{a}, \boldsymbol{b})$ is a finite, semi-infinite or infinite box in $\mathbb{R}^d$. This integral does not have an analytic expression for general Σ, so cubatures are required.

Genz [23] introduced a variable transformation to transform (2.25) into an integral on the unit cube. Not only does this variable transformation accommodate domains that are (semi-)infinite, it also tends to smooth out the integrand better, which expedites the cubature. Let $\Sigma = \mathsf{L}\mathsf{L}^T$ be the Cholesky decomposition where $\mathsf{L} = (l_{jk})_{j,k=1}^d$ is a lower triangular matrix. Iteratively define

$$\alpha_1 = \Phi(a_1), \qquad \beta_1 = \Phi(b_1),$$

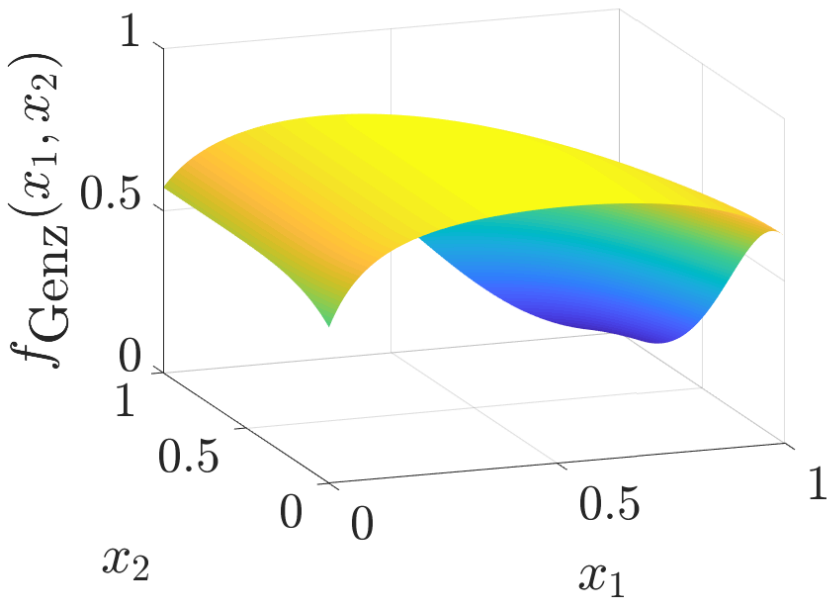

Figure 2.3. The $d = 3$ multivarivate normal probability transformed to an integral of f_{Genz} with $d = 2$. This plot can be reproduced using `IntegrandPlots.m` in GAIL.

$$\alpha_\ell(x_1, ..., x_{\ell-1}) = \Phi\left(\frac{1}{l_{\ell\ell}}\left(a_\ell - \sum_{k=1}^{\ell-1} l_{\ell k}\Phi^{-1}(\alpha_k + x_k(\beta_k - \alpha_k))\right)\right), \quad \ell = 2, ..., d,$$

$$\beta_\ell(x_1, ..., x_{\ell-1}) = \Phi\left(\frac{1}{l_{\ell\ell}}\left(b_\ell - \sum_{k=1}^{\ell-1} l_{\ell k}\Phi^{-1}(\alpha_k + x_k(\beta_k - \alpha_k))\right)\right), \quad \ell = 2, ..., d,$$

$$f_{\text{Genz}}(\boldsymbol{x}) = \prod_{\ell=1}^{d}[\beta_\ell(\boldsymbol{x}) - \alpha_\ell(\boldsymbol{x})]. \tag{2.26}$$

where Φ is the cumulative standard normal distribution function. Then,

$$\mu = \int_{[0,1]^{d-1}} f_{\text{Genz}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

This approach transforms a d' dimensional integral into a $d = d' - 1$ dimensional integral.

We use the following parameter values in the simulation:

$$d = 3, \quad \boldsymbol{a} = \begin{pmatrix} -6 \\ -2 \\ -2 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix}, \quad \mathsf{L} = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 1 & 0.5 \\ 0 & 0 & 0.25 \end{pmatrix}.$$

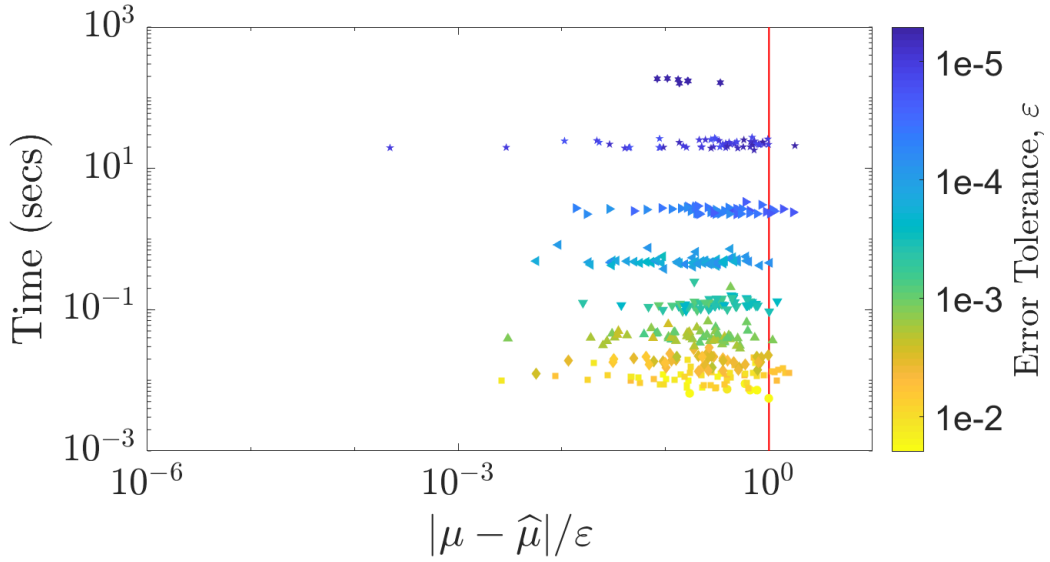

Figure 2.4. Multivariate Gaussian probability: Guaranteed integration using Matérn kernel in $d = 2$ using empirical Bayes stopping criterion within error tolerance ε. This figure can be conditionally reproduced using `matern_guaranteed_plots.m` in GAIL.

The node sets are randomly scrambled Sobol' points [24, 25]. The results are for 400 randomly chosen ε in the interval $[10^{-5}, 10^{-2}]$ as shown in Figure 2.4. In each run, the nodes are randomly scrambled. We observe the algorithm meets the error criterion 95% of the time even-though we used 99% credible intervals. One possible explanation is that the matrix inversions in the algorithm are ill-conditioned leading to numerical inaccuracies. Another possible explanation is that this Matérn covariance kernel is not a good match for the integrand.

On our test computer, it took more than an hour to compute $\widehat{\mu}_n$ with $n = 2^{14}$. As shown in Figure 2.5, the computation time increases rapidly with n. The empirical

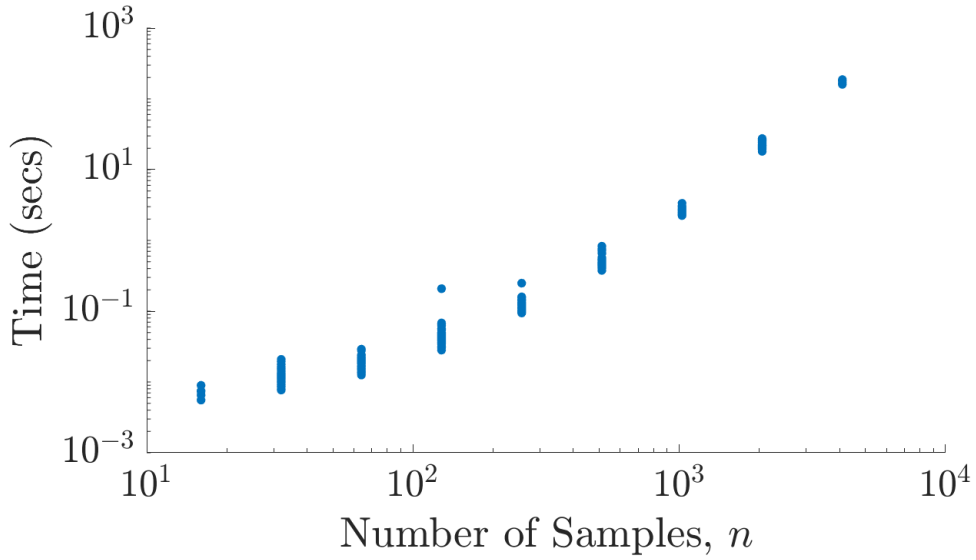

Figure 2.5. Multivariate Gaussian probability estimated using Matérn kernel in $d = 2$ using empirical Bayes stopping criterion. Computation time rapidly increases with increase of n. This figure can be conditionally reproduced using `matern_guaranteed_plots.m` in GAIL.

Bayes estimation of $\boldsymbol{\theta}$, which requires repeated evaluation of the objective function, is the most time consuming of all. This is due to fact that the objective function needs to be computed multiple times in every iteration to find its minimum. It takes tens of seconds to compute $\widehat{\mu}_n$ with $\varepsilon = 10^{-5}$. In contrast, this example in Chapter 7 take less than a hundredth of a second to compute $\widehat{\mu}_n$ with the same ε using our new algorithm. Not only is the Bayesian cubature with the Matérn kernel slow, but also $\mathsf{C}_{\boldsymbol{\theta}}$ becomes highly ill-conditioned as n increases. So, Algorithm 1 in its current form is impractical when n must be large.

CHAPTER 3

FAST AUTOMATIC BAYESIAN CUBATURE

The generic automatic Bayesian cubature algorithm described in the previous section requires $\mathcal{O}\big(n\$(f) + N_{\text{opt}}[n^2\$(C_{\boldsymbol{\theta}}) + n^3]\big)$ operations to compute the cubature. Now we explain how to speed up the calculations. A key is to choose covariance kernels that match the nodes, $\{\boldsymbol{x}_i\}_{i=1}^n$, so that the vector-matrix operations required by Bayesian cubature can be accomplished using *fast Bayesian transforms* at a computational cost of $\mathcal{O}\big(n\$(f) + N_{\text{opt}}[n\$(C_{\boldsymbol{\theta}}) + n\log(n)]\big)$. We develop the concept of fast Bayesian transform and show how matching kernels and nodes with three key assumptions are used.

3.1 Fast Bayesian Transform Kernel

We make some assumptions about the relationship between the covariance kernel and the nodes. In Chapter 4 these assumptions are shown to hold for rank-1 lattices and shift-invariant kernels and again in Chapter 5 to hold for Sobol' nodes and Walsh kernels. Although the integrands and covariance kernels are real, it is convenient to allow related vectors and matrices to be complex. A relevant example is the fast Fourier transform (FFT) of a real-valued vector, which is a complex-valued vector.

We introduce some further notation

$$\begin{aligned}
\mathsf{C} = \mathsf{C}_{\boldsymbol{\theta}} &= \Big(C_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\Big)_{i,j=1}^n = (\boldsymbol{C}_1, ..., \boldsymbol{C}_n) \\
&= \frac{1}{n}\mathsf{V}\Lambda\mathsf{V}^H, \qquad \mathsf{V}^H = n\mathsf{V}^{-1}, \\
\mathsf{V} &= (\boldsymbol{v}_1, ..., \boldsymbol{v}_n)^T = (\boldsymbol{V}_1, ..., \boldsymbol{V}_n) \\
\mathsf{C}^p &= \frac{1}{n}\mathsf{V}\Lambda^p\mathsf{V}^H, \qquad \forall p \in \mathbb{Z},
\end{aligned} \tag{3.1}$$

where V^H is the Hermitian of V, $\boldsymbol{C}_1, \cdots, \boldsymbol{C}_n$ are columns of C, $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n$ are

columns of V, and $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ are rows of V. The columns of matrix V are eigenvectors of C, and Λ is a diagonal matrix of eigenvalues of C. In this and later sections, we drop the $\boldsymbol{\theta}$ dependence of various quantities for simplicity of notation. The normalization of V assumed in (3.1) conveniently allows the first eigenvector, $\boldsymbol{V}_1$, to be the vector of ones in (3.2b) below. For any $n \times 1$ vector $\boldsymbol{b}$, define the notation $\widetilde{\boldsymbol{b}} := \mathsf{V}^H \boldsymbol{b}$.

We make three assumptions that allow the fast computation:

$$\mathsf{V} \text{ may be identified analytically,} \tag{3.2a}$$

$$\boldsymbol{v}_1 = \boldsymbol{V}_1 = \mathbf{1}, \tag{3.2b}$$

$$\text{Computing } \mathsf{V}^H \boldsymbol{b} \text{ requires only } \mathcal{O}(n \log n) \text{ operations } \forall \boldsymbol{b}. \tag{3.2c}$$

We call the transformation $\boldsymbol{b} \mapsto \mathsf{V}^H \boldsymbol{b}$ a *fast Bayesian transform* and $C_{\boldsymbol{\theta}}$ a *fast Bayesian transform kernel* for the matching nodes $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$.

Under assumptions (3.2) the eigenvalues may be identified as the fast Bayesian transform of the first column of C:

$$\begin{aligned} \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \Lambda \mathbf{1} = \Lambda \boldsymbol{v}_1^* = \underbrace{\left(\frac{1}{n} \mathsf{V}^H \mathsf{V} \right)}_{\mathsf{I}} \Lambda \boldsymbol{v}_1^* \\ = \mathsf{V}^H \left(\frac{1}{n} \mathsf{V} \Lambda \boldsymbol{v}_1^* \right) = \mathsf{V}^H \boldsymbol{C}_1 = \widetilde{\boldsymbol{C}}_1, \end{aligned} \tag{3.3}$$

where I is the identity matrix and $\boldsymbol{v}_1^*$ is the complex conjugate of the first row of V. Also note that the fast Bayesian transform of $\mathbf{1}$ has a simple form

$$\widetilde{\mathbf{1}} = \mathsf{V}^H \mathbf{1} = \mathsf{V}^H \boldsymbol{V}_1 = \begin{pmatrix} n \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Many of the terms that arise in the calculations in Algorithm 1 take the form $\boldsymbol{a}^T \mathsf{C}^p \boldsymbol{b}$ for real $\boldsymbol{a}$ and $\boldsymbol{b}$ and integer p. These can be calculated via the transforms

$\widetilde{\boldsymbol{a}} = \mathsf{V}^H\boldsymbol{a}$ and $\widetilde{\boldsymbol{b}} = \mathsf{V}^H\boldsymbol{b}$ as

$$\boldsymbol{a}^T\mathsf{C}^p\boldsymbol{b} = \frac{1}{n}\boldsymbol{a}^T\mathsf{V}\Lambda^p\mathsf{V}^H\boldsymbol{b} = \frac{1}{n}\widetilde{\boldsymbol{a}}^H\Lambda^p\widetilde{\boldsymbol{b}} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i^p\widetilde{a}_i^*\widetilde{b}_i,$$

Note that $\widetilde{\boldsymbol{a}}^*$ appears on the right side of this equation because $\boldsymbol{a}^T\mathsf{V} = (\mathsf{V}^H\boldsymbol{a})^* = \widetilde{\boldsymbol{a}}^*$. In particular,

$$\begin{aligned}
\mathbf{1}^T\mathsf{C}^{-p}\mathbf{1} &= \frac{n}{\lambda_1^p}, & \mathbf{1}^T\mathsf{C}^{-p}\boldsymbol{y} &= \frac{\widetilde{y}_1}{\lambda_1^p},\\
\boldsymbol{y}^T\mathsf{C}^{-p}\boldsymbol{y} &= \frac{1}{n}\sum_{i=1}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^p}, & \boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1} &= \frac{\widetilde{c}_1}{\lambda_1},\\
\boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{y} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{c}_i^*\widetilde{y}_i}{\lambda_i}, & \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c} &= \frac{1}{n}\sum_{i=1}^{n}\frac{|\widetilde{c}_i|^2}{\lambda_i},
\end{aligned}$$

where $\widetilde{\boldsymbol{y}} = \mathsf{V}^H\boldsymbol{y}$ and $\widetilde{\boldsymbol{c}} = \mathsf{V}^H\boldsymbol{c}$. For any real $\boldsymbol{b}$, with $\widetilde{\boldsymbol{b}} = \mathsf{V}^H\boldsymbol{b}$, it follows that $\widetilde{b}_1$ is real since the first row of V^H is $\mathbf{1}$.

The covariance kernel used in practice also may satisfy an additional assumption:

$$\int_{[0,1]^d} C(\boldsymbol{t},\boldsymbol{x})\,\mathrm{d}\boldsymbol{t} = 1 \qquad \forall \boldsymbol{x} \in [0,1]^d, \tag{3.4}$$

which implies that $c_{0\boldsymbol{\theta}} = 1$ and $\boldsymbol{c}_{\boldsymbol{\theta}} = \mathbf{1}$. Under (3.4), the expressions above may be further simplified:

$$\boldsymbol{c}^T\mathsf{C}^{-1}\mathbf{1} = \boldsymbol{c}^T\mathsf{C}^{-1}\boldsymbol{c} = \frac{n}{\lambda_1}.$$

We use the fast Bayesian transform to speedup the computation of the hyperparameter $\boldsymbol{\theta}$, the credible interval width $\mathrm{err}_{\mathrm{CI}}$, and the integral estimate $\widehat{\mu}$ that we presented in Theorem 2.5.1 as shown next. The assumptions and results in this chapter lead to the following theorem.

Theorem 3.1.1. *Under assumptions* (3.2)*, the parameters and credible interval half-widths in Theorem 2.5.1 may be expressed in terms of the fast Bayesian transforms of the integrand data, the first column of the Gram matrix,* c_0*, and* $\boldsymbol{c}$ *as follows:*

$$m_{\mathrm{EB}} = m_{\mathrm{full}} = m_{\mathrm{GCV}} = \frac{\widetilde{y}_1}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$
\begin{aligned}
s^2_{\mathrm{EB}} &= \frac{1}{n^2}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i},\\
\widehat{\sigma}^2_{\mathrm{full}} &= \frac{1}{n(n-1)}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\left[\frac{\lambda_1}{n}\left(1-\frac{\widetilde{c}_1}{\lambda_1}\right)^2+\left(c_0-\frac{1}{n}\sum_{i=1}^{n}\frac{|\widetilde{c}_i|^2}{\lambda_i}\right)\right],\\
s^2_{\mathrm{GCV}} &= \frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\left[\sum_{i=1}^{n}\frac{1}{\lambda_i}\right]^{-1},
\end{aligned}
$$

$$
\boldsymbol{\theta}_{\mathrm{EB}} = \operatorname*{argmin}_{\boldsymbol{\theta}}\left[\log\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right)+\frac{1}{n}\sum_{i=1}^{n}\log(\lambda_i)\right], \tag{3.5a}
$$

$$
\boldsymbol{\theta}_{\mathrm{GCV}} = \operatorname*{argmin}_{\boldsymbol{\theta}}\left[\log\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right)-2\log\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)\right], \tag{3.5b}
$$

$$
\begin{aligned}
\widehat{\mu}_{\mathrm{EB}} = \widehat{\mu}_{\mathrm{full}} = \widehat{\mu}_{\mathrm{GCV}} &= \frac{\widetilde{y}_1}{n}+\frac{1}{n}\sum_{i=2}^{n}\frac{\widetilde{c}_i^{*}\widetilde{y}_i}{\lambda_i},\\
\mathrm{err}_{\mathrm{EB}} &= \frac{2.58}{n}\sqrt{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\left(c_0-\frac{1}{n}\sum_{i=1}^{n}\frac{|\widetilde{c}_i|^2}{\lambda_i}\right)},\\
\mathrm{err}_{\mathrm{full}} &= t_{n-1,0.995}\widehat{\sigma}_{\mathrm{full}},\\
\mathrm{err}_{\mathrm{GCV}} &= \frac{2.58}{n}\left\{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i}\right]^{-1}\left(c_0-\frac{1}{n}\sum_{i=1}^{n}\frac{|\widetilde{c}_i|^2}{\lambda_i}\right)\right\}^{1/2}.
\end{aligned}
$$

Under the further assumption (3.4)*, it follows that*

$$
\widehat{\mu}_{\mathrm{EB}} = \widehat{\mu}_{\mathrm{full}} = \widehat{\mu}_{\mathrm{GCV}} = \frac{\widetilde{y}_1}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3.6}
$$

and so $\widehat{\mu}$ *is simply the sample mean. Also, under assumption* (3.4)*, the credible interval half-widths simplify to*

$$
\mathrm{err}_{\mathrm{EB}} = \frac{2.58}{n}\sqrt{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\left(1-\frac{n}{\lambda_1}\right)}, \tag{3.7a}
$$

$$
\mathrm{err}_{\mathrm{full}} = t_{n-1,0.995}\sqrt{\frac{1}{n(n-1)}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\left(\frac{\lambda_1}{n}-1\right)}, \tag{3.7b}
$$

$$\mathrm{err}_{\mathrm{GCV}} = \frac{2.58}{n}\left\{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i}\right]^{-1}\left(1-\frac{n}{\lambda_1}\right)\right\}^{1/2}. \tag{3.7c}$$

In the formulas for the credible interval half-widths and $\boldsymbol{\lambda}$ *depends on* $\boldsymbol{\theta}$*, and* $\boldsymbol{\theta}$ *is assumed to take on the values* $\boldsymbol{\theta}_{\mathrm{EB}}$ *or* $\boldsymbol{\theta}_{\mathrm{GCV}}$ *as appropriate.*

The remaining part of the chapter proves this theorem. We apply the fast Bayesian transform to speedup empirical Bayes, full Bayes and Generalized cross validation stopping criteria.

3.2 Empirical Bayes

Under assumptions (3.2), the empirical Bayes parameters in (2.13), (2.14), (2.16) (2.18), and (2.20) can be expressed in terms of the fast Bayesian transforms of the function data, the first column of the Gram matrix, and $\boldsymbol{c}$ as follows:

$$\begin{aligned}
m_{\mathrm{EB}} &= \frac{\widetilde{y}_1}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i,\\
s^2_{\mathrm{EB}} &= \frac{1}{n^2}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i},\\
\boldsymbol{\theta}_{\mathrm{EB}} &= \operatorname*{argmin}_{\boldsymbol{\theta}}\left[\log\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right)+\frac{1}{n}\sum_{i=1}^{n}\log(\lambda_i)\right],\\
\widehat{\mu}_{\mathrm{EB}} &= \frac{\widetilde{y}_1}{n}+\frac{1}{n}\sum_{i=2}^{n}\frac{\widetilde{c}_i^{*}\widetilde{y}_i}{\lambda_i},\\
\mathrm{err}_{\mathrm{EB}} &= \frac{2.58}{n}\sqrt{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\left(c_0-\frac{1}{n}\sum_{i=1}^{n}\frac{|\widetilde{c}_i|^2}{\lambda_i}\right)},
\end{aligned}$$

The quantities on the right hand sides can be obtained in $\mathcal{O}(n\log n)$ operations by fast Bayesian transforms.

Under the further assumption (3.4) it follows that

$$\widehat{\mu}_{\mathrm{EB}} = \frac{\widetilde{y}_1}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\mathrm{err}_{\mathrm{EB}} = \frac{2.58}{n}\sqrt{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\left(1-\frac{n}{\lambda_1}\right)}.$$

Thus, in this case $\widehat{\mu}$ is simply the sample mean.

3.2.1 Gradient of the objective function using fast Bayesian transform.

We refer back to Section 2.3.1, where we discuss about using gradient descent for hyperparameter search but the computational cost is of $\mathcal{O}(N_{\mathrm{opt}}n^3)$. Here we develop a techniques to speed up the computation. If V does not depend on $\boldsymbol{\theta}$ then one can fast compute the derivative of Gram matrix C. Starting from the definition (3.1) and taking derivative w.r.t. θ_ℓ,

$$\frac{\partial \mathsf{C}}{\partial \theta_\ell} = \frac{1}{n}\mathsf{V}\frac{\partial \mathsf{\Lambda}}{\partial \theta_\ell}\mathsf{V}^H = \frac{1}{n}\mathsf{V}\bar{\mathsf{\Lambda}}_{(\ell)}\mathsf{V}^H,$$

$$\text{where} \quad \bar{\mathsf{\Lambda}}_{(\ell)} = \mathrm{diag}(\bar{\boldsymbol{\lambda}}_{(\ell)}), \quad \text{and}$$

$$\bar{\boldsymbol{\lambda}}_{(\ell)} = \frac{\partial \boldsymbol{\lambda}}{\partial \theta_\ell} = \left(\frac{\partial \lambda_i}{\partial \theta_\ell}\right)_{i=1}^{n} = \left(\frac{\partial}{\partial \theta_\ell}\mathsf{V}^H\boldsymbol{C}_1\right) = \mathsf{V}^H\left(\frac{\partial}{\partial \theta_\ell}C_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_i)\right)_{i=1}^{n}, \tag{3.8}$$

where we used the fast Bayesian transform property (3.3). We use the notation $\bar{\boldsymbol{\lambda}}_{(\ell)} = \mathsf{V}^H\bar{\boldsymbol{C}}_{1(\ell)}$ to denote the derivative of the eigenvalue $\boldsymbol{\lambda}_{(\ell)}$, where $\bar{\boldsymbol{C}}_{1(\ell)}$ denotes the first row of the gram matrix after taking the derivative in the ℓth variable, i.e.

$$\bar{\boldsymbol{C}}_{1(\ell)} = \left(\frac{\partial}{\partial \theta_\ell}C_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_i)\right)_{i=1}^{n}.$$

The goal is to compute the derivative of the objective function faster. First, let's rewrite the objective function from (3.5a) in two parts,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})_{\mathrm{EB}} &= \underbrace{\frac{1}{n}\log(\det\ \mathsf{C})}_{\mathcal{L}_{|\mathsf{C}|}} + \underbrace{\log\left((\boldsymbol{y}-m_{\mathrm{EB}}\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y}-m_{\mathrm{EB}}\mathbf{1})\right)}_{\mathcal{L}_{\boldsymbol{y}}},\\ &=: \mathcal{L}_{|\mathsf{C}|} + \mathcal{L}_{\boldsymbol{y}}.\end{aligned}$$

Now, take the derivative:

$$\frac{\partial}{\partial \theta_\ell}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial}{\partial \theta_\ell}\mathcal{L}_{|\mathsf{C}|} + \frac{\partial}{\partial \theta_\ell}\mathcal{L}_{\boldsymbol{y}}\ .$$

Now we tackle the individual terms,

$$\begin{aligned}\frac{\partial}{\partial\theta_\ell}\mathcal{L}_{|\mathsf{C}|} &= \frac{\partial}{\partial\theta_\ell}\frac{1}{n}\log(\det\,\mathsf{C}),\\ &= \frac{1}{n}\text{trace}\left(\mathsf{C}^{-1}\frac{\partial\mathsf{C}}{\partial\boldsymbol{\theta}_\ell}\right) = \frac{1}{n}\text{trace}\left(\mathsf{V}\boldsymbol{\Lambda}^{-1}\mathsf{V}^H\frac{1}{n}\mathsf{V}\overline{\boldsymbol{\Lambda}}_{(\ell)}\mathsf{V}^H\right),\\ &= \frac{1}{n}\text{trace}\big(\mathsf{V}\boldsymbol{\Lambda}^{-1}\overline{\boldsymbol{\Lambda}}_{(\ell)}\mathsf{V}^H\big), \quad \text{where we used } \mathsf{V}^H\mathsf{V}=n,\\ &= \frac{1}{n}\text{trace}\left(\mathsf{V}\,\text{diag}\left(\frac{\overline{\lambda}_{i(\ell)}}{\lambda_i}\right)_{i=1}^{n}\mathsf{V}^H\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{\overline{\lambda}_{i(\ell)}}{\lambda_i},\end{aligned}$$

where we used the fact from [26],

$$\log(\det\,\mathsf{C}) = \text{trace}(\log(\mathsf{C})).$$

Part of the $\mathcal{L}_{\boldsymbol{y}}$ was already simplified using the fast Bayesian transform,

$$(\boldsymbol{y}-m_{\text{EB}}\mathbf{1})^T\mathsf{C}^{-1}(\boldsymbol{y}-m_{\text{EB}}\mathbf{1}) = \frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}.$$

Using the above result,

$$\begin{aligned}\frac{\partial}{\partial\theta_\ell}\mathcal{L}_{\boldsymbol{y}} &= \frac{\partial}{\partial\boldsymbol{\theta}_\ell}\log\left(\frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right),\\ &= \left(\frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right)^{-1}\frac{\partial}{\partial\boldsymbol{\theta}_\ell}\left(\frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right),\\ &= \left(\frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right)^{-1}\frac{1}{n}\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\left(-\frac{\partial\lambda_i}{\partial\boldsymbol{\theta}_\ell}\right),\\ &= -\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i}\right)^{-1}\left(\sum_{i=2}^{n}|\widetilde{y}_i|^2\frac{\bar{\lambda}_{i(\ell)}}{\lambda_i^2}\right).\end{aligned}$$

Finally, using the above results,

$$\frac{\partial}{\partial\theta_\ell}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})_{\text{EB}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\bar{\lambda}_{i(\ell)}}{\lambda_i} - \left(\sum_{i=2}^{n}\frac{|\tilde{\boldsymbol{y}}_i|^2\,\bar{\lambda}_{i(\ell)}}{\lambda_i^2}\right)\left(\sum_{i=2}^{n}\frac{|\tilde{\boldsymbol{y}}_i|^2}{\lambda_\ell}\right)^{-1}, \tag{3.9}$$

where $\bar{\lambda}_{i(\ell)}$ is the derivative of the ith eigenvalue of C in the ℓth variable. Please recollect the gradient descent proposed in (2.7) can be computed faster in $\mathcal{O}(n\log n)$ using the result (3.9). A technique to compute this faster is discussed in Section 3.5.

3.3 Full Bayes

For the full Bayes approach the cubature is the same as for empirical Bayes. We also defer to empirical Bayes to estimate the parameter $\boldsymbol{\theta}$. The width of the confidence interval is $\mathrm{err}_{\mathrm{full}} := t_{n_j-1,0.995}\widehat{\sigma}_{\mathrm{full}}$, where $\widehat{\sigma}^2_{\mathrm{full}}$ can also be computed swiftly under assumptions (3.2):

$$\widehat{\sigma}^2_{\mathrm{full}} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i} \left[\frac{\lambda_1}{n} \left(1 - \frac{\widetilde{c}_1}{\lambda_1} \right)^2 + \left(c_0 - \frac{1}{n} \sum_{i=1}^{n} \frac{|\widetilde{c}_i|^2}{\lambda_i} \right) \right],$$

Under assumption (3.4) further simplification can be made:

$$\widehat{\sigma}^2_{\mathrm{full}} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i} \left(\frac{\lambda_1}{n} - 1 \right),$$

It follows that

$$\mathrm{err}_{\mathrm{full}} = t_{n_j-1,0.995} \sqrt{\frac{1}{n(n-1)} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i} \left(\frac{\lambda_1}{n} - 1 \right)}.$$

3.4 Generalized Cross-Validation

GCV yields a different cubature, which nevertheless can also be computed quickly using the fast Bayesian transform. Under assumptions (3.2):

$$\begin{aligned}
m_{\mathrm{GCV}} &= m_{\mathrm{EB}} = \frac{\widetilde{y}_1}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i, \\
s^2_{\mathrm{GCV}} &:= \frac{1}{n} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i^2} \left[\sum_{i=1}^{n} \frac{1}{\lambda_i} \right]^{-1}, \\
\boldsymbol{\theta}_{\mathrm{GCV}} &= \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\log \left(\sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i^2} \right) - 2 \log \left(\sum_{i=1}^{n} \frac{1}{\lambda_i} \right) \right], \\
\widehat{\mu}_{\mathrm{GCV}} &= \widehat{\mu}_{\mathrm{EB}} = \frac{\widetilde{y}_1}{n} + \frac{1}{n} \sum_{i=2}^{n} \frac{\widetilde{c}_i^* \widetilde{y}_i}{\lambda_i}, \\
\mathrm{err}_{\mathrm{GCV}} &= \frac{2.58}{n} \left\{ \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i^2} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i} \right]^{-1} \times \left(c_0 - \frac{1}{n} \sum_{i=1}^{n} \frac{|\widetilde{c}_i|^2}{\lambda_i} \right) \right\}^{1/2}.
\end{aligned} \tag{3.10}$$

Moreover, under further assumption (3.4) it follows that

$$\widehat{\mu}_{\mathrm{GCV}} = \widehat{\mu}_{\mathrm{EB}} = \widehat{\mu}_{\mathrm{full}} = \frac{\widetilde{y}_1}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

$$\mathrm{err}_{\mathrm{GCV}} = \frac{2.58}{n}\left\{\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i}\right]^{-1}\times\left(1-\frac{n}{\lambda_1}\right)\right\}^{1/2}.$$

In this case too, $\widehat{\mu}$ is simply the sample mean.

3.4.1 Gradient of the objective function. Using the results obtained from the Section Section 3.2.1 with empirical Bayes, one can reduce the computational cost of the derivative of the objective function in (3.10),

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})_{\mathrm{GCV}} = \log\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right) - 2\log\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right).$$

Using the similar techniques from Section 3.2.1, the derivative of the objective function w.r.t θ_ℓ:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_\ell}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})_{\mathrm{GCV}}\\
&= \left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right)^{-1}\frac{\partial}{\partial\theta_\ell}\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right) - 2\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)^{-1}\frac{\partial}{\partial\theta_\ell}\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)\\
&= \left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right)^{-1}\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^3}(-2)\frac{\partial\lambda_i}{\partial\theta_\ell}\right)\\
&\qquad\qquad - 2\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{1}{\lambda_i^2}(-1)\frac{\partial\lambda_i}{\partial\theta_\ell}\right)\\
&= -2\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right)^{-1}\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2\,\bar{\lambda}_{i(\ell)}}{\lambda_i^3}\right) + 2\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\bar{\lambda}_{i(\ell)}}{\lambda_i^2}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\partial}{\partial\theta_\ell}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})_{\mathrm{GCV}} = -2\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2}{\lambda_i^2}\right)^{-1}&\left(\sum_{i=2}^{n}\frac{|\widetilde{y}_i|^2\,\bar{\lambda}_{i(\ell)}}{\lambda_i^3}\right)\\
&+ 2\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\bar{\lambda}_{i(\ell)}}{\lambda_i^2}\right), \quad (3.11)
\end{aligned}$$

where $\bar{\lambda}_{i(\ell)}$ is the derivative of the ith eigenvalue of the Gram matrix, C, in the ℓth variable. We discuss a technique to compute $\bar{\lambda}_{i(\ell)}$ in the next section below.

3.5 Product Kernels

In this research, we use product kernels in the demonstrations and numerical implementations. They got nice properties which are helpful to obtain analytical results easily. Product kernels in d dimensions are of the form,

$$C_{\boldsymbol{\theta}}(\boldsymbol{t},\boldsymbol{x}) = \prod_{\ell=1}^{d}\Big[1-\eta_\ell\ \mathfrak{C}(x_\ell,t_\ell)\Big] \tag{3.12}$$

where η_ℓ is called shape parameter in the ℓth variable for $\ell = 1,\cdots,d$, and $\mathfrak{C}$ is some positive definite function. Our goal is to compute $\bar{\lambda}_{i(\ell)}$ for which the kernel derivative is necessary. The derivative of the product kernels can be obtained easily. Please note that $\boldsymbol{\theta}$ denotes all the hyper parameters of the kernel C where η is one of them and called the shape parameter.

3.5.1 Derivative of the product kernel when $\eta_1 = \cdots = \eta_d = \eta$. It was suggested to use gradient descent to find optimal shape parameter in Section 2.3.1. In this section, we compute the gradient for product kernels. When the $\eta_1 = \cdots = \eta_d = \eta$, the derivative of a product kernel w.r.t. η can be obtained as below,

$$\begin{aligned}
\frac{\partial}{\partial\eta}C_{\boldsymbol{\theta}}(\boldsymbol{t},\boldsymbol{x}) &= \frac{\partial}{\partial\eta}\prod_{j=1}^{d}\Big[1-\eta\mathfrak{C}(x_j,t_j)\Big]\\
&= \sum_{\ell=1}^{d}\prod_{j=1,j\neq\ell}^{d}\Big[1-\eta\mathfrak{C}(x_j,t_j)\Big]\Big(-\mathfrak{C}(x_\ell,t_\ell)\Big)\\
&= \prod_{j=1}^{d}\Big[1-\eta\mathfrak{C}(x_j,t_j)\Big]\sum_{\ell=1}^{d}\frac{\Big(-\mathfrak{C}(x_\ell,t_\ell)\Big)}{1-\eta\mathfrak{C}(x_\ell,t_\ell)}\\
&= C_{\boldsymbol{\theta}}(\boldsymbol{t},\boldsymbol{x})\frac{1}{\eta}\sum_{\ell=1}^{d}\frac{\Big(1-\eta\mathfrak{C}(x_\ell,t_\ell)-1\Big)}{1-\eta\mathfrak{C}(x_\ell,t_\ell)}\\
&= C_{\boldsymbol{\theta}}(\boldsymbol{t},\boldsymbol{x})\frac{1}{\eta}\sum_{\ell=1}^{d}\Big(1-\frac{1}{1-\eta\mathfrak{C}(x_\ell,t_\ell)}\Big)\\
&= (d/\eta)\underbrace{\Big(\prod_{j=1}^{d}\Big[1-\eta\mathfrak{C}(x_j,t_j)\Big]\Big)}_{C_{\boldsymbol{\theta}}(\boldsymbol{t},\boldsymbol{x})}\Big(1-\frac{1}{d}\sum_{\ell=1}^{d}\frac{1}{1-\eta\mathfrak{C}(x_\ell,t_\ell)}\Big).
\end{aligned}$$

Thus,

$$\frac{\partial}{\partial \eta} C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) = (d/\eta) C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) \left(1 - \frac{1}{d} \sum_{\ell=1}^{d} \frac{1}{1 - \eta \mathfrak{C}(x_\ell, t_\ell)} \right).$$

3.5.1.1 When η_ℓ is different for each $\ell = 1, \cdots, d$. In this case, we will have a vector of length d shape parameters. Derivative of the kernel, $C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x})$ (3.12), with respect to η_ℓ is,

$$\begin{aligned}
\frac{\partial}{\partial \eta_\ell} C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) &= \frac{\partial}{\partial \eta_\ell} \prod_{j=1}^{d} \Big[1 - \eta_j \mathfrak{C}(x_j, t_j) \Big], \quad \text{where} \quad \ell = 1, \cdots, d \\
&= \prod_{j=1, j \neq \ell}^{d} \Big[1 - \eta_j \mathfrak{C}(x_j, t_j) \Big] \Big(-\mathfrak{C}(x_\ell, t_\ell) \Big) \\
&= \prod_{j=1}^{d} \Big[1 - \eta_j \mathfrak{C}(x_j, t_j) \Big] \frac{\Big(-\mathfrak{C}(x_\ell, t_\ell) \Big)}{1 - \eta_\ell \mathfrak{C}(x_\ell, t_\ell)} \\
&= C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) \frac{1}{\eta_\ell} \frac{\Big(1 - \eta_\ell \mathfrak{C}(x_\ell, t_\ell) - 1 \Big)}{1 - \eta_\ell \mathfrak{C}(x_\ell, t_\ell)} \\
&= C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) \frac{1}{\eta_\ell} \left(1 - \frac{1}{1 - \eta_\ell \mathfrak{C}(x_\ell, t_\ell)} \right) \\
&= \frac{1}{\eta_\ell} \underbrace{\left(\prod_{j=1}^{d} \Big[1 - \eta \mathfrak{C}(x_j, t_j) \Big] \right)}_{C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x})} \left(1 - \frac{1}{1 - \eta_\ell \mathfrak{C}(x_\ell, t_\ell)} \right).
\end{aligned}$$

Thus,

$$\frac{\partial}{\partial \eta_\ell} C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) = \frac{1}{\eta_\ell} C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) \left(1 - \frac{1}{1 - \eta_\ell \mathfrak{C}(x_\ell, t_\ell)} \right).$$

Please note that the above derivatives do not depend on $\mathfrak{C}(x, t)$ and most importantly these computations are applicable to any product kernel of the form (3.12). The $\bar{\lambda}_{i(\ell)}$ can be computed now using (3.8) with the computed kernel derivative, $\frac{\partial}{\partial \eta_\ell} C_{\boldsymbol{\theta}}$.

3.5.2 Shape parameter search using steepest descent. Using the obtained derivative of the eigenvalues, $\bar{\lambda}_{i(\ell)}$, one can easily compute the gradient of the objective

function (3.9) or (3.11). This can be further used to implement the steepest descent search as introduced in Section 2.3.1

$$\eta_\ell^{(j+1)} = \eta_\ell^{(j)} - \nu \frac{\partial}{\partial \eta_\ell} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}), \quad j = 0, 1, \cdots, \quad \ell = 1, \cdots, d$$

where ν is the step size for the gradient descent, j is the iteration index, and $\frac{\partial}{\partial \eta_\ell} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y})$ is either (3.9) or (3.11) depending on the choice of the hyperparameter search method. The parameter η_ℓ is usually searched in the whole $\mathbb{R}$ by using the simple domain transformation as explained in Section 6.2.

CHAPTER 4

INTEGRATION LATTICES AND SHIFT INVARIANT KERNELS

The preceding sections lay out an automatic Bayesian cubature algorithm whose computational cost is drastically reduced. However, this algorithm relies on covariance kernel functions, $C_{\boldsymbol{\theta}}$ and node sets, $\{\boldsymbol{x}_i\}_{i=1}^n$ that satisfy assumptions (3.2). In this chapter, we demonstrate such a covariance kernel and matching design. When periodic shift-invariant kernels are combined with rank-1 lattice nodes, the resulting Gram matrix is symmetric and circulant. This combination also satisfies assumption (3.4). To conveniently facilitate the fast Bayesian transform, it is assumed in this section and the next that n is power of 2.

4.1 Extensible Integration Lattice Node Sets

We choose set of nodes defined by a shifted extensible integration lattice node sequence, which takes the form

$$\boldsymbol{x}_i = \boldsymbol{h}\phi(i-1) + \boldsymbol{\Delta} \mod \mathbf{1}, \qquad i \in \mathbb{N}. \tag{4.1}$$

Here, $\boldsymbol{h}$ is a d-dimensional generating vector of positive integers, $\boldsymbol{\Delta}$ is some point in $[0,1)^d$, often chosen at random, and $\{\phi(i)\}_{i=0}^n$ is the van der Corput sequence, defined by reflecting the binary digits of the integer about the decimal point, i.e.,

i	0	1	2	3	4	5	6	7	$\cdots$
i	0_2	1_2	10_2	11_2	100_2	101_2	110_2	111_2	$\cdots$
$\phi(i)$	$_2.0$	$_2.1$	$_2.01$	$_2.11$	$_2.001$	$_2.101$	$_2.011$	$_2.111$	$\cdots$
$\phi(i)$	0	0.5	0.25	0.75	0.125	0.625	0.375	0.875	$\cdots$

(4.2)

Note that

$$n\phi : \{0,\dots,n-1\} \to \{0,\dots,n-1\} \quad \text{is one-to-one,} \tag{4.3}$$

assuming n is a power of 2.

These node sets are called shifted rank-1 lattice node sets. A random shift $\boldsymbol{\Delta}$ is added to $\boldsymbol{h}\phi(i-1)$ to get $\{\boldsymbol{x}_i\}_{i=1}^n$ which is to avoid zero at the origin in the node sets. However, this shift does not disturb the discrepancy properties of $\{\boldsymbol{x}_i\}_{i=1}^n$. The rank-1 lattices with the modulo one addition have a very desirable group structure that helps to satisfy fast Bayesian transform kernel assumptions.

An example of 64 nodes is given in Figure 4.1. The even coverage of the unit cube is ensured by a well chosen generating vector $\boldsymbol{h}$. The choice of generating vector is typically done offline by computer search. Please refer to [24, 27] for more on extensible integration lattices. Lattice rules are designed to integrate the class of certain sinusoidal functions without error.

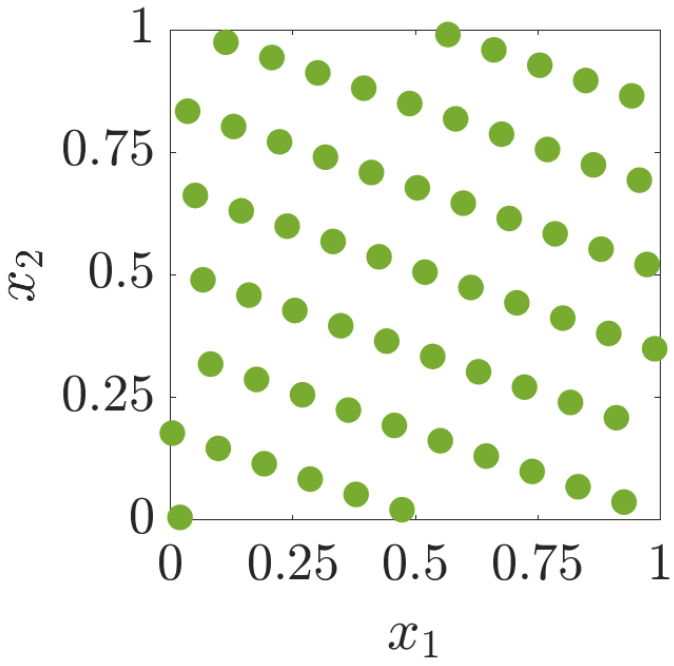

Figure 4.1. Example of a shifted integration lattice node set in $d = 2$

4.2 Shift Invariant Kernels

The covariance functions $C_{\boldsymbol{\theta}}$ that match integration lattice node sets have the form

$$C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) = K_{\boldsymbol{\theta}}(\boldsymbol{t} - \boldsymbol{x} \bmod \mathbf{1}). \tag{4.4}$$

This is called a *shift invariant kernel* because shifting both arguments of the covariance function by the same amount leaves the value unchanged. By a proper scaling of the function $K_{\boldsymbol{\theta}}$, the kernel satisfies the assumption (3.4). Here, $K_{\boldsymbol{\theta}}$ is chosen such that to ensure $C_{\boldsymbol{\theta}}$ is symmetric and positive definite, as assumed in (2.1).

A family of shift invariant kernels is constructed via even degree Bernoulli polynomials. Symmetric, periodic, positive definite kernels of this form appear in [24] and [28]:

$$C_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{t}) := \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \alpha_{\boldsymbol{k},\boldsymbol{\theta}} e^{2\pi\sqrt{-1}\boldsymbol{k}^T\boldsymbol{x}} e^{-2\pi\sqrt{-1}\boldsymbol{k}^T\boldsymbol{t}}, \quad \alpha_{-\boldsymbol{k},\boldsymbol{\theta}} = \alpha_{\boldsymbol{k},\boldsymbol{\theta}}$$

where d is the number of dimensions and $\alpha_{\boldsymbol{k}}$ is a positive scalar. The Gram matrix formed by this kernel is symmetric and positive definite. The *shape parameter* η_ℓ changes the kernel's shape, so that the integrand is in the middle of the function space spanned by the kernel. If the coefficients are chosen as

$$\alpha_{\boldsymbol{k},\boldsymbol{\theta}} := \prod_{\ell=1, k_\ell \neq 0}^{d} \frac{\eta_\ell}{|k_\ell|^r}, \quad \text{with } \alpha_{\mathbf{0},\boldsymbol{\theta}} = 1,\ r \in \mathbb{N},$$

then there exists a simpler closed form expression.

$$K_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{\ell=1}^{d} \Big[1 - (-1)^r \eta_\ell B_{2r}(x_\ell)\Big],$$

$$\forall \boldsymbol{x} \in [0,1]^d, \boldsymbol{\theta} := (r, \boldsymbol{\eta}),\ r \in \mathbb{N},\ \eta_\ell > 0. \tag{4.5}$$

Larger r implies a greater degree of smoothness of the kernel. Larger η_ℓ implies greater fluctuations of the output with respect to the input x_ℓ. The Bernoulli polynomials $B_r(x)$ are described in [29, Chapter 24]

$$B_r(x) = \frac{-r!}{(2\pi\sqrt{-1})^r} \sum_{\substack{k \neq 0, \\ k=-\infty}}^{\infty} \frac{e^{2\pi\sqrt{-1}kx}}{k^r} \quad \begin{cases} \text{for } \ r = 1, \ \ 0 < x < 1 \\ \text{for } \ r = 2, 3, \ldots \ \ 0 \leq x \leq 1 \end{cases}$$

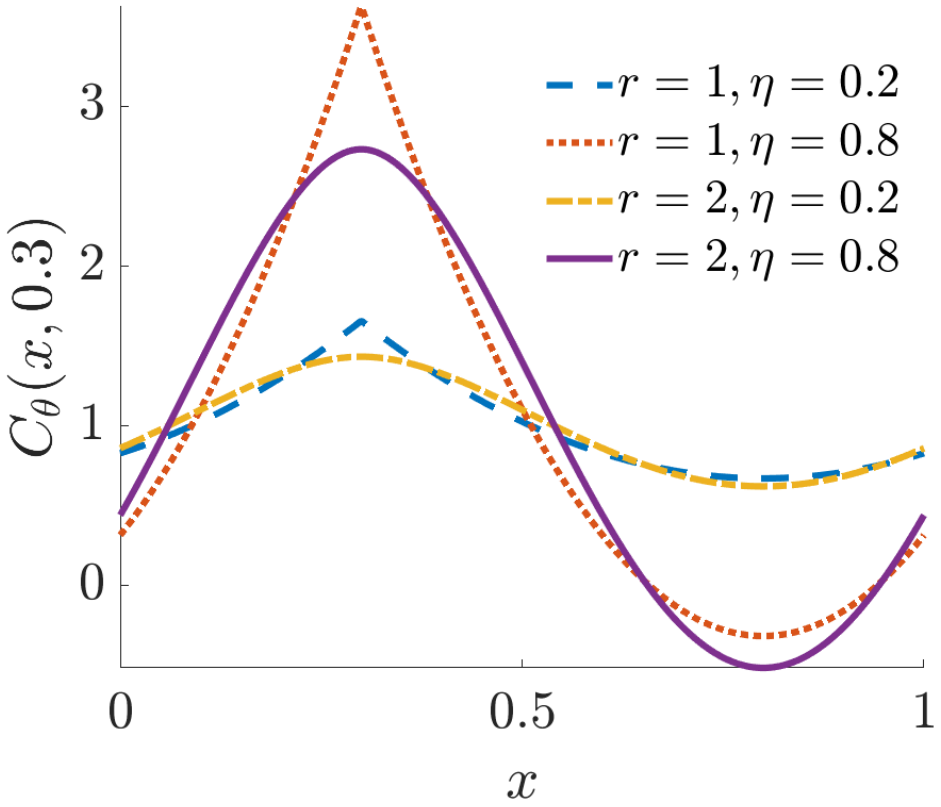

Figure 4.2. Shift invariant kernel in $d = 1$ shifted by 0.3 to show the discontinuity

Plots of $C(\cdot, 0.3)$ are given in Figure 4.2 for $d = 1$ and for various r and η_1 values.

Lattice cubature rules are known to have convergence rates that depend on the smoothness of the integrands, but that are rather independent of the choice of the integration lattice [24]. Thus, we expect integration lattice node sets to perform well regardless of the smoothness of the covariance kernel. The bigger concern is whether the derivatives of the integrand are as smooth as the covariance kernel implies. This topic is touched upon again in Section 4.5.

4.2.1 Eigenvectors. For general shift-invariance covariance functions the Gram matrix

$$\mathsf{C}_{\boldsymbol{\theta}} = \big(C_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\big)_{i,j=1}^{n} \tag{4.6}$$

can be shown that to have the eigenvector matrix

$$\mathsf{V} = \left(\mathrm{e}^{2\pi n\sqrt{-1}\phi(i-1)\phi(j-1)}\right)_{i=1}^{n}. \tag{4.7}$$

One can interpret the sequence reordering from $\{\phi(i-1)\}_{i=1}^{n}$ to $(0,\ldots,1-1/n)$, for n a power of 2, as a permutation. Let

$$\mathsf{P} = \left(\delta_{n\phi(i-1),j-1}\right)_{i,j=1}^{n} \tag{4.8}$$

be a permutation matrix, where $\delta_{\cdot,\cdot}$ is the Kronecker delta function. Then,

$$\begin{aligned}
\mathsf{C}_{\boldsymbol{\theta}} &= \left(C_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)_{i,j=1}^{n} \\
&= \Big(K_{\boldsymbol{\theta}}\big(\boldsymbol{h}(\phi(i-1)-\phi(j-1)) \bmod \mathbf{1}\big)\Big)_{i,j=1}^{n} \qquad \text{by (4.1) and (4.4)} \\
&= \left(\sum_{i',j'=1}^{n} \delta_{n\phi(i-1),i'-1}\, K_{\boldsymbol{\theta}}\big(\boldsymbol{h}(i'-j')/n \bmod \mathbf{1}\big)\delta_{j'-1,n\phi(j-1)}\right)_{i,j=1}^{n} \\
&= \mathsf{P}\mathsf{K}_{\boldsymbol{\theta}}\mathsf{P}^{T}, \quad \text{by (4.8)}
\end{aligned} \tag{4.9}$$

where

$$\mathsf{K}_{\boldsymbol{\theta}} = \left(K_{\boldsymbol{\theta}}\big(\boldsymbol{h}(i-j)/n \bmod \mathbf{1}\big)\right)_{i,j=1}^{n}. \tag{4.10}$$

Because $\mathsf{K}_{\boldsymbol{\theta}}$ is circulant, we know the form of it's eigenvector-eigenvalue decomposition:

$$\mathsf{K}_{\boldsymbol{\theta}} = \frac{1}{n}\mathsf{W}\Lambda_{\boldsymbol{\theta}}\mathsf{W}^{H}, \quad \text{where} \quad \mathsf{W} = \left(\mathrm{e}^{2\pi\sqrt{-1}(i-1)(j-1)/n}\right)_{i,j=1}^{n} \tag{4.11}$$

where $\Lambda_{\boldsymbol{\theta}}$ is a diagonal matrix. By (4.9) we then have the eigenvector-eigenvalue decomposition for $\mathsf{C}_{\boldsymbol{\theta}}$ assumed in (3.1), namely

$$\begin{aligned}
\mathsf{C}_{\boldsymbol{\theta}} &= \mathsf{P}\mathsf{K}_{\boldsymbol{\theta}}\mathsf{P}^{T} \\
&= \frac{1}{n}\mathsf{P}\mathsf{W}\Lambda_{\boldsymbol{\theta}}\mathsf{W}^{H}\mathsf{P}^{T} = \frac{1}{n}\underbrace{\mathsf{P}\mathsf{W}\mathsf{P}^{T}}\,\Lambda_{\boldsymbol{\theta}}\,\underbrace{\mathsf{P}\mathsf{W}^{H}\mathsf{P}^{T}} \\
&= \frac{1}{n}\mathsf{V}\Lambda_{\boldsymbol{\theta}}\mathsf{V}^{H}.
\end{aligned}$$

Thus

$$\mathsf{C}_{\boldsymbol{\theta}} = \frac{1}{n}\mathsf{V}\Lambda_{\boldsymbol{\theta}}\mathsf{V}^H, \qquad \mathsf{V} = \mathsf{P}\mathsf{W}\mathsf{P}^T, \tag{4.12}$$

where the eigenvalues of $\mathsf{C}_{\boldsymbol{\theta}}$ and $\mathsf{K}_{\boldsymbol{\theta}}$ are identical. Note that the matrix multiplication by V can be performed in $\mathcal{O}(n \log n)$ operations using the FFT.

4.3 Continuous Valued Kernel Order

In the previous sections, we assumed that the shift-invariant kernel's order is an even valued integer and also fixed. It requires the practitioner to be aware of the integrand's smoothness to precisely handpick the kernel order to match the integrand's smoothness. However, it is not possible to know the integrand's smoothness in most of the practical applications. The constraint to have an integer-valued kernel order also limits the ability to continuously vary the kernel's smoothness to match the integrand like the shape parameter is varied to match.

The integer kernel order is not suitable to optimally search by standard optimization algorithm. As a consequence, one usually ends up choosing a higher kernel order when the integrand is not smooth or lower kernel order when the integrand is very smooth. Often it leads to longer computation time or poor accuracy in the numerical integration. Here we explore two alternative forms of the kernel which allow the kernel order to be positive continuous value greater than one or a continuous value in the range $(0, 1)$. Let us recall the infinite series expression that was used to construct the kernel (4.5):

$$C_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{t}) := \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \alpha_{\boldsymbol{k},\boldsymbol{\theta}} e^{2\pi\sqrt{-1}\boldsymbol{k}^T\boldsymbol{x}} e^{-2\pi\sqrt{-1}\boldsymbol{k}^T\boldsymbol{t}}, \quad \text{where } \alpha_{\boldsymbol{k},\boldsymbol{\theta}} = \prod_{\ell=1}^{d} \frac{\eta_\ell}{|k_\ell|^r}$$

and $\boldsymbol{\theta} = (r, \boldsymbol{\eta})$. This form is convenient for analytical derivations. To make the derivations easier to follow, we fix the dimension $d = 1$,

$$C_{\boldsymbol{\theta}}(x, t) = 1 + \eta \sum_{k\in\mathbb{Z}, k\neq 0} \frac{1}{|k|^r} e^{2\pi\sqrt{-1}kx} e^{-2\pi\sqrt{-1}kt}.$$

4.3.1 Truncated series kernel. The following variation to the infinite series kernel (4.5) has the kernel order in the interval $(1, \infty)$. This kernel provides algebraic decay but it is more robust in the hyperparameter search. We reuse the original definition of the infinite kernel (4.5) but truncate to a finite length. This allows the kernel order r continuous valued so that it does not have to be an even integer, which was a constraint previously. For $d = 1$,

$$C_{\boldsymbol{\theta}}(x,t) = 1 + \eta \sum_{k\in\mathbb{Z},k\neq 0} \frac{1}{|k|^r} e^{2\pi\sqrt{-1}k(x-t)},$$

where $\theta = (r, \eta)$. Since the infinite sum cannot be used directly, we truncate to length n,

$$C_{\boldsymbol{\theta},n}(x,t) = 1 + \eta \sum_{k=-n/2}^{n/2-1} \frac{1}{|k|^r} e^{2\pi\sqrt{-1}k(x-t)}.$$

The Gram matrix is written as

$$\mathsf{C}_{\boldsymbol{\theta},n} = \Big(C_{\boldsymbol{\theta},n}(\boldsymbol{x}_i, \boldsymbol{x}_j) \Big)_{i,j=1}^{n},$$

where n is the number of samples. The reason for having the truncation length and the number of samples equal will be obvious as we proceed further. The first column of the Gram matrix is

$$\begin{aligned}\boldsymbol{C}_{\boldsymbol{\theta},n} &= \Big(C_{\boldsymbol{\theta},n}(\boldsymbol{x}_i, \boldsymbol{x}_1) \Big)_{i=1}^{n} \\ &= \left(\prod_{\ell=1}^{d} \left[1 + \eta_\ell \sum_{k=-n/2,k\neq 0}^{n/2-1} \frac{1}{|k_l|^r} e^{2\pi\sqrt{-1}k_l(x_{i\ell}-x_{1\ell})} \right] \right)_{i=1}^{n},\end{aligned}$$

where d is the number of dimensions. However the direct computation involves n^2 computations since we have chosen the truncation length to n. We can reduce the computations to $\mathcal{O}(n \log n)$ using the FFT. Define

$$\mathfrak{C}_r(t) := \sum_{k=-n/2,k\neq 0}^{n/2-1} \frac{1}{|k|^r} e^{2\pi\sqrt{-1}k\,t}.$$

Using the $\mathfrak{C}_r$, rewrite

$$\boldsymbol{C}_{\boldsymbol{\theta},n} = \left(\prod_{l=1}^{d} \left[1 + \eta \mathfrak{C}_r(|x_{il} - x_{1l}|) \right] \right)_{i=1}^{n}. \tag{4.13}$$

By the definition of lattice points from (4.1), we can observe $|x_{i\ell} - x_{1\ell}| \in \{0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{n-1}{n}\}$. This can be used to rewrite $\mathfrak{C}_r$ in a much simpler form,

$$\mathfrak{C}_r\left(\frac{j}{n}\right) = \sum_{k=-n/2, k\neq 0}^{n/2-1} \frac{1}{|k|^r} e^{2\pi\sqrt{-1}k(\frac{j}{n})}, \quad \text{where } j = 0, 1, \ldots n-1.$$

This notation is very convenient to show that $\widetilde{\mathfrak{C}}_r$, the discrete Fourier transform of $\mathfrak{C}_r$, can be computed analytically

$$\begin{aligned}
\widetilde{\mathfrak{C}}_r(m) &= \sum_{j=0}^{n-1} \mathfrak{C}_r(j/n) e^{-2\pi\sqrt{-1}jm/n} \\
&= \sum_{k=-n/2, k\neq 0}^{n/2-1} \sum_{j=0}^{n-1} \frac{1}{|k|^r} e^{2\pi\sqrt{-1}(k-m)j/n}, \quad \text{by (4.15)} \\
&= \sum_{k=-n/2, k\neq 0}^{n/2-1} \frac{n}{|k|^r}\ \delta_{k-m \bmod n, 0}\ .
\end{aligned}$$

This is the reason we have chosen the truncation length to n. Based on the above result, it is evident that $\widetilde{\mathfrak{C}}_r$ can be computed analytically,

$$\widetilde{\mathfrak{C}}_r(m) = \begin{cases} 0, & \text{for} \quad m = 0 \\ \frac{n}{|m|^r}, & \text{for} \quad m = 1, \ldots, n/2 - 1 \\ \frac{n}{|n-m|^r}, & \text{for} \quad m = n/2, \ldots, n-1 \end{cases} \tag{4.14}$$

where we used the fact,

$$\sum_{i=0}^{n-1} e^{2\pi\sqrt{-1}ij/n} = \begin{cases} \frac{1-e^{2\pi\sqrt{-1}jn/n}}{1-e^{2\pi\sqrt{-1}j/n}} = 0, & j \neq 0 \bmod n \\ n, & j = 0 \bmod n. \end{cases} \tag{4.15}$$

Having these results, we can easily back-compute $\mathfrak{C}$ using inverse discrete Fourier transform. It can be shown that inverse DFT of $\widetilde{\mathfrak{C}}_r$ returns $\mathfrak{C}$,

$$\frac{1}{n} \sum_{m=0}^{n-1} \widetilde{\mathfrak{C}}_r(m) e^{2\pi\sqrt{-1}lm/n}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{m=0}^{n-1}\sum_{j=0}^{n-1}\mathfrak{C}_r(j/n)e^{-2\pi\sqrt{-1}jm/n}e^{2\pi\sqrt{-1}lm/n}, \quad \text{by (4.15)}\\
&= \frac{1}{n}\sum_{j=0}^{n-1}\mathfrak{C}_r(j/n)n\delta_{(l-j) \bmod n,0}\\
&= \mathfrak{C}_r(l/n), \quad \text{for} \quad l = 0,\ldots,n-1
\end{aligned}
$$

This implies that to compute n values of $\left(C_{\boldsymbol{\theta},n}(\boldsymbol{x}_i,\boldsymbol{x}_1)\right)_{i=1}^{n}$, we need to have the number of samples and the truncation length the same. The above results are summarized as an algorithm to compute $\mathfrak{C}$ using FFT in Algorithm 2.

Algorithm 2 The kernel with continuous valued order

Require: Number of points to use, n;

1: Analytically compute $\left(\widetilde{\mathfrak{C}}_r(m)\right)_{m=0}^{n-1}$, the discrete Fourier coefficients of $(\mathfrak{C}_r(j/n))_{j=0}^{n-1}$ using (4.14)

2: Take the inverse FFT of $\left(\widetilde{\mathfrak{C}}_r(m)\right)_{m=0}^{n-1}$ to get $(\mathfrak{C}_r(j/n))_{j=0}^{n-1}$

In Algorithm 2, the computational cost of computing $(\mathfrak{C}_r(j/n))_{j=0}^{n-1}$ is $\mathcal{O}(n\log n)$ instead of $\mathcal{O}(n^2)$. Plugging-in the values of $\mathfrak{C}$ in (4.13) gives the kernel. Another major benefit is that the FFT approach is numerically more stable than the direct sum approach. Please note that these kernels evolve with the truncation length n. The larger n value the closer the kernel resembles the original infinite series kernel. One disadvantage is, the truncated series kernels obtain algebraic order decay at best. The infinite series kernel with little modification can be enhanced to obtain exponential decay as shown next.

4.3.2 Exponentially decaying kernel. We propose the following alternative form of the kernel. This kernel can provide exponential decay,

$$
C_{\boldsymbol{\theta}}(x,t) = 1 + \eta\sum_{k\in\mathbb{Z},k\neq 0} q^{|k|}e^{2\pi\sqrt{-1}k(x-t)}, \quad \text{with} \quad 0<q<1
$$

where q is used to denote the kernel order to distinguish it from the notation in (4.13). This can be rewritten as

$$\begin{aligned}
C_{\boldsymbol{\theta}}(x,t) =& 1+\eta \sum_{k\in\mathbb{Z},k\neq 0} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)} \\
=& 1+\eta\left(\sum_{k=1}^{\infty} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)} + \sum_{k=\infty}^{-1} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)}\right) \\
=& 1+\eta\left(\sum_{k=1}^{\infty} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)} + \sum_{k=-\infty}^{-1} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)}\right) \\
=& 1+\eta\left(\underbrace{\sum_{k=1}^{\infty} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)}}_{*} + \sum_{k=1}^{\infty} e^{-2\pi\sqrt{-1}k(x-t)+|k|\log(q)}\right).
\end{aligned}$$

Let us focus on the first term $(*)$ within the parenthesis in the previous equation,

$$\begin{aligned}
\sum_{k=1}^{\infty} e^{2\pi\sqrt{-1}k(x-t)+|k|\log(q)} &= \sum_{k=1}^{\infty}\left[e^{2\pi\sqrt{-1}(x-t)+\log(q)}\right]^k \\
&= \frac{e^{2\pi\sqrt{-1}(x-t)+\log(q)}}{1-e^{2\pi\sqrt{-1}(x-t)+\log(q)}} = \frac{1}{e^{-2\pi\sqrt{-1}(x-t)-\log(q)}-1} \\
&= \frac{1}{q^{-1}e^{-2\pi\sqrt{-1}(x-t)}-1}
\end{aligned}$$

Using this result

$$\begin{aligned}
C_{\boldsymbol{\theta}}(x,t) &= 1+\eta\left(\frac{1}{q^{-1}e^{-2\pi\sqrt{-1}(x-t)}-1} + \frac{1}{q^{-1}e^{2\pi\sqrt{-1}(x-t)}-1}\right) \\
&= 1+\eta\left(\frac{q^{-1}\left(e^{2\pi\sqrt{-1}(x-t)}+e^{-2\pi\sqrt{-1}(x-t)}\right)-2}{q^{-2}-q^{-1}\left(e^{2\pi\sqrt{-1}(x-t)}+e^{-2\pi\sqrt{-1}(x-t)}\right)+1}\right) \\
&= 1+\eta\left(\frac{2q^{-1}\cos(2\pi\sqrt{-1}(x-t))-2}{q^{-2}-2q^{-1}\cos(2\pi\sqrt{-1}(x-t))+1}\right) \\
&= 1+2\eta q\left(\frac{\cos(2\pi\sqrt{-1}(x-t))-q}{q^2-2q\cos(2\pi\sqrt{-1}(x-t))+1}\right).
\end{aligned}$$

Using the fact $\cos^2(t)+\sin^2(t)=1$,

$$C_{\boldsymbol{\theta}}(x,t) = 1+2\eta q\left(\frac{\cos(2\pi\sqrt{-1}(x-t))-q}{\left[\cos(2\pi\sqrt{-1}(x-t))-q\right]^2+\sin^2(2\pi\sqrt{-1}(x-t))}\right),$$

which shows that the kernel order q can be continuously varied while searching for the optimal value. The hyperparameters need to be $\eta > 0$ and $0 < q < 1$ while searching for the optimum value, so we use the transformations demonstrated in Section 6.2 to map the values to or from $\mathbb{R}$, where the search is usually done. One disadvantage of this kernel is that it is very sensitive to the changes in kernel order $q \in (0, 1)$, for even small values, which might cause the hyperparameter search to miss the global minima.

4.4 Summary

We summarize the results of this and the previous chapter as a theorem below.

Theorem 4.4.1. *Let $\mathsf{C}_{\boldsymbol{\theta}}$ be any symmetric, positive definite, shift-invariant covariance kernel of the form* (4.4)*, where $K_{\boldsymbol{\theta}}$ has period one in every variable. Furthermore, let $K_{\boldsymbol{\theta}}$ be scaled to satisfy* (3.4)*. When matched with rank-1 lattice data-sites, $\mathsf{C}_{\boldsymbol{\theta}}$ must satisfy assumptions* (3.2)*. The cubature, $\widehat{\mu}$, is just the sample mean. The* fast Fourier transform *(FFT) can be used to expedite the estimates of $\boldsymbol{\theta}$ in* (6.1) *and the credible interval widths* (6.2) *in $\mathcal{O}(n \log n)$ operations.*

Although the third part of the computational cost has the largest dependence on n, in practice it need not be the largest contributor to the computational cost. If function values are the result of an expensive simulation, then the first part may consume most of the computation time.

We have implemented the fast adaptive Bayesian cubature algorithm in MATLAB as part of the Guaranteed Adaptive Integration Library (GAIL) [30] as `cubBayesLattice_g`. This algorithm uses the kernel defined in (4.5) with $r = 1, 2$ or the continuous valued order kernel (4.13), and the periodizing variable transforms in Section 4.5. The rank-1 lattice node generator is taken from [31] (`exod2_base2_m20`).

4.5 Periodizing Variable Transformations

The shift-invariant covariance kernels underlying our `cubBayesLattice_g` Bayesian cubature assume that the integrand has a degree of periodicity, with the smoothness assumed depending on the smoothness of the kernel. In other-words, non-periodic functions do not live in the space spanned by the shift-invariant covariance kernels. While integrands arising in practice may be smooth, they might not be periodic. Variable transformation or periodization transform techniques are typically used to enforce the periodicity in multi-dimensional numerical integrations where boundary conditions needs to be enforced. These transformations could be either polynomial, exponential and also trigonometric in nature. Some of the most popular transformation are provided here for reference.

Suppose that the original integral has been expressed as

$$\mu := \int_{[0,1]^d} g(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t},$$

where g has sufficient smoothness, but lacks periodicity. The goal is to transform the integral above to the form of (1.1), where the integrand f—and perhaps its derivatives—are periodic.

The Baker's transform, also called tent transform,

$$\boldsymbol{\Psi} : \boldsymbol{x} \mapsto (\Psi(x_1), \ldots, \Psi(x_d)), \quad \Psi(x) = 1 - 2\,|x - 1/2|\,, \tag{4.16}$$

allows us to write μ in the form of (1.1), where $f(\boldsymbol{x}) = g(\boldsymbol{\Psi}(\boldsymbol{x}))$. Since $\Psi'(x)$ is not continuous, f does not have continuous derivatives.

A family of smoother variable transforms that can also preserve continuity of derivatives from the original integrand g takes the form

$$\boldsymbol{\Psi} : \boldsymbol{x} \mapsto (\Psi(x_1), \ldots, \Psi(x_d)), \quad \Psi : [0,1] \mapsto [0,1].$$

This allows us to write μ in the form of (1.1) with

$$f(\boldsymbol{x}) = g(\boldsymbol{\Psi}(\boldsymbol{x})) \prod_{\ell=1}^{d} \Psi'(x_l).$$

For $r \in \mathbb{N}_0$, if the following hold:

- $\Psi \in C^{r+1}[0,1]$,
- $\lim_{x\downarrow 0} x^{-r-1}\Psi'(x) = \lim_{x\uparrow 1}(1-x)^{-r-1}\Psi'(x) = 0$, and
- $g \in C^{(r,\dots,r)}[0,1]^d$,

then f has continuous, periodic mixed partial derivatives of up to order r in each direction. Examples of this kind of transform include [32]:

$$
\begin{aligned}
C^0 : \Psi(x) &= 3x^2 - 2x^3, \quad \Psi'(x) = 6x(1-x),\\
C^1 : \Psi(x) &= x^3(10 - 15x + 6x^2),\\
\Psi'(x) &= 30x^2(1-x)^2\\
\text{Sidi's } C^1 : \Psi(x) &= x - \frac{\sin(2\pi x)}{2\pi},\\
\Psi'(x) &= 1 - \cos(2\pi x),\\
\text{Sidi's } C^2 : \Psi(x) &= \frac{8 - 9\cos(\pi x) + \cos(3\pi x)}{16},\\
\Psi'(x) &= \frac{3\pi[3\sin(\pi x) - \sin(3\pi x)]}{16}.
\end{aligned}
$$

These transforms vary in terms of computational complexity and accuracy and shall be chosen to match the covariance kernel and integrand accordingly. Choosing an optimal periodizing is a topic of future research. Baker's transform is the least complex of all which is a tent map in each coordinate. It preserves only continuity but it is easier to compute and it does not include product term up to the length dimension of the integrand, making it more numerically stable. C^0 is a polynomial transformation only and ensures periodicity of function. C^1 is a polynomial transformation and preserving the first derivative. Sidi's C^1, a transform which uses trigonometric Sine, preserves the first derivative and is, in general, a better option than C^1. Sidi's C^2, also a transform which uses trigonometric Sine, preserves up to second derivative.

We use this when smoothness of Sidi's C^1 is not sufficient and need to preserve up to second derivative.

Periodizing variable transforms are used in the numerical examples in Section 7. In some cases, they can speed the convergence of the Bayesian cubature because they allow one to take advantage of smoother covariance kernels. However, there is a trade-off. Smoother periodizing transformations tend to give integrands f with larger inferred s values and thus wider credible intervals.

CHAPTER 5

SOBOL' NETS AND WALSH KERNELS

The previous section shows an automatic Bayesian cubature algorithm using rank-1 lattice nodes and shift-invariant kernels. In this chapter, we demonstrate a second approach to formulate fast Bayesian transform using matching kernel and point sets. Scrambled Sobol' nets and Walsh kernels are paired to achieve $\mathcal{O}(n^{-1+\epsilon})$ order error convergence where n is the sample size. Sobol' nets [33] are low discrepancy points, used extensively in numerical integration, simulation, and optimization. The results of this chapter can be summarized as a theorem,

Theorem 5.0.1. *Any symmetric, positive definite, digital shift-invariant covariance kernel of the form* (5.5) *scaled to satisfy* (3.4)*, when matched with digital net data-sites, satisfies assumptions* (3.2)*. The* fast Walsh-Hadamard transform *(FWHT) can be used to expedite the estimates of* $\boldsymbol{\theta}$ *in* (6.1) *and the credible interval widths* (6.2) *in* $\mathcal{O}(n \log n)$ *operations. The cubature,* $\widehat{\mu}$*, is just the sample mean.*

We introduce the necessary concepts and prove this theorem in the remaining of this chapter.

5.1 Sobol' Nets

Nets were developed to provide deterministic sample points for quasi-Monte Carlo rules [34]. Nets are defined geometrically using elementary intervals, which are subintervals of the unit cube $[0,1)^d$. The (t,m,d)-nets in base b, introduced by Niederreiter, whose quality is governed by t. Lower values of t correspond to (t,m,d)-nets of higher quality [35].

Definition 1. *Let* $\mathcal{A}$ *be the set of all elementary intervals* $\mathcal{A} \subset [0,1)^d$ *where* $\mathcal{A} = \prod_{\ell=1}^{d} [\alpha_\ell b^{-\gamma_\ell}, (\alpha_\ell + 1) b^{-\gamma_\ell})$*, with* $d, b, \gamma_\ell \in \mathbb{N}, b \geq 2$ *and* $b^{\gamma_\ell} > \alpha_\ell \geq 0$*. For* $m, t \in$

$\mathbb{N}, m \geq t \geq 0$, *the point set* $\mathcal{P}_m \in [0,1)^d$ *with* $n = b^m$ *points is a* (t, m, d) *– net in base* b *if every* $\mathcal{A}$ *with volume* b^{t-m} *contains* b^t *points of* $\mathcal{P}_m$.

Digital (t, m, d)-nets are a special case of (t, m, d)-nets, constructed using matrix-vector multiplications over finite fields. Digital sequences are infinite length digital nets, i.e., the first $n = b^m$ points of a digital sequence comprise a digital net for all integer $m \in \mathbb{N}_0$.

Definition 2. *For any non-negative integer* $i = \ldots i_3 i_2 i_1 (\text{base } b)$, *define the* $\infty \times 1$ *vector* $\vec{i}$ *as the vector of its digits, that is,* $\vec{i} = (i_1, i_2, \ldots)^T$. *For any point* $z = 0.z_1 z_2 \ldots (\text{base } b) \in [0,1)$, *define the* $\infty \times 1$ *vector of the digits of* z, *that is,* $\vec{z} = (z_1, z_2, \ldots)^T$. *Let* $\mathsf{G}_1, \ldots, \mathsf{G}_d$ *denote predetermined* $\infty \times \infty$ *generator matrices. The digital sequence in* base b *is* $\{\boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots\}$, *where each* $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{id})^T \in [0,1)^d$ *is defined by*

$$\vec{z}_{i\ell} = \mathsf{G}_\ell \, \vec{i}, \quad \ell = 1, \ldots, d, \quad i = 0, 1, \ldots \, .$$

The value of t *as mentioned in Definition 1 depends on the choice of* G_ℓ.

Digital nets have a group structure under digitwise addition, which is a very useful property exploited in our algorithm, especially to develop a fast Bayesian transform that speedups computations. Digitwise addition, $\oplus$, and subtraction $\ominus$, are defined in terms of b-ary expansions of points in $[0,1)^d$,

$$\boldsymbol{z} \oplus \boldsymbol{y} = \left(\sum_{j=1}^{\infty} [z_{\ell j} + y_{\ell j} \bmod b] b^{-j} \bmod 1 \right)_{\ell=1}^{d},$$
$$\boldsymbol{z} \ominus \boldsymbol{y} = \left(\sum_{j=1}^{\infty} [z_{\ell j} - y_{\ell j} \bmod b] b^{-j} \bmod 1 \right)_{\ell=1}^{d},$$

where

$$\boldsymbol{z} = \left(\sum_{j=1}^{\infty} z_{\ell j} b^{-j} \right)_{\ell=1}^{d}, \quad \boldsymbol{y} = \left(\sum_{j=1}^{\infty} y_{\ell j} b^{-j} \right)_{\ell=1}^{d}, \quad z_{\ell j}, y_{\ell j} \in \{0, \cdots, b-1\}.$$

Similarly for integer values in $\mathbb{N}_0^d$, the digitwise addition, $\oplus$, and subtraction $\ominus$, are defined in terms of their b-ary expansions,

$$\boldsymbol{k} \oplus \boldsymbol{l} = \left(\sum_{j=0}^{\infty} [k_{\ell j} + l_{\ell j} \bmod b] b^j \bmod 1 \right)_{\ell=1}^{d},$$
$$\boldsymbol{k} \ominus \boldsymbol{l} = \left(\sum_{j=0}^{\infty} [k_{\ell j} - l_{\ell j} \bmod b] b^j \bmod 1 \right)_{\ell=1}^{d},$$

where

$$\boldsymbol{k} = \left(\sum_{j=0}^{\infty} k_{\ell j} b^j \right)_{\ell=1}^{d}, \quad \boldsymbol{l} = \left(\sum_{j=0}^{\infty} l_{\ell j} b^j \right)_{\ell=1}^{d}, \quad \boldsymbol{k}_{\ell j}, \boldsymbol{l}_{\ell j} \in \{0, \cdots, b-1\}.$$

Let $\{\boldsymbol{z}_i\}_{i=0}^{b^m-1}$ be a digital net. Then

$$\forall i_1, i_2 \in \{0, \cdots, b^m - 1\}, \quad \boldsymbol{z}_{j_1} \oplus \boldsymbol{z}_{i_2} = \boldsymbol{z}_{i_3}, \quad \text{for some } i_3 \in \{0, \cdots, b^m - 1\}.$$

The following very useful result, which will be further used to obtain the fast Bayesian transform, arises from the fundamental property of digital nets.

Lemma 5.1.1. *Let $\{\boldsymbol{z}_i\}_{i=0}^{b^m-1}$ be the digital-net and the corresponding digitally shifted net be $\{\boldsymbol{x}_i\}_{i=0}^{b^m-1}$, i.e.,*

$$\vec{x}_{i\ell} = \vec{z}_{i\ell} + \vec{\Delta}_l \bmod 1,$$

where $\vec{x}_{i\ell}$ is the ℓth component of ith digital net and $\vec{\Delta}_\ell$ is the digital shift for the ℓth component. Then,

$$\boldsymbol{x}_i \ominus \boldsymbol{x}_j = \boldsymbol{z}_i \ominus \boldsymbol{z}_j = \boldsymbol{z}_{i \ominus j}, \quad \forall i, j \in \mathbb{N}_0. \tag{5.1}$$

Also the digital subtraction is symmetric,

$$\boldsymbol{x}_i \ominus \boldsymbol{x}_i = \boldsymbol{0}, \qquad \boldsymbol{x}_i \ominus \boldsymbol{x}_j = \boldsymbol{x}_j \ominus \boldsymbol{x}_i, \quad \forall i, j \in \mathbb{N}_0. \tag{5.2}$$

Proof. The proof can be obtained from the definition of digital nets which stated that the digital nets are obtained using generator matrices, $\vec{z}_{i\ell} = \mathsf{G}_\ell \, \vec{i} \bmod b$. Rewriting the subtraction using the generating matrix provides the result,

$$\begin{aligned}
\vec{z}_{i\ell} - \vec{z}_{j\ell} \bmod b &= (\mathsf{G}_\ell \vec{i} \bmod b) - (\mathsf{G}_\ell \vec{j} \bmod b) \\
&= (\mathsf{G}_\ell \vec{i} - \mathsf{G}_\ell \vec{j}) \bmod b \\
&= \mathsf{G}_\ell (\vec{i} - \vec{j}) \bmod b \\
&= \mathsf{G}_\ell (\overrightarrow{i \ominus j}) \bmod b \\
&= \vec{z}_{i \ominus j \; \ell}.
\end{aligned}$$

The rest of the lemma is obvious from the definition of digital nets. □

We chose digitally shifted and scrambled nets [36] for our Bayesian cubature algorithm. Digital shifts help to avoid having nodes at the origin, similar to the random shift used with lattice nodes. Scrambling helps to eliminate bias while retaining the low-discrepancy properties. A proof that a scrambled net preserves the property of (t, m, d)-net almost surely can be found in Owen [37]. The scrambling method proposed by Matoušek [38] is preferred since it is more efficient than the Owen's scrambling.

Sobol' nets [39] are a special case of (t, m, d)-nets when base $b = 2$. An example of 64 Sobol' nets in $d = 2$ is given in Figure 5.1. The even coverage of the unit cube is ensured by a well chosen generating matrix. The choice of generating vector is typically done offline by computer search. See [40] and [41] for more on generating matrices. We use randomly scrambled and digitally shifted Sobol' sequences in this research [42].

5.2 Walsh Kernels

Walsh kernels are product kernels based on the Walsh functions. We introduce

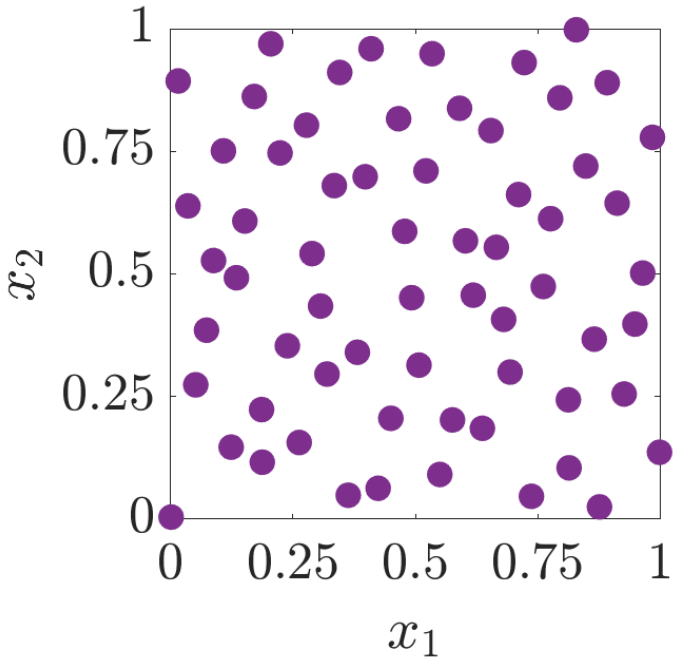

Figure 5.1. Example of a scrambled Sobol' node set in $d = 2$

the necessary concepts in this section.

5.2.1 Walsh functions. Like the Fourier transform used with lattice points (Section 4.2), the Walsh-Hadamard transform, which we will simply call Walsh transform, is used for the digital nets. The Walsh transform is defined using Walsh functions. Recall $\mathbb{N}_0 := \{0, 1, 2, \cdots\}$. The one-dimensional Walsh functions in base b are defined as

$$\text{wal}_{b,k}(x) := e^{2\pi\sqrt{-1}(x_1k_0+x_2k_1+\cdots)/b} = e^{2\pi\sqrt{-1}\vec{k}^T\vec{x}/b}, \tag{5.3}$$

for $x \in [0, 1)$ and $k \in \mathbb{N}_0$ and the unique base b expansions $x = \sum_{j\geq 1} x_j b^{-i} = (0.x_1x_2\cdots)_b$, $\vec{x} = (x_1, x_2, \cdots)^T$ $k = \sum_{j\geq 0} k_j b^j = (\cdots k_1k_0)_b$, $\vec{k} = (k_0, k_1, \cdots)^T$, and $\vec{k}^T\vec{x} = x_1k_0 + x_2k_1 + \cdots$ where the number of digits used in (5.3) are limited to the length required to represent x or k, i.e., $\max(\lceil -\log_b x\rceil, \lceil \log_b k\rceil)$. Multivariate Walsh

functions are defined as the product of the one-dimensional Walsh functions,

$$\mathrm{wal}_{b,\boldsymbol{k}}(\boldsymbol{x}) := \prod_{\ell=1}^{d} \mathrm{wal}_{b,k_\ell}(x_\ell)$$

As shown in (5.3), for the case of $b = 2$, the Walsh functions only take the values in $\{1, -1\}$, i.e., $\mathrm{wal}_{b,\boldsymbol{k}} : [0,1)^d \to \{-1, 1\},\ k \in \mathbb{N}_0^d$. Walsh functions form an orthonormal basis of the Hilbert space $L^2[0,1)^d$,

$$\int_{[0,1)^d} \mathrm{wal}_{b,\boldsymbol{l}}(\boldsymbol{x})\mathrm{wal}_{b,\boldsymbol{k}}(\boldsymbol{x})\mathrm{d}x = \delta_{\boldsymbol{l},\boldsymbol{k}}, \quad \forall \boldsymbol{l}, \boldsymbol{k} \in \mathbb{N}_0^d$$

Digital nets are designed to integrate certain Walsh functions without error. Thus our Bayesian cubature algorithm integrates linear combinations of certain Walsh functions without error. Functions that are well approximated by such linear combinations are then integrated with small errors.

In this research we use Sobol' nodes which are digital nets with base $b = 2$. So here afterwards base $b = 2$ is assumed. In this case, the Walsh function is simply

$$\mathrm{wal}_{2,\boldsymbol{k}}(\boldsymbol{x}) = (-1)^{\vec{\boldsymbol{k}}^T \vec{\boldsymbol{x}}}.$$

5.2.2 Walsh kernels. Consider the covariance kernels of the form,

$$C_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{t}) = K_{\boldsymbol{\theta}}(\boldsymbol{x} \ominus \boldsymbol{t}) \tag{5.4}$$

where $\ominus$ is bitwise subtraction. This is called a *digitally shift invariant kernel* because shifting both arguments of the covariance function by the same amount leaves the value unchanged. By a proper scaling of the function $K_{\boldsymbol{\theta}}$, it follows that assumption (3.4) is satisfied. The function $K_{\boldsymbol{\theta}}$ must be of the form that ensures that $C_{\boldsymbol{\theta}}$ is symmetric and positive definite, as assumed in (2.1). We drop the $\boldsymbol{\theta}$ sometimes to make the notation simpler. The Walsh kernels are of the form,

$$K_{\boldsymbol{\theta}}(\boldsymbol{x} \ominus \boldsymbol{t}) = \prod_{\ell=1}^{d} 1 + \eta_\ell \omega_r(x_\ell \ominus t_\ell), \quad \boldsymbol{\eta} = (\eta_1, \cdots, \eta_d), \quad \boldsymbol{\theta} = (r, \boldsymbol{\eta}) \tag{5.5}$$

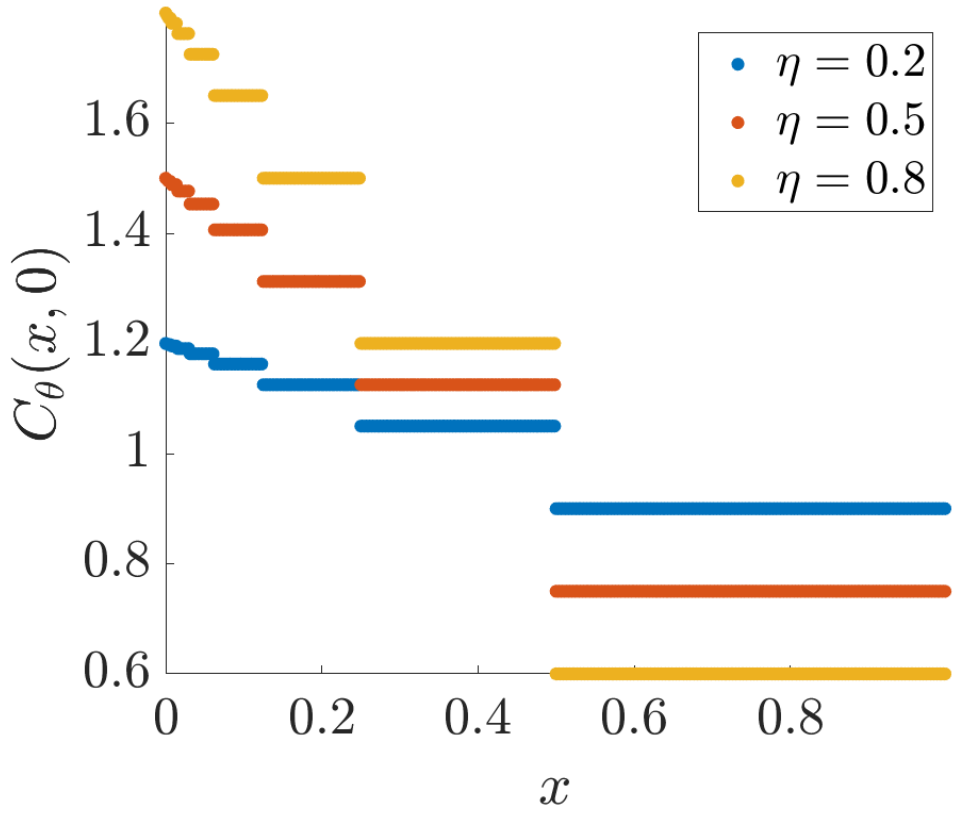

Figure 5.2. Walsh kernel of order $r = 1$ in dimension $d = 1$. This figure can be reproduced using `plot_walsh_kernel.m`.

where r is the kernel order, $\boldsymbol{\eta}$ is the kernel shape parameter, and

$$\omega_r(x) = \sum_{k=1}^{\infty} \frac{\mathrm{wal}_{2,k}(x)}{2^{2r\lfloor \log_2 k \rfloor}}.$$

Explicit expression is available for ω_r in the case of order $r = 1$ [43],

$$\omega_1(x) = 6\left(\frac{1}{6} - 2^{\lfloor \log_2 x \rfloor - 1}\right). \tag{5.6}$$

The Figure 5.2 shows the Walsh kernel (5.5) of order $r = 1$ in the interval $[0, 1)$. Unlike the shift-invariant kernels used with lattice nodes, low order Walsh kernels are discontinuous and are only piecewise constant. Smaller η_ℓ implies lesser variation in the amplitude of the kernel. Also, the Walsh kernels are digitally shift invariant but not periodic.

5.3 Eigenvectors

We show the eigenvectors V in (3.1) of the Gram matrix formed by the covariance kernel (5.5) and Sobol' nets are the columns of the Walsh-Hadamard matrix. First we introduce the necessary concepts.

5.3.1 Walsh transform. The Walsh-Hadamard transform (WHT) is a generalized class of discrete Fourier transform (DFT) and is much simpler to compute than the DFT. The WHT matrices are comprised of only ± 1 values, so the computation usually involves only ordinary additions and subtractions. Hence, the WHT is also sometimes called the integer transform. In comparison, the DFT that was used with lattice nodes, uses complex exponential functions and the computation involves complex, non-integer multiplications.

The WHT involves multiplications by $2^m \times 2^m$ Walsh-Hadamard matrices, which is constructed recursively, starting with $\mathsf{H}^{(0)} = 1$,

$$\mathsf{H}^{(1)} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\mathsf{H}^{(2)} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

$$\vdots$$

$$\mathsf{H}^{(m)} = \begin{pmatrix} \mathsf{H}^{(m-1)} & \mathsf{H}^{(m-1)} \\ \mathsf{H}^{(m-1)} & -\mathsf{H}^{(m-1)} \end{pmatrix} = \underbrace{\mathsf{H}^{(1)} \bigotimes \cdots \bigotimes \mathsf{H}^{(1)}}_{m \text{ times}} = \mathsf{H}^{(1)} \bigotimes \mathsf{H}^{(m-1)} \tag{5.7}$$

where $\bigotimes$ is Kronecker product. Alternatively for base $b = 2$, these matrices can be

directly obtained by,

$$\mathsf{H}^{(m)} = \left((-1)^{(\vec{i}^T\vec{j})}\right)_{i,j=0}^{2^m-1},$$

where the notation $\vec{i}^T\vec{j}$ indicates the bitwise dot product.

5.3.2 Eigenvectors of C are columns of Walsh-Hadamard matrix. The Gram matrix $\mathsf{C}_{\boldsymbol{\theta}}$ formed by Walsh kernels and Sobol' nodes have a special structure called block-Toeplitz, which can be used to construct the fast Bayesian transform. A Toeplitz matrix is a diagonal-constant matrix in which each descending diagonal from left to right is constant. A block Toeplitz matrix is a special block matrix, which contains blocks that are repeated down the diagonals of the matrix. We prove that the eigenvectors of $\mathsf{C}_{\boldsymbol{\theta}}$ are columns of a Walsh-Hadamard matrix in two theorems.

Theorem 5.3.1. *Let $(\boldsymbol{x}_i)_{i=0}^{n-1}$ be digitally shifted Sobol' nodes and K be any function, then the Gram matrix,*

$$\mathsf{C}_{\boldsymbol{\theta}} = \big(C(\boldsymbol{x}_i, \boldsymbol{x}_j)\big)_{i,j=0}^{n-1} = \big(K(\boldsymbol{x}_i \ominus \boldsymbol{x}_j)\big)_{i,j=0}^{n-1},$$
$$\text{where} \qquad n = 2^m, \quad C(\boldsymbol{x}, \boldsymbol{t}) = K(\boldsymbol{x} \ominus \boldsymbol{t}), \quad \boldsymbol{x}, \boldsymbol{t} \in [0,1)^d,$$

is a 2×2 block-Toeplitz matrix and all the sub-blocks and their sub-sub-blocks, etc. are also 2×2 block-Toeplitz.

Proof. We prove this theorem by induction. Let $\mathsf{C}_{\boldsymbol{\theta}}^{(m)}$ denote the Gram matrix of size $2^m \times 2^m$. The relation between sub-block matrices can be deciphered using the properties of digital nets. To help with the proof of block-Toeplitz structure, consider the digital net properties (5.1), (5.2), and notations,

$$\mathsf{K}^{(m)} := \left(K(\boldsymbol{z}_i \ominus \boldsymbol{z}_j)\right)_{i,j=0}^{2^m-1} = \left(K(\boldsymbol{z}_{i\ominus j})\right)_{i,j=0}^{2^m-1}, \quad m = 1, 2, \cdots,$$
$$\mathsf{K}^{(m,q)} := \left(K(\boldsymbol{z}_{i\ominus j+q2^m})\right)_{i,j=0}^{2^m-1}, \quad q = 0, 1, \cdots.$$

These two notations are related by $\mathsf{K}^{(m)} = \mathsf{K}^{(m,0)}$. Please note that $\mathsf{C}_{\boldsymbol{\theta}}^{(m)} = \mathsf{K}^{(m,0)}$. We will prove $\mathsf{K}^{(m,q)}$ is a 2×2 block-toeplitz matrix for all $m \in \mathbb{N}, q \in \mathbb{N}$.

As the first step, we verify the property holds for $m = 1$,

$$\mathsf{K}^{(1,q)} = \begin{pmatrix} K(\boldsymbol{z}_{0\ominus 0+q2^1}) & K(\boldsymbol{z}_{1\ominus 0+q2^1}) \\ K(\boldsymbol{z}_{0\ominus 1+q2^1}) & K(\boldsymbol{z}_{1\ominus 1+q2^1}) \end{pmatrix} = \begin{pmatrix} K(\boldsymbol{z}_{2q}) & K(\boldsymbol{z}_{1+2q}) \\ K(\boldsymbol{z}_{1+2q}) & K(\boldsymbol{z}_{2q}) \end{pmatrix}, \quad \text{by (5.1)}$$

has diagonal elements repeated. Thus by definition, it is a 2×2 block-Toeplitz.

Now assume that $\mathsf{K}^{(m,q)}$ is block-Toeplitz. We need to prove $\mathsf{K}^{(m+1,q)}$ is also a 2×2 block-Toeplitz. Let $n = 2^m$,

$$\mathsf{K}^{(m+1)} = \begin{pmatrix} K(\boldsymbol{z}_{0\ominus 0}) & \dots & K(\boldsymbol{z}_{0\ominus n-1}) & K(\boldsymbol{z}_{0\ominus n}) & \dots & K(\boldsymbol{z}_{0\ominus 2n-1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ K(\boldsymbol{z}_{n-1\ominus 0}) & \dots & K(\boldsymbol{z}_{n-1\ominus n-1}) & K(\boldsymbol{z}_{n-1\ominus n}) & \dots & K(\boldsymbol{z}_{n-1\ominus 2n-1}) \\ K(\boldsymbol{z}_{n\ominus 0}) & \dots & K(\boldsymbol{z}_{n\ominus n-1}) & K(\boldsymbol{z}_{n\ominus n}) & \dots & K(\boldsymbol{z}_{n\ominus 2n-1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ K(\boldsymbol{z}_{2n-1\ominus 0}) & \dots & K(\boldsymbol{z}_{2n-1\ominus n-1}) & K(\boldsymbol{z}_{2n-1\ominus n}) & \dots & K(\boldsymbol{z}_{2n-1\ominus 2n-1}) \end{pmatrix}$$

$$= \begin{pmatrix} \begin{pmatrix} K(\boldsymbol{z}_0) & \dots & K(\boldsymbol{z}_{n-1}) \\ \vdots & \vdots & \vdots \\ K(\boldsymbol{z}_{n-1}) & \dots & K(\boldsymbol{z}_0) \end{pmatrix} & \begin{pmatrix} K(\boldsymbol{z}_n) & \dots & K(\boldsymbol{z}_{2n-1}) \\ \vdots & \vdots & \vdots \\ K(\boldsymbol{z}_{2n-1}) & \dots & K(\boldsymbol{z}_n) \end{pmatrix} \\ \begin{pmatrix} K(\boldsymbol{z}_n) & \dots & K(\boldsymbol{z}_{2n-1}) \\ \vdots & \vdots & \vdots \\ K(\boldsymbol{z}_{2n-1}) & \dots & K(\boldsymbol{z}_n) \end{pmatrix} & \begin{pmatrix} K(\boldsymbol{z}_0) & \dots & K(\boldsymbol{z}_{n-1}) \\ \vdots & \vdots & \vdots \\ K(\boldsymbol{z}_{n-1}) & \dots & K(\boldsymbol{z}_0) \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} \mathsf{K}^{(m)} & \mathsf{K}^{(m,1)} \\ \\ \mathsf{K}^{(m,1)} & \mathsf{K}^{(m)} \end{pmatrix}$$

is a 2×2 block-Toeplitz, where we used the properties (5.1), (5.2) and facts $2n-1\ominus n = n-1$, $2n-1\ominus n-1 = n$, and $n\ominus n-1 = 2n-1$. Thus $\mathsf{K}^{(m+1)}$ is a 2×2 block-Toeplitz. Similarly

$$\mathsf{K}^{(m+1,q)} = \begin{pmatrix} \mathsf{K}^{(m,q)} & \mathsf{K}^{(m,q+1)} \\ \\ \mathsf{K}^{(m,q+1)} & \mathsf{K}^{(m,q)} \end{pmatrix}$$

is a 2×2 block-Toeplitz. Thus $\mathsf{C}_{\boldsymbol{\theta}}^{(m)}$ of size $2^m \times 2^m$, for $m \in \mathbb{N}$, is a 2×2 block-Toeplitz and every block and it's sub-blocks of size 2^p, $p \in \mathbb{N}$, $p \leq m$ are also 2×2 block-Toeplitz. □

Theorem 5.3.2. *The Walsh-Hadamard matrix* $\mathsf{H}^{(m)}$ *factorizes* $\mathsf{C}_{\boldsymbol{\theta}}^{(m)}$ *so that the columns of Walsh-Hadamard matrix are the eigenvectors of* $\mathsf{C}_{\boldsymbol{\theta}}^{(m)}$*, i.e.,*

$$\mathsf{H}^{(m)}\mathsf{C}_{\boldsymbol{\theta}}^{(m)} = \Lambda^{(m)}\mathsf{H}^{(m)}, \quad m \in \mathbb{N}.$$

Proof. Again we use the proof-by-induction technique to show that the Walsh-Hadamard matrix factorizes $\mathsf{K}^{(m,q)}$. We can easily see the Hadamard matrix $\mathsf{H}^{(1)}$ diagonalizes $\mathsf{K}^{(1,q)}$,

$$\begin{aligned} \mathsf{H}^{(1)}\mathsf{K}^{(1,q)} &= \begin{pmatrix} 1 & 1 \\ \\ 1 & -1 \end{pmatrix} \begin{pmatrix} K(\boldsymbol{z}_{0+q2^1}) & K(\boldsymbol{z}_{1+q2^1}) \\ \\ K(\boldsymbol{z}_{1+q2^1}) & K(\boldsymbol{z}_{0+q2^1}) \end{pmatrix}, \quad \text{by Theorem 5.3.1} \\ &= \begin{pmatrix} K(\boldsymbol{z}_{2q}) + K(\boldsymbol{z}_{2q+1}) & K(\boldsymbol{z}_{2q}) + K(\boldsymbol{z}_{2q+1}) \\ \\ K(\boldsymbol{z}_{2q}) - K(\boldsymbol{z}_{2q+1}) & K(\boldsymbol{z}_{2q+1}) - K(\boldsymbol{z}_{2q}) \end{pmatrix} \\ &= \begin{pmatrix} K(\boldsymbol{z}_{2q}) + K(\boldsymbol{z}_{2q+1}) & 0 \\ \\ 0 & K(\boldsymbol{z}_{2q}) - K(\boldsymbol{z}_{2q+1}) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \\ 1 & -1 \end{pmatrix} \end{aligned}$$

$$= \Lambda^{(1,q)}\mathsf{H}^{(1)},$$

where $\Lambda^{(1,q)}$ is a diagonal matrix, thus $\mathsf{H}^{(1)}$ factorizes $\mathsf{K}^{(1,q)}$.

Now assume $\mathsf{H}^{(m)}$ factorizes $\mathsf{K}^{(m,q)}$, so $\mathsf{H}^{(m)}\mathsf{K}^{(m,q)} = \Lambda^{(m,q)}\mathsf{H}^{(m)}$ where $\Lambda^{(m,q)}$ is diagonal. We need to prove $\mathsf{H}^{(m+1)}$ factorizes $\mathsf{K}^{(m+1,q)}$,

$$\begin{aligned}
\mathsf{H}^{(m+1)}\mathsf{K}^{(m+1,q)} &= \begin{pmatrix} \mathsf{H}^{(m)} & \mathsf{H}^{(m)} \\ \mathsf{H}^{(m)} & -\mathsf{H}^{(m)} \end{pmatrix} \begin{pmatrix} \mathsf{K}^{(m,q)} & \mathsf{K}^{(m,q+1)} \\ \mathsf{K}^{(m,q+1)} & \mathsf{K}^{(m,q)} \end{pmatrix}, \quad \text{by Theorem 5.3.1} \\
&= \begin{pmatrix} \mathsf{H}^{(m)}(\mathsf{K}^{(m,q)} + \mathsf{K}^{(m,q+1)}) & \mathsf{H}^{(m)}(\mathsf{K}^{(m,q)} + \mathsf{K}^{(m,q+1)}) \\ \mathsf{H}^{(m)}(\mathsf{K}^{(m,q)} - \mathsf{K}^{(m,q+1)}) & \mathsf{H}^{(m)}(\mathsf{K}^{(m,q+1)} - \mathsf{K}^{(m,q)}) \end{pmatrix} \\
&= \begin{pmatrix} (\Lambda^{(m,q)} + \Lambda^{(m,q+1)})\mathsf{H}^{(m)} & (\Lambda^{(m,q)} + \Lambda^{(m,q+1)})\mathsf{H}^{(m)} \\ (\Lambda^{(m,q)} - \Lambda^{(m,q+1)})\mathsf{H}^{(m)} & (\Lambda^{(m,q+1)} - \Lambda^{(m,q)})\mathsf{H}^{(m)} \end{pmatrix} \\
&= \begin{pmatrix} \Lambda^{(m,q)} + \Lambda^{(m,q+1)} & 0 \\ 0 & \Lambda^{(m,q)} - \Lambda^{(m,q+1)} \end{pmatrix} \begin{pmatrix} \mathsf{H}^{(m)} & \mathsf{H}^{(m)} \\ \mathsf{H}^{(m)} & -\mathsf{H}^{(m)} \end{pmatrix} \\
&= \Lambda^{(m+1,q)}\mathsf{H}^{(m+1)}.
\end{aligned}$$

Thus, $\mathsf{H}^{(m+1)}$ factorizes $\mathsf{K}^{(m+1,q)}$ to a diagonal matrix $\Lambda^{(m+1,q)}$. This implies $\mathsf{H}^{(p)}$ factorizes $\mathsf{C}_{\boldsymbol{\theta}}^{(p)}$ for $p \in \mathbb{N}$. Please recall $\mathsf{C}_{\boldsymbol{\theta}}^{(p)} = \mathsf{K}^{(p,0)}$. Here we used the fact that both H and K are symmetric positive definite. □

5.3.3 Fast Bayesian transform. We can easily show that the Walsh-Hadamard matrices satisfy the assumptions of fast Bayesian transform (3.2). As shown in Section 5.3.2 the columns of $\mathsf{H}^{(m)}$ are the eigenvectors. Since the Gram matrix C is symmetric, the columns/rows of Walsh-Hadamard matrices are mutually orthogonal.

Thus the Gram matrix can be written as

$$\mathsf{C}^{(m)} = \frac{1}{n}\mathsf{H}^{(m)}\Lambda^{(m)}\mathsf{H}^{(m)}, \quad \text{where} \quad \mathsf{H}^{(m)} = \underbrace{\mathsf{H}^{(1)} \bigotimes \cdots \bigotimes \mathsf{H}^{(1)}}_{m \text{ times}}. \tag{5.8}$$

Assumption (3.2b) follows automatically by the fact that Walsh-Hadamard matrices can be constructed analytically. Assumption (3.2a) can also be verified as the first row/column are one vectors. Finally, assumption (3.2c) is satisfied due to the fact that fast Walsh transform can be computed in $\mathcal{O}(n \log n)$ operations using fast Walsh-Hadamard transform. Thus the Walsh-Hadamard transform is a fast Bayesian transform, $\mathsf{V} := \mathsf{H}$, as per (3.2).

We have implemented a fast adaptive Bayesian cubature algorithm using the kernel (5.5) with $r = 1$ and Sobol' points [44] in MATLAB as part of the Guaranteed Adaptive Integration Library (GAIL) [30] as `cubBayesNet_g`. The Sobol' points used in this algorithm are generated using MATLAB's builtin function `sobolset` and scrambled using MATLAB function `scramble` [42]. The fast Walsh-Hadamard transform (5.8) is computed using MATLAB's builtin function `fwht` with *hadamard* ordering.

5.3.4 Iterative Computation of Walsh Transform. In every iteration of our algorithm, we double the number of function values. Using the technique described here, we have to only compute the Walsh transform for the newly added function values. Similar to the lattice points, Sobol' points are extensible by definition. This property is used in our algorithm to improve the integration accuracy till the required error tolerance is met. Sobol' nodes can be combined with Hadamard matrices as demonstrated here for iterative computation. Let $\widetilde{\boldsymbol{y}} = \mathsf{H}^{(m+1)}\boldsymbol{y}$ for some arbitrary

$\boldsymbol{y} \in \mathbb{R}^{2n}$, $n = 2^m$. Define,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{2n} \end{pmatrix}, \quad \boldsymbol{y}^{(1)} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{y}^{(2)} = \begin{pmatrix} y_{n+1} \\ \vdots \\ y_{2n} \end{pmatrix},$$

$$\widetilde{\boldsymbol{y}}^{(1)} = \mathsf{H}^{(m)}\boldsymbol{y}^{(1)} = \begin{pmatrix} \widetilde{y}_1^{(1)} \\ \widetilde{y}_2^{(1)} \\ \vdots \\ \widetilde{y}_n^{(1)} \end{pmatrix}, \quad \widetilde{\boldsymbol{y}}^{(2)} = \mathsf{H}^{(m)}\boldsymbol{y}^{(2)} = \begin{pmatrix} \widetilde{y}_1^{(2)} \\ \widetilde{y}_2^{(2)} \\ \vdots \\ \widetilde{y}_n^{(2)} \end{pmatrix}.$$

Then,

$$\begin{aligned} \widetilde{\boldsymbol{y}} &= \mathsf{H}^{(m+1)}\boldsymbol{y} \\ &= \begin{pmatrix} \mathsf{H}^{(m)} & \mathsf{H}^{(m)} \\ \mathsf{H}^{(m)} & -\mathsf{H}^{(m)} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}^{(1)} \\ \boldsymbol{y}^{(2)} \end{pmatrix}, \qquad \text{by (5.7)} \\ &= \begin{pmatrix} \mathsf{H}^{(m)}\boldsymbol{y}^{(1)} + \mathsf{H}^{(m)}\boldsymbol{y}^{(2)} \\ \mathsf{H}^{(m)}\boldsymbol{y}^{(1)} - \mathsf{H}^{(m)}\boldsymbol{y}^{(2)} \end{pmatrix} \\ &= \begin{pmatrix} \widetilde{\boldsymbol{y}}^{(1)} + \widetilde{\boldsymbol{y}}^{(2)} \\ \widetilde{\boldsymbol{y}}^{(1)} - \widetilde{\boldsymbol{y}}^{(2)} \end{pmatrix} =: \widetilde{\boldsymbol{y}} \quad . \end{aligned}$$

As before with the lattice nodes, the computational cost to compute $\mathsf{V}^{(m+1)H}\boldsymbol{y}$ is twice the cost of computing $\mathsf{V}^{(m)H}\boldsymbol{y}^{(1)}$ plus $2n$ additions, where $n = 2^m$. An inductive argument shows that for any $m \in \mathbb{N}$, $\mathsf{V}^{(m)H}\boldsymbol{y}$ requires only $\mathcal{O}(n \log n)$ operations. Usually the multiplications in $\mathsf{V}^{(m)H}\boldsymbol{y}^{(1)}$ are multiplications by -1 which are simply accomplished using sign change or negation, requiring no multiplications at all.

5.4 Higher Order Nets Higher order digital nets are an extension of (t, m, d)-nets, introduced in [45]. They can be used to numerically integrate smoother functions

which are not necessarily periodic, but have square integrable mixed partial derivatives of order α, at a rate of $\mathcal{O}(n^{-\alpha})$ multiplied by a power of a $\log n$ factor using rules corresponding to the modified (t, m, d)-nets. We want to emphasize that quasi-Monte Carlo rules based on these point sets can achieve convergence rates faster than $\mathcal{O}(n^{-1})$. Higher order digital nets are constructed using matrix-vector multiplications over finite fields.

One could develop matching digitally shift invariant kernels to formulate the fast Bayesian cubature. Bayesian cubatures using higher order digital nets are a topic for future research.

CHAPTER 6

NUMERICAL IMPLEMENTATION

6.1 Overcoming Cancellation Error

We now refer back to general setting for the fast automatic Bayesian cubature in Section 3. For the covariance kernels used in our computation, it often happens that n/λ_1 is close to 1, especially for larger n. Thus, the term $1 - n/\lambda_1$, which appears in the credible interval widths, $\mathrm{err}_{\mathrm{EB}}$, $\mathrm{err}_{\mathrm{full}}$, and $\mathrm{err}_{\mathrm{GCV}}$ (3.7), may suffer from cancellation error. We can avoid this cancellation error by modifying how we compute the Gram matrix and its eigenvalues.

Any shift-invariant or digital shift-invariant covariance kernel satisfying (3.4) can be written as $C_{\boldsymbol{\theta}} = 1 + \mathring{C}_{\boldsymbol{\theta}}$, where $\mathring{C}_{\boldsymbol{\theta}}$ is also symmetric and positive definite. The associated Gram matrix for $\mathring{C}_{\boldsymbol{\theta}}$ is then $\mathring{\mathsf{C}}_{\boldsymbol{\theta}} = \mathsf{C}_{\boldsymbol{\theta}} - \mathbf{1}\mathbf{1}^T$, and the eigenvalues of $\mathring{\mathsf{C}}_{\boldsymbol{\theta}}$ are $\mathring{\lambda}_1 = \lambda_1 - n, \lambda_2, \ldots, \lambda_n$, which follows because $\mathbf{1}$ is the first eigenvector of both $\mathsf{C}_{\boldsymbol{\theta}}$ and $\mathring{\mathsf{C}}_{\boldsymbol{\theta}}$. Note that $\mathring{C}_{\boldsymbol{\theta}}$ inherits the shift-invariant properties of $C_{\boldsymbol{\theta}}$. Then,

$$1 - \frac{n}{\lambda_1} = \frac{\lambda_1 - n}{\lambda_1} = \frac{\mathring{\lambda}_1}{\mathring{\lambda}_1 + n},$$

where now the right hand side is free of cancellation error.

We show how to compute $\mathring{C}_{\boldsymbol{\theta}}$ without introducing round-off error. The covariance functions that we use in both Chapter 4 and 5 are of product form, namely,

$$C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) = \prod_{\ell=1}^{d} \left[1 + \mathring{C}_{\boldsymbol{\theta},\ell}(t_\ell, x_\ell)\right], \qquad \mathring{C}_{\boldsymbol{\theta},\ell} : [0,1] \times [0,1] \to \mathbb{R}.$$

Direct computation of $\mathring{C}_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) = C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) - 1$ introduces cancellation error if the $\mathring{C}_\ell$ are small. So, we employ the iteration,

$$\begin{aligned}
\mathring{C}^{(1)}_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) &= \mathring{C}_{\boldsymbol{\theta},1}(t_1, x_1), \\
\mathring{C}^{(\ell)}_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x}) &= \mathring{C}^{(\ell-1)}_{\boldsymbol{\theta}}[1 + \mathring{C}_{\boldsymbol{\theta},\ell}(t_\ell, x_\ell)] + \mathring{C}_{\boldsymbol{\theta},\ell}(t_\ell, x_\ell), \qquad \ell = 2, \ldots, d,
\end{aligned}$$

$$\mathring{C}_{\boldsymbol{\theta}}(\boldsymbol{t},\boldsymbol{x}) = \mathring{C}_{\boldsymbol{\theta}}^{(d)}(\boldsymbol{t},\boldsymbol{x}).$$

In this way, the Gram matrix $\mathring{\mathsf{C}}_{\boldsymbol{\theta}}$, whose i,j-element is $\mathring{C}_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ can be constructed with minimal round-off error because we avoid subtraction.

Computing the eigenvalues of $\mathring{\mathsf{C}}_{\boldsymbol{\theta}}$ via the procedure given in (3.3) yields $\mathring{\lambda}_1 = \lambda_1 - n, \lambda_2, \ldots, \lambda_n$. The estimates of $\boldsymbol{\theta}$ are computed in terms of the eigenvalues of $\mathring{\mathsf{C}}_{\boldsymbol{\theta}}$. So (3.5a) and (3.5b) become

$$\boldsymbol{\theta}_{\mathrm{EB}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\log\left(\sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i} \right) + \frac{1}{n} \sum_{i=1}^{n} \log(\lambda_i) \right], \tag{6.1a}$$

$$\boldsymbol{\theta}_{\mathrm{GCV}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\log\left(\sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i^2} \right) - 2\log\left(\sum_{i=1}^{n} \frac{1}{\lambda_i} \right) \right], \tag{6.1b}$$

where $\lambda_1 = n + \mathring{\lambda}_1$. The widths of the credible intervals in (3.7a), (3.7b), and (3.7c) become,

$$\mathrm{err}_{\mathrm{EB}} = \frac{2.58}{n} \sqrt{ \frac{\mathring{\lambda}_1}{\lambda_1} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i} }, \tag{6.2a}$$

$$\mathrm{err}_{\mathrm{full}} = \frac{t_{n_j-1,0.995}}{n} \sqrt{ \frac{\mathring{\lambda}_1}{n-1} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i} }, \tag{6.2b}$$

$$\mathrm{err}_{\mathrm{GCV}} = \frac{2.58}{n} \sqrt{ \frac{\mathring{\lambda}_1}{\lambda_1} \sum_{i=2}^{n} \frac{|\widetilde{y}_i|^2}{\lambda_i^2} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i} \right]^{-1} }. \tag{6.2c}$$

Since $\mathring{\lambda}_1 = \lambda_1 - n$ and $\lambda_1 \sim n$ it follows $\mathring{\lambda}_1/\lambda_1 \approx \mathring{\lambda}_1/(n-1)$ and is small for large n. Moreover, for large n, the credible intervals via empirical Bayes and full Bayes are similar, since $t_{n_j-1,0.995}$ is approximately 2.58.

The computational steps for the improved, faster, automatic Bayesian cubature are detailed in Algorithm 3. In comparison to Algorithm 1, the second and third components of the computational cost of Algorithm 3 are substantially reduced. The Algorithm 3 has a computational cost which is the sum of the following:

Algorithm 3 Fast Automatic Bayesian Cubature

Require: A choice of generator for the point set $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ and a matching kernel $C_{\boldsymbol{\theta}}$ from, 1) rank-1 Lattice points and a matching shift-invariant kernel, 2) Sobol' sequence and a matching digital shift-invariant kernel; a black-box function, f; an absolute error tolerance, $\varepsilon > 0$; the positive initial sample size, n_0, that is a power of 2; the maximum sample size $n_{\max}$

$n \leftarrow n_0,\ n' \leftarrow 0,\ \mathrm{err}_{\mathrm{CI}} \leftarrow \infty$

while $\mathrm{err}_{\mathrm{CI}} > \varepsilon$ and $n \leq n_{\max}$ **do**

    Generate $\{\boldsymbol{x}_i\}_{i=n'+1}^{n}$ and sample $\{f(\boldsymbol{x}_i)\}_{i=n'+1}^{n}$

    Compute $\boldsymbol{\theta}$ by (3.5a) or (3.5b) by using the techniques from Chapter 4 or 5

    Compute $\mathrm{err}_{\mathrm{CI}}$ according to (6.2a), (6.2b), or (6.2c) by using the techniques from Chapter 4 or 5

    $n' \leftarrow n,\ n \leftarrow 2n'$

end while

Update sample size to compute $\widehat{\mu}$, $n \leftarrow n'$

Compute $\widehat{\mu}$, the approximate integral, according to (3.6)

return $\widehat{\mu}$, n and $\mathrm{err}_{\mathrm{CI}}$

- $\mathcal{O}\big(n\$(f)\big)$ for the integrand data, where $\$(f)$ is the computational cost of a single $f(\boldsymbol{x})$
- $\mathcal{O}\big(N_{\text{opt}} n\$(C_{\boldsymbol{\theta}})\big)$ for the evaluations of the vector $\boldsymbol{C}_1$, where N_{opt} is the number of optimization steps required, and $\$(C_{\boldsymbol{\theta}})$ is the computational cost of a single $C_{\boldsymbol{\theta}}(\boldsymbol{t}, \boldsymbol{x})$
- $\mathcal{O}\big(N_{\text{opt}} n \log(n)\big)$ for the FFT calculations; there is no d dependence in these calculations

6.2 Kernel Hyperparameters Search

The various hyperparameters introduced and used by our algorithms need to be optimally chosen. The parameter search can be done in two major ways. Bounded minima search, if the search interval is known, else unbounded search. Most of the scenarios, the search interval is unknown. So the natural choice is to use unbounded search over the unbound domain such as `fminsearch` provided by MATLAB. However hyperparameters need to live in a domain that is bounded or semi-bounded. There are some simple domain transformations available to achieve this.

6.2.1 Positive kernel shape parameter. The following parameter map is used to ensure that the shape parameter values are positive real numbers. For $\eta > 0$ as introduced in Section 3.5.1, let

$$\eta(t_1) = \quad e^{t_1}, \qquad\qquad \eta : (-\infty, \infty) \to (0, \infty).$$

Instead of searching for $\eta \in (0, \infty)$, we may search for the optimal $t_1 = \log(\eta)$ over the whole real line $\mathbb{R}$. The optimal value $t_{1,\text{opt}}$ can be transformed back to the $(0, \infty)$ interval using

$$\eta_{\text{opt}} = e^{t_{1,\text{opt}}}.$$

6.2.2 Kernel order $1 < r < \infty$. The following map is used to ensure that the kernel order values are positive real number and greater than one, i.e., in the $(1, \infty)$ interval as required in Section 4.3.1,

$$r(t_2) = \quad 1 + e^{t_2}, \qquad\qquad r : (-\infty, \infty) \to (1, \infty).$$

So one may search for the optimal $t_2 = -\log(r-1)$ in the whole real line $\mathbb{R}$. The optimal value $t_{2,\text{opt}}$ can be transformed back to the desired interval $(0, 1)$ using

$$r_{\text{opt}} = 1 + e^{t_{2,\text{opt}}}.$$

6.2.3 Kernel order $0 < q < 1$. The following multivariate map is used to ensure that the kernel order values are positive real and less than one, i.e., in the $(0, 1)$ interval to use with exponentially decaying kernel, as introduced in Section 4.3.2,

$$q(t_3) = \quad \frac{1}{1 + e^{t_3}}, \qquad\qquad q : (-\infty, \infty) \to (0, 1).$$

So one may search for the optimal $t_3 = \log(q^{-1} - 1)$ in the whole real line $\mathbb{R}$. The optimal value $t_{3,\text{opt}}$ can be transformed back to the desired interval $(0, 1)$ by using

$$q_{\text{opt}} = \frac{1}{1 + e^{t_{3,\text{opt}}}}.$$

6.2.4 Combined searching of kernel order r and shape parameter $\boldsymbol{\eta}$. Instead of searching $\boldsymbol{\eta}$ and r separately one would prefer to search them together so that the most optimal values can be obtained, where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_d)$ such that $\eta_l \neq \eta_k$, for $l \neq k$. We can combine the parameter maps used above to ensure that the kernel order values in $(1, \infty)$ and shape parameter $\boldsymbol{\eta}$ in $(0, \infty)^d$ as required in Section 4.3.1,

$$\boldsymbol{\theta}(\boldsymbol{t}) = \begin{pmatrix} r(t_1) \\ \eta(t_2) \\ \vdots \\ \eta(t_{d+1}) \end{pmatrix} = \begin{pmatrix} 1 + e^{t_1} \\ e^{t_2} \\ \vdots \\ e^{t_{d+1}} \end{pmatrix}, \quad \boldsymbol{\theta} : \mathbb{R}^{d+1} \to (1, \infty) \times (0, \infty)^d.$$

So instead of searching for $\boldsymbol{\theta}_{\text{opt}}$ in $(1,\infty)\times(0,\infty)^d$, one may search for the optimal

$$\boldsymbol{t}=\boldsymbol{t}^{-1}(\boldsymbol{\theta})=\begin{pmatrix}\log(r-1)\\ \log(\eta_1)\\ \vdots\\ \log(\eta_d)\end{pmatrix}$$

in the whole real line $\mathbb{R}^{d+1}$. The optimal value $\boldsymbol{t}_{\text{opt}}$ can be transformed back to the desired interval $(1,\infty)\times(0,\infty)^d$ using

$$\boldsymbol{\theta}_{\text{opt}}=\begin{pmatrix}1+e^{t_{1,\text{opt}}}\\ e^{t_{2,\text{opt}}}\\ \vdots\\ e^{t_{d+1,\text{opt}}}\end{pmatrix}.$$

Similarly one can map the kernel order $q\in(0,1)$ Section 4.3.2, and $\boldsymbol{\eta}$ in to a multivariate hyperparameter search.

CHAPTER 7

NUMERICAL RESULTS AND OBSERVATIONS

Fast Bayesian cubature algorithms developed in this research are demonstrated using three commonly used integration examples. These integrals were evaluated using both the algorithms `cubBayesLattice_g` and `cubBayesNet_g`. The first example shows evaluating a multivariate Gaussian probability given the interval. The second example shows integrating the Keister's function, and the final example shows computing an Asian arithmetic option pricing.

7.1 Testing Methodology

Four hundred different error tolerances, ε, were randomly chosen from a fixed interval for each example. The intervals for error tolerance were chosen depending on the difficulty of the problem. The nodes used in `cubBayesLattice_g` were the randomly shifted lattice points supplied by GAIL, whereas the nodes used in `cubBayesNet_g` were the randomly scrambled and shifted Sobol' points supplied by MATLAB's Sobol' sequence generator.

For each integral example, and each stopping criteria—empirical Bayes, full Bayes, and generalized cross-validation—our algorithm is run with each randomly chosen error tolerance as mentioned above. For each test, the execution time is plotted against $|\mu - \widehat{\mu}|/\varepsilon$. We expect $|\mu - \widehat{\mu}|/\varepsilon$ to be no greater than one, but hope that it is not too much smaller than one, which would indicate a stopping criterion that is too conservative.

Periodization variable transforms are used in the examples with `cubBayesLattice_g`, which assumes the integrands to be periodic in $[0,1]^d$. But the `cubBayesNet_g` does not need this additional requirement, so the integrands are used directly.

7.2 Multivariate Gaussian Probability

This example is introduced in Section 2.8, where we use the Matérn covariance kernel. We reuse f_{Genz} (2.26) and apply a periodization transform to obtain f_{GenzP} when required.

7.2.1 Using `cubBayesLattice_g`. As required by the algorithm, we apply Sidi's C^2 periodization to f_{Genz} (2.26), and chose $d = 3$ and $r = 2$. The simulation results for this example integrand are summarized in Figures 7.1, 7.2, and 7.3. In all cases, `cubBayesLattice_g` returns an approximation within the prescribed error tolerance. We used the same setting as before with generic slow Bayesian cubature in Section 2.8 for comparision. For error threshold $\varepsilon = 10^{-5}$ with empirical stopping criterion, our fast algorithm takes 0.001 seconds as shown in Figure 7.1 whereas the basic algorithm takes 30 seconds as shown in Figure 2.4. Amongst the three stopping criteria, GCV achieved the results faster than others but it is less conservative. One can also observe from the figures that the credible intervals are wider, causing true error much smaller than requested. This could be due to the periodization transformed integrand, f_{GenzP}, being smoother than the $r = 2$ kernel approximation. Using a kernel of matching smoothness could produce right credible intervals.

7.2.2 Using `cubBayesNet_g`. Here we use f_{Genz} (2.26) without any periodization, and chose $d = 3$ and $r = 1$. The simulation results for this example integrand are summarized in Figures 7.4, 7.5, and 7.6. In all cases, `cubBayesNet_g` returns an approximation within the prescribed error tolerance. We used the same setting as before with generic slow Bayesian cubature in Section 2.8 for comparision. For error threshold $\varepsilon = 10^{-5}$ with empirical stopping criterion, our fast algorithm takes about 2 seconds as shown in Figure 7.1 whereas the basic algorithm takes 30 seconds as shown in Figure 2.4. `cubBayesNet_g` uses fast Walsh transform which is slower in MATLAB due to the way it was implemented. This is reason it takes more longer the

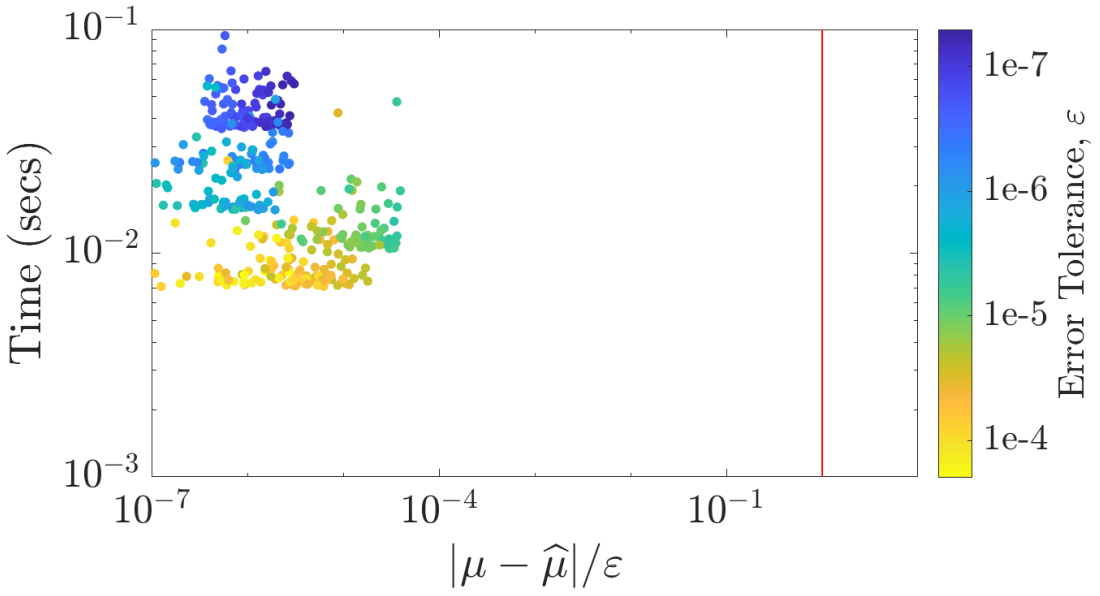

Figure 7.1. `cubBayesLattice_g`: Multivariate normal probability example using the empirical Bayes stopping criterion.

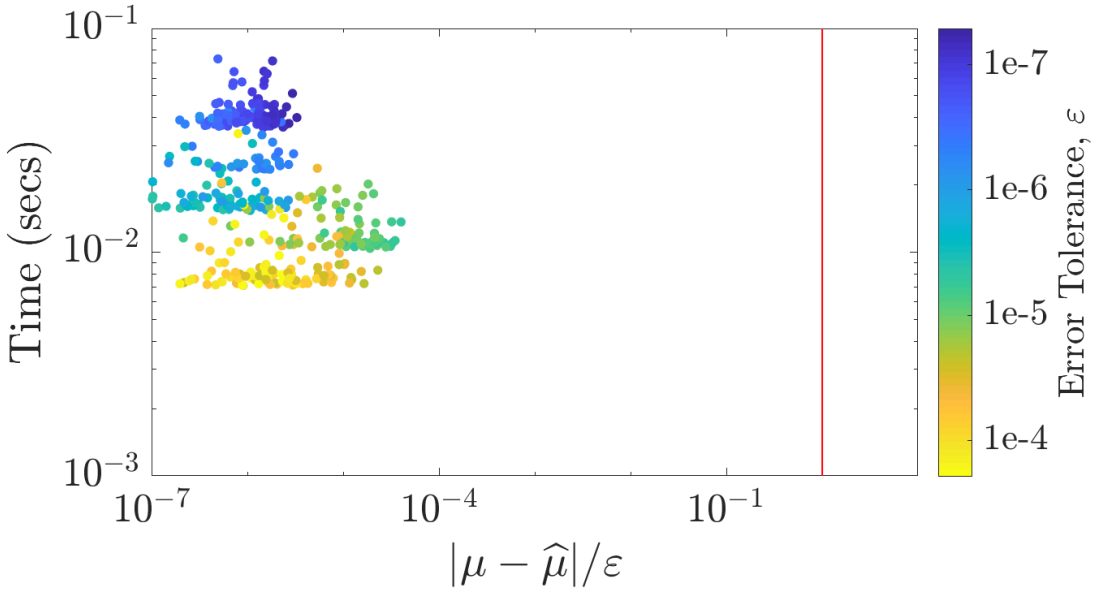

Figure 7.2. `cubBayesLattice_g`: Multivariate normal probability example using the full Bayes stopping criterion.

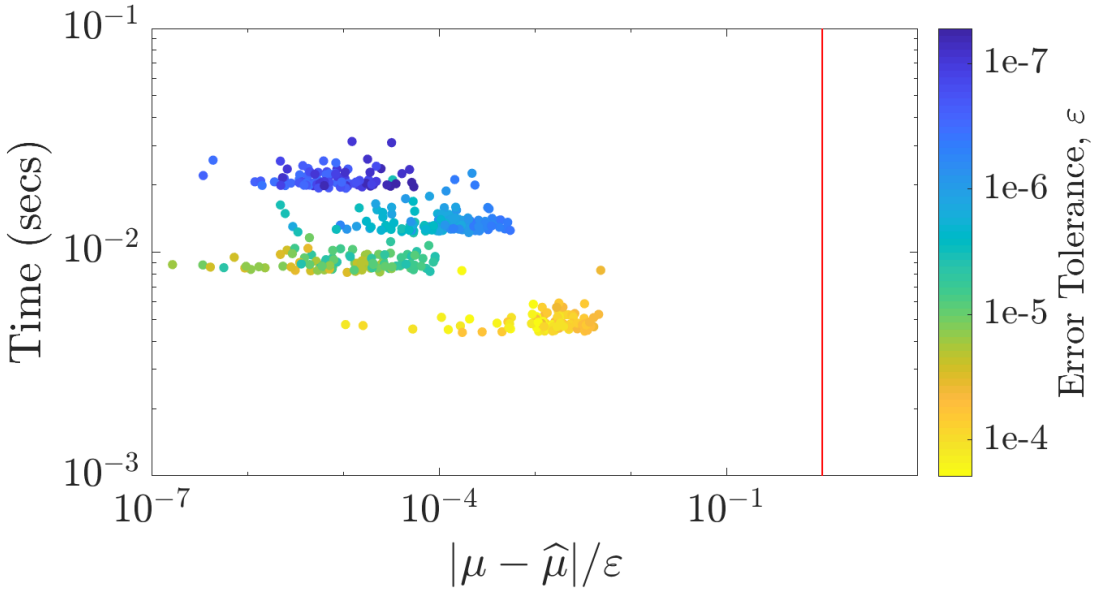

Figure 7.3. `cubBayesLattice_g`: Multivariate normal probability example using the GCV stopping criterion.

`cubBayesLattice_g`. But comparing the number of samples, n, used for integration provides more insight which directly relates to the alogrithm's computational cost. The `cubBayesLattice_g` used $n = 16384$ samples whereas `cubBayesNet_g` used $n = 32768$ samples even with $r = 1$ order kernel.

Amongst the three stopping criteria, GCV achieved the results faster than others but it is less conservative. One can also observe from the figures that the credible intervals are narrower than in Figure 7.1. This shows that `cubBayesNet_g` with $r = 1$ kernel more accurately approximates the integrand.

7.3 Keister's Example

This multidimensional integral function comes from [46] and is inspired by a physics application:

$$\begin{aligned} \mu &= \int_{\mathbb{R}^d} \cos(\|\boldsymbol{t}\|) \exp(-\|\boldsymbol{t}\|^2) \, \mathrm{d}\boldsymbol{t} \\ &= \int_{[0,1]^d} f_{\text{Keister}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{7.1}$$

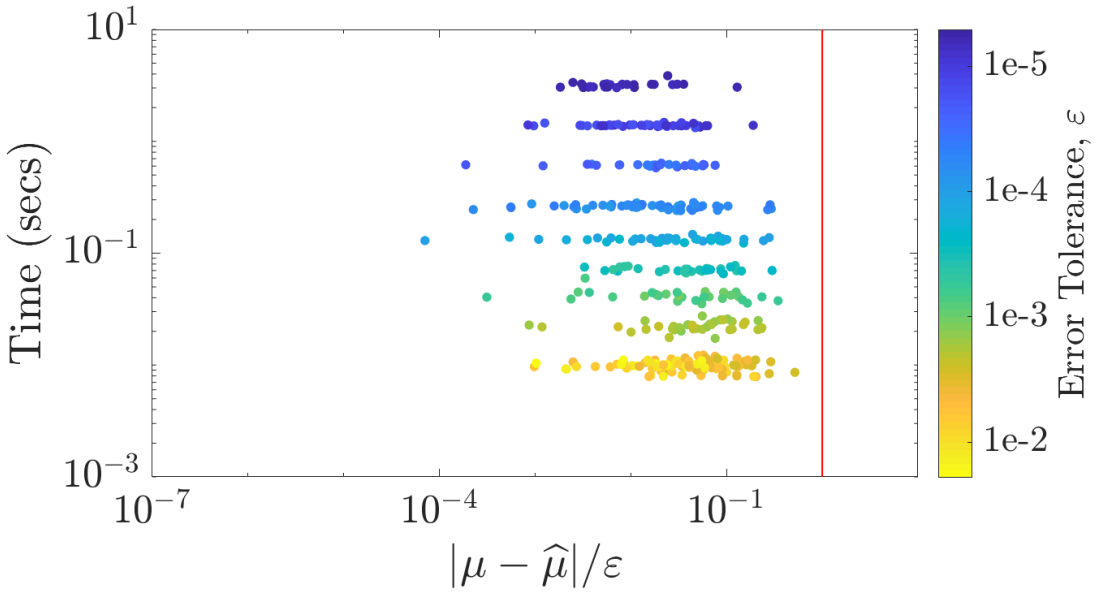

Figure 7.4. `cubBayesNet_g`: Multivariate normal probability example with empirical Bayes stopping criterion.

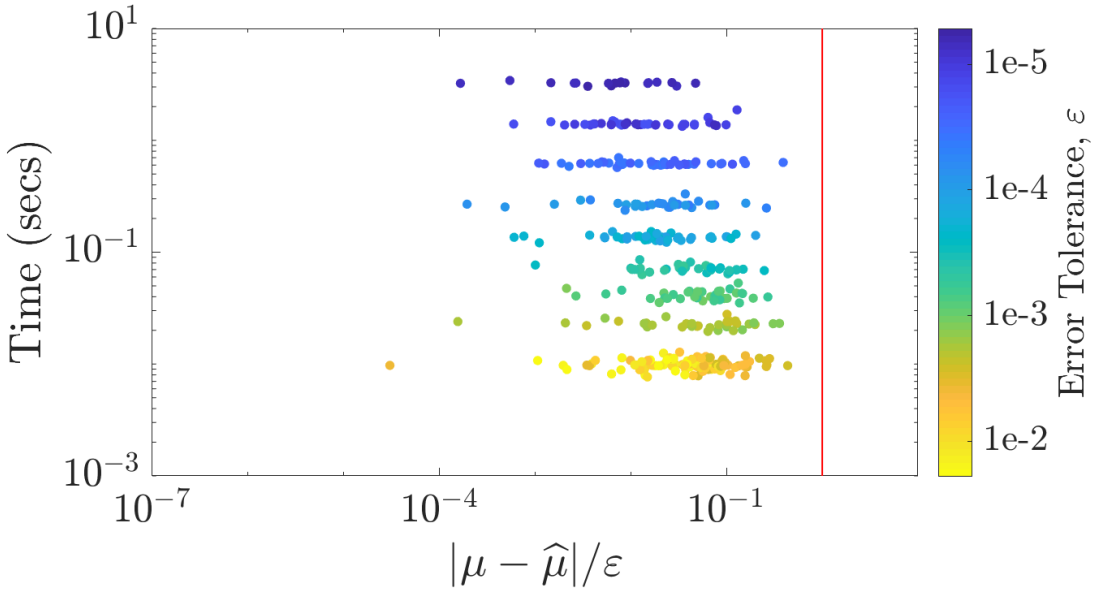

Figure 7.5. `cubBayesNet_g`: Multivariate normal probability example with the full-Bayes stopping criterion.

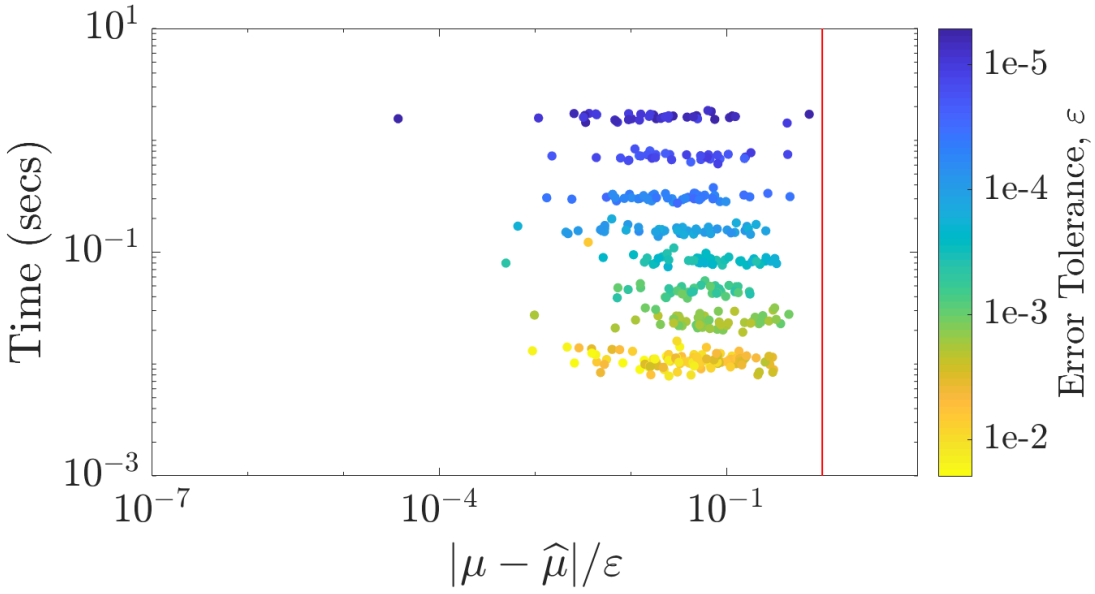

Figure 7.6. `cubBayesNet_g`: Multivariate normal probability example with the GCV stopping criterion.

where

$$f_{\text{Keister}}(\boldsymbol{x}) = \pi^{d/2} \cos\left(\left\|\Phi^{-1}(\boldsymbol{x})/2\right\|\right),$$

and Φ is the standard normal distribution. The true value of μ can be calculated iteratively in terms of a quadrature as follows:

$$\mu = \frac{2\pi^{d/2} I_c(d)}{\Gamma(d/2)}, \quad d = 1, 2, \ldots$$

where Γ denotes the gamma function, and

$$\begin{aligned}
I_c(1) &= \frac{\sqrt{\pi}}{2\exp(1/4)}, \\
I_s(1) &= \int_{x=0}^{\infty} \exp(-\boldsymbol{x}^T\boldsymbol{x}) \sin(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&= 0.4244363835020225, \\
I_c(2) &= \frac{1 - I_s(1)}{2}, \qquad I_s(2) = \frac{I_c(1)}{2} \\
I_c(j) &= \frac{(j-2)I_c(j-2) - I_s(j-1)}{2}, \qquad j = 3, 4, \ldots \\
I_s(j) &= \frac{(j-2)I_s(j-2) - I_c(j-1)}{2}, \qquad j = 3, 4, \ldots.
\end{aligned}$$

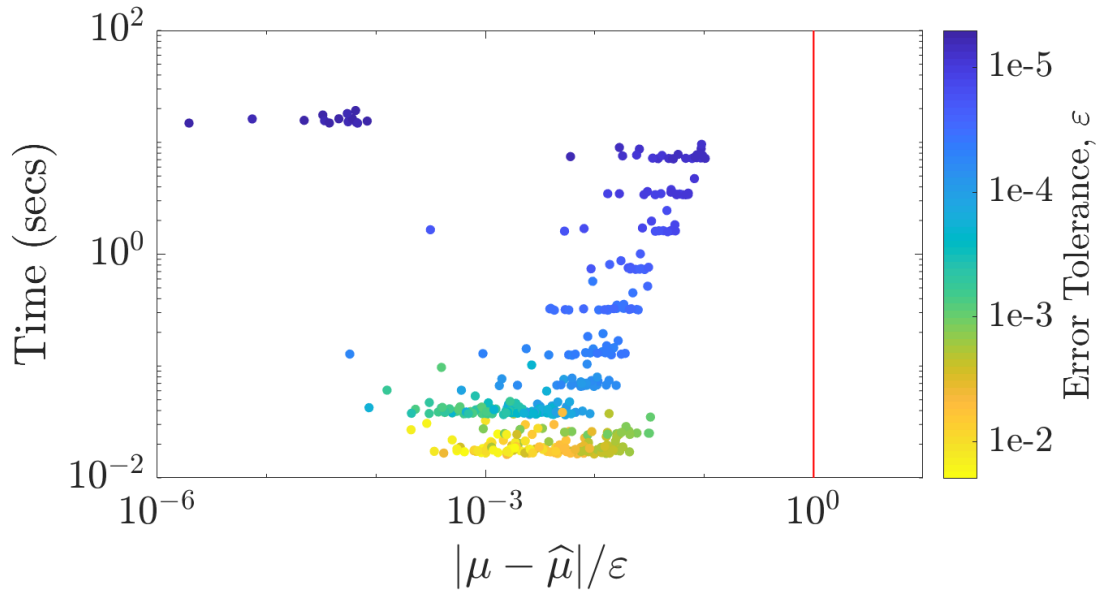

Figure 7.7. `cubBayesLattice_g`: Keister example using the empirical Bayes stopping criterion.

7.3.1 Using `cubBayesLattice_g`. Figures 7.7, 7.8 and 7.9 summarize the numerical tests for this integral. We used the Sidi's C^1 periodization, dimension $d = 4$, and $r = 2$. As we can see the GCV stopping criterion achieved the results faster than the others but it is less conservative similar to the multivariate Gaussian case.

7.3.2 Using `cubBayesNet_g`. Figures 7.10, 7.11 and 7.12 summarize the numerical tests for this case. We used dimension $d = 4$, and $r = 1$. No periodization transform was used as the integrand need not be periodic. In this example, we use $r = 1$ order kernel whereas in Section 7.3.1, $r = 2$ kernel was used. This necessitates `cubBayesNet_g` to use more samples for integration. As observed from the figures, the GCV stopping criterion achieved the results faster than the others but it is less conservative which is also the case with the multivariate Gaussian example.

7.4 Option Pricing

The price of financial derivatives can often be modeled by high dimensional

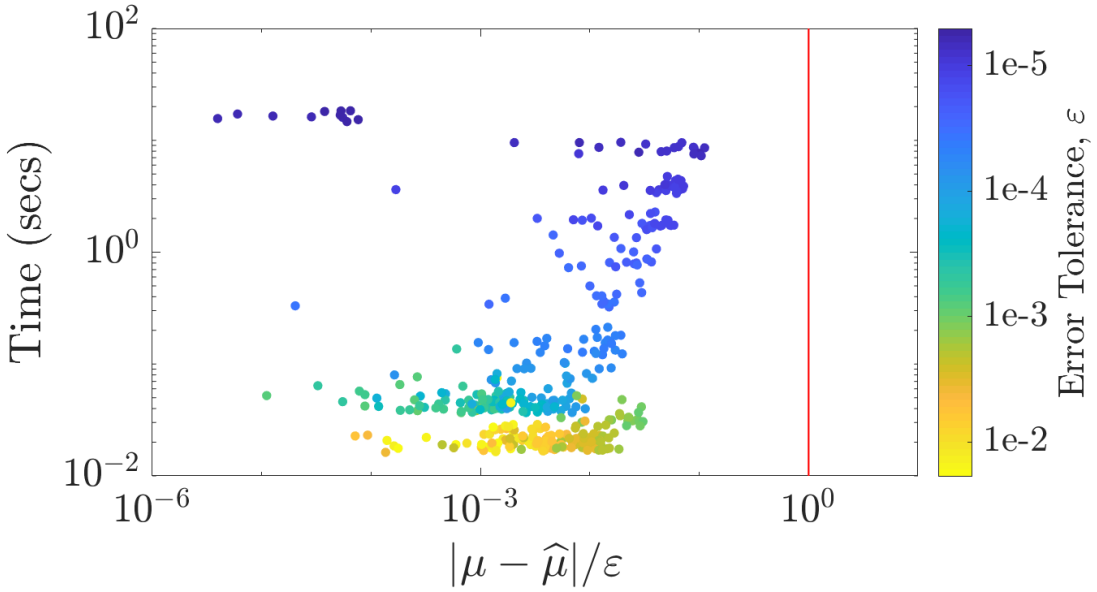

Figure 7.8. `cubBayesLattice_g`: Keister example using the full Bayes stopping criterion.

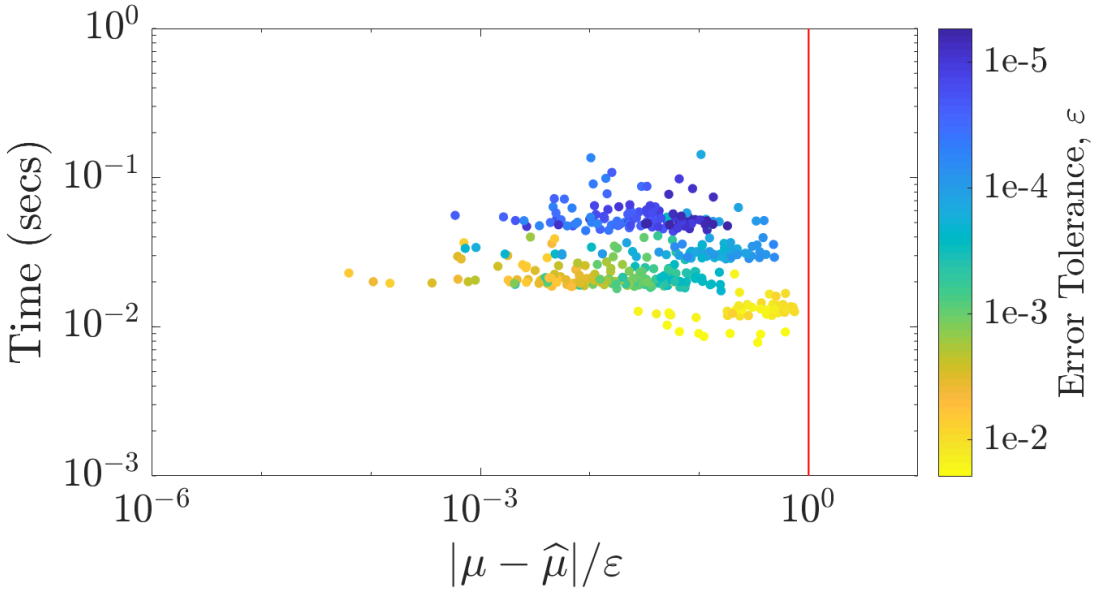

Figure 7.9. `cubBayesLattice_g`: Keister example using the GCV stopping criterion.

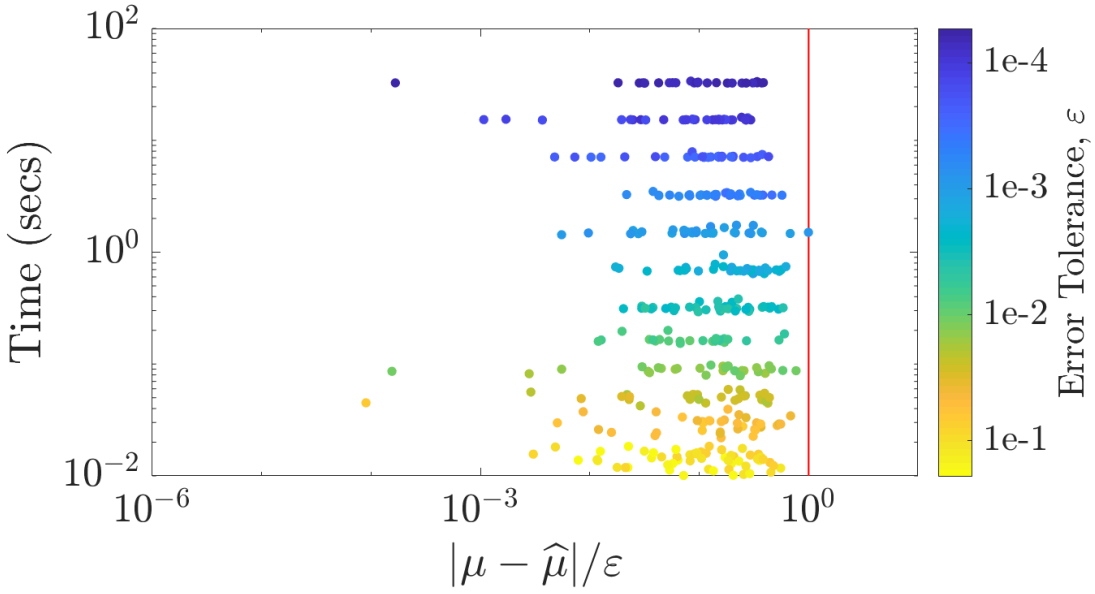

Figure 7.10. `cubBayesNet_g`: Keister example using the empirical Bayes stopping criterion.

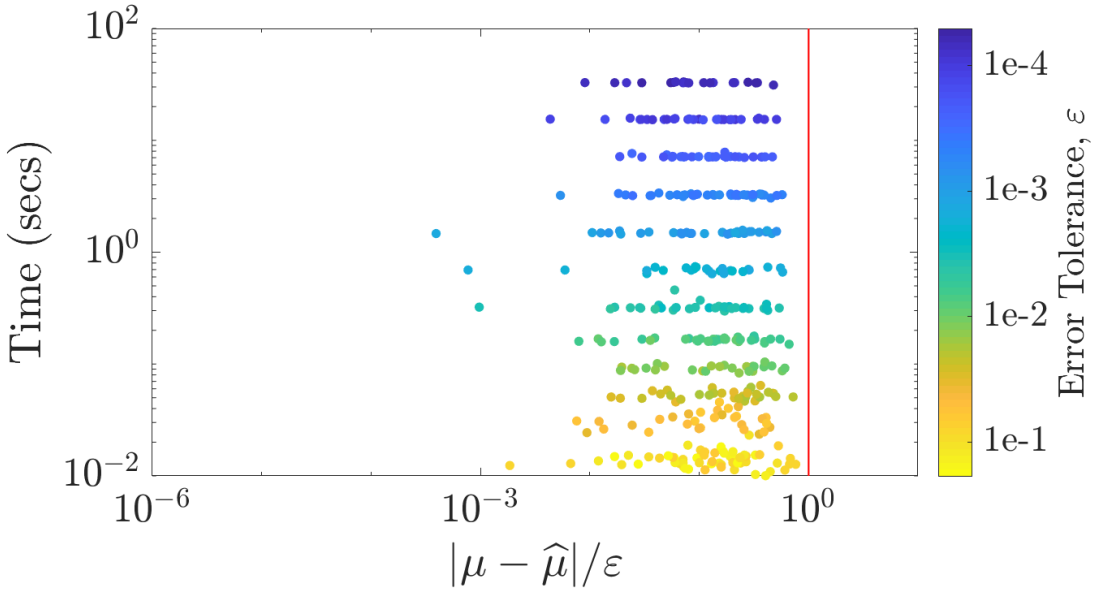

Figure 7.11. `cubBayesNet_g`: Keister example using the full-Bayes stopping criterion.

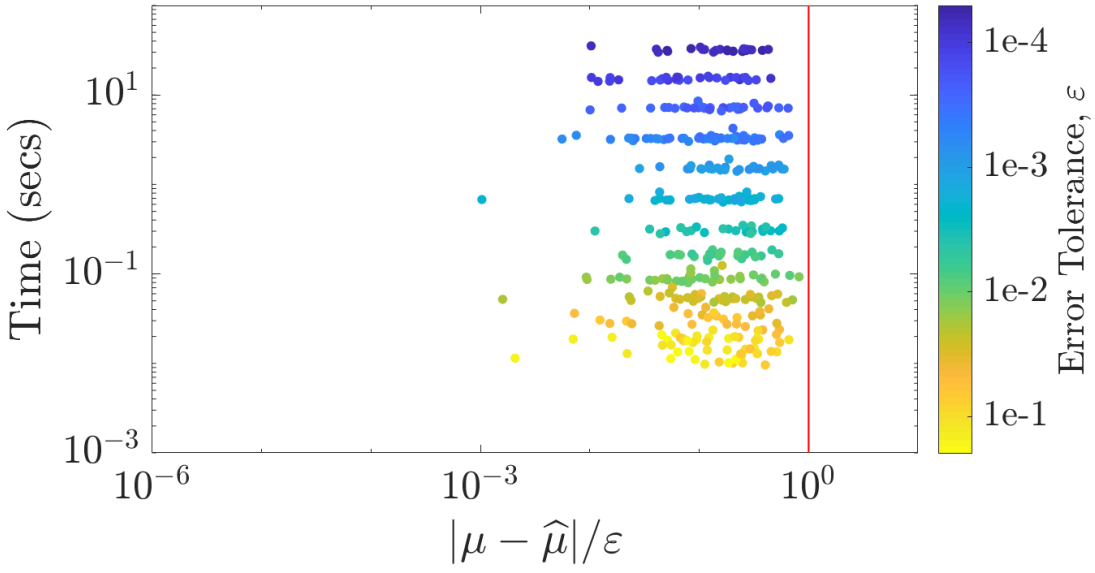

Figure 7.12. `cubBayesNet_g`: Keister example using the GCV stopping criterion.

integrals. If the underlying asset is described in terms of a discretized geometric Brownian motion, then the fair price of the option is:

$$\mu = \int_{\mathbb{R}^d} \text{payoff}(\boldsymbol{z}) \frac{\exp(\frac{1}{2}\boldsymbol{z}^T \Sigma^{-1} \boldsymbol{z})}{\sqrt{(2\pi)^d \det(\Sigma)}} \, \mathrm{d}\boldsymbol{z} = \int_{[0,1]^d} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

where payoff$(\cdot)$ defines the discounted payoff of the option,

$$\Sigma = (T/d)\big(\min(j,k)\big)_{j,k=1}^{d} = \mathsf{L}\mathsf{L}^T,$$

$$f(\boldsymbol{x}) = \text{payoff}\left(\mathsf{L}\begin{pmatrix} \Phi^{-1}(x_1) \\ \vdots \\ \Phi^{-1}(x_d) \end{pmatrix}\right).$$

The Asian arithmetic mean call option has a payoff of the form

$$\text{payoff}(\boldsymbol{z}) = \max\left(\frac{1}{d}\sum_{j=1}^{d} S_j(\boldsymbol{z}) - K, 0\right) \mathrm{e}^{-rT},$$

$$S_j(\boldsymbol{z}) = S_0 \exp\big((r - \sigma^2/2)jT/d + \sigma\sqrt{T/d}\, z_j\big).$$

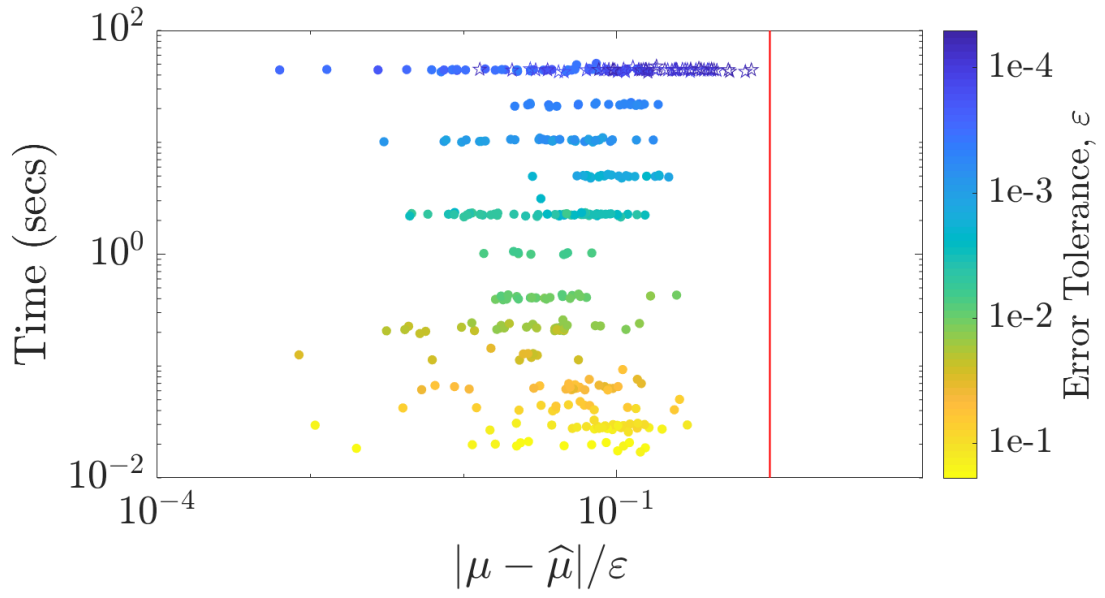

Figure 7.13. `cubBayesLattice_g`: Option pricing using the empirical Bayes stopping criterion. The hollow stars indicate the algorithm has not met the error threshold ϵ even with using maximum n.

Here, T denotes the time to maturity of the option, d the number of time steps, S_0 the initial price of the stock, r the interest rate, σ the volatility, and K the strike price.

7.4.1 Using `cubBayesLattice_g`. The Figures 7.13, 7.14 and 7.15 summarize the numerical results for this example using $T = 1/4$, $d = 13$, $S_0 = 100$, $r = 0.05$, $\sigma = 0.5$, $K = 100$. Moreover, L is chosen to be the matrix of eigenvectors of $\mathsf{\Sigma}$ times the square root of the diagonal matrix of eigenvalues of $\mathsf{\Sigma}$. Because the integrand has a kink caused by the max function, it does not help to use a periodizing transform that is very smooth. We chose the baker's transform (4.16) and $r = 1$.

7.4.2 Using `cubBayesNet_g`. The Figures 7.16, 7.17 and 7.18 summarize the numerical results for the option pricing example using the same values for, T, d, S_0, r, σ, K, as in Section 7.4.1. As mentioned before, this integrand has a kink caused by the max function, so, `cubBayesNet_g` could be more efficient than `cubBayesLat-`

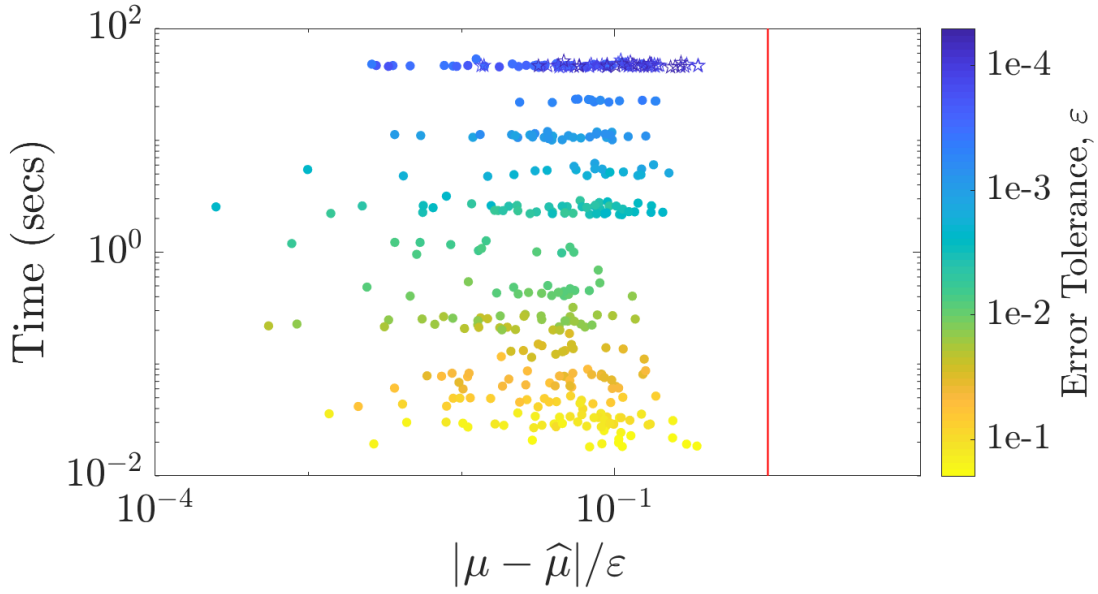

Figure 7.14. `cubBayesLattice_g`: Option pricing using the full Bayes stopping criterion. The hollow stars indicate the algorithm has not met the error threshold ϵ even with using maximum n.

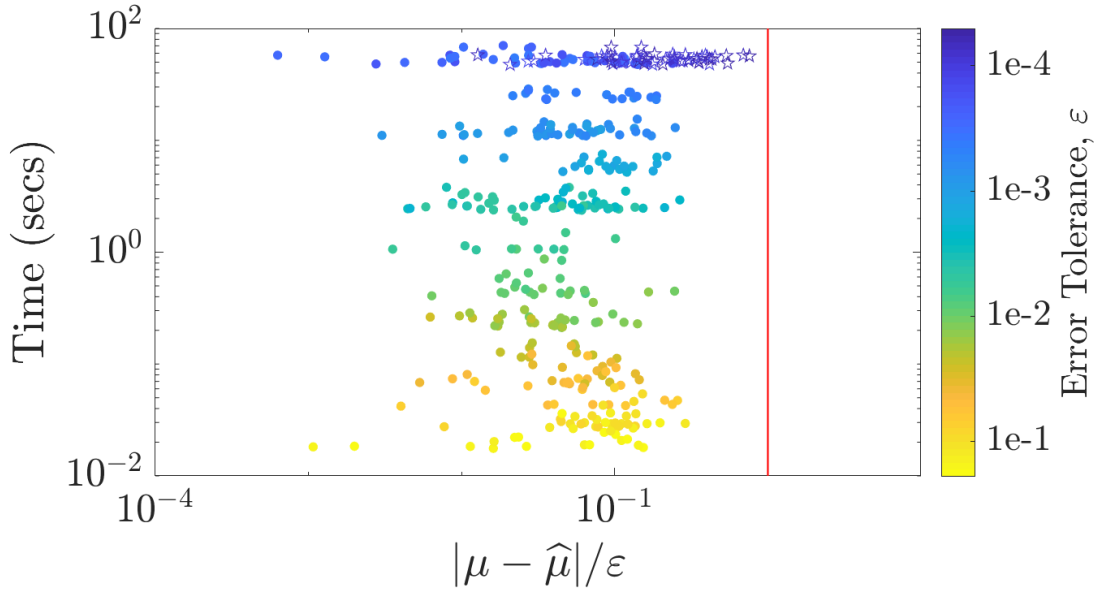

Figure 7.15. `cubBayesLattice_g`: Option pricing using the GCV stopping criterion. The hollow stars indicate the algorithm has not met the error threshold ϵ even with using maximum n.

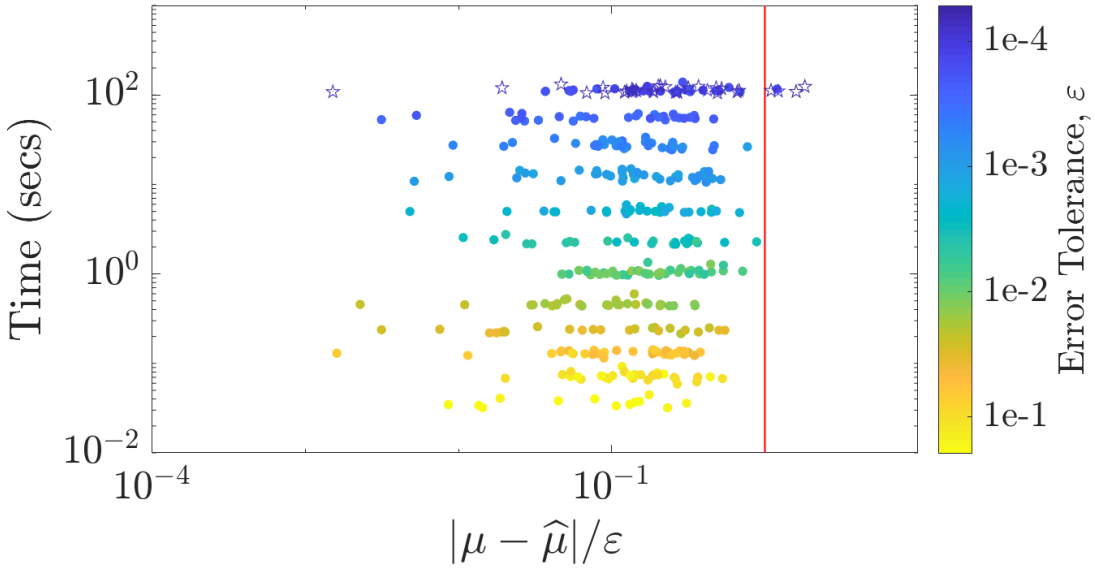

Figure 7.16. `cubBayesNet_g`: Option pricing using the empirical Bayes stopping criterion. The hollow stars indicate the algorithm has not met the error threshold ϵ even with using maximum n.

`tice_g`, as no periodization transform is required. This can be observed from the number of samples used for intgration to meet the same error threshold. For the error tolerance, $\varepsilon = 10^{-3}$, `cubBayesLattice_g` used $n = 2^{20}$ samples, whereas `cubBayesNet_g` used $n = 2^{17}$ samples.

7.5 Discussion

As shown in Figures 7.1 to 7.18, both the algorithms computed the integral within user specified threshold most of the time except on a few occasions. This is especially the case with option pricing example due to the complexity and high dimension of the integrand. Also notice that the `cubBayesLattice_g` algorithm finished within 10 seconds for Keister and multivariate Gaussian. Option pricing took closer to 70 seconds due to the complexity of the integrand.

Another noticeable aspect from the plots of `cubBayesLattice_g` is how much the error bounds differ from the true error. For option pricing example, the error

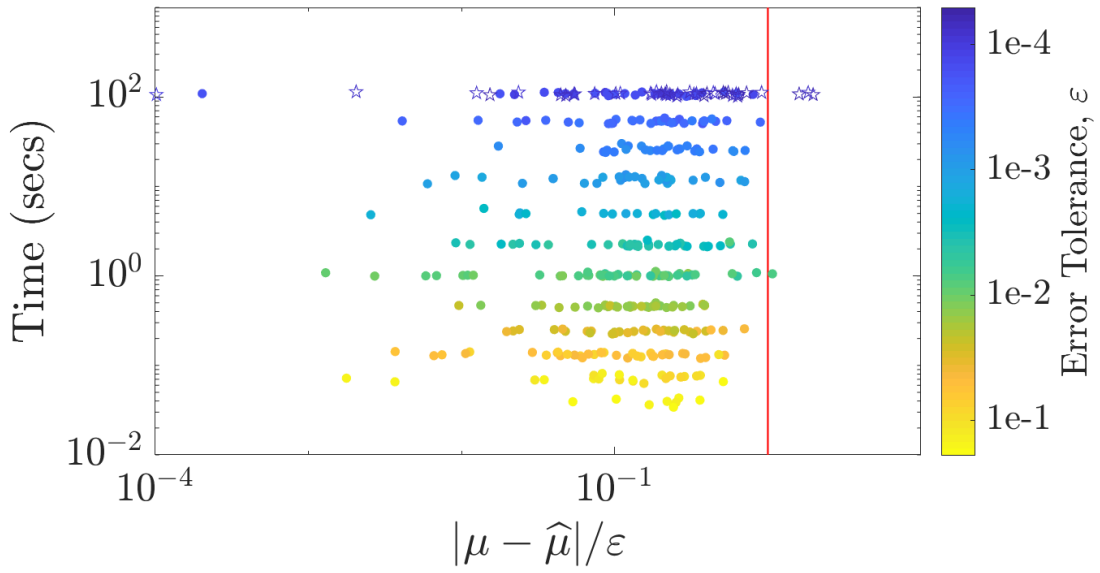

Figure 7.17. `cubBayesNet_g`: Option pricing using the full-Bayes stopping criterion. The hollow stars indicate the algorithm has not met the error threshold ϵ even with using maximum n.

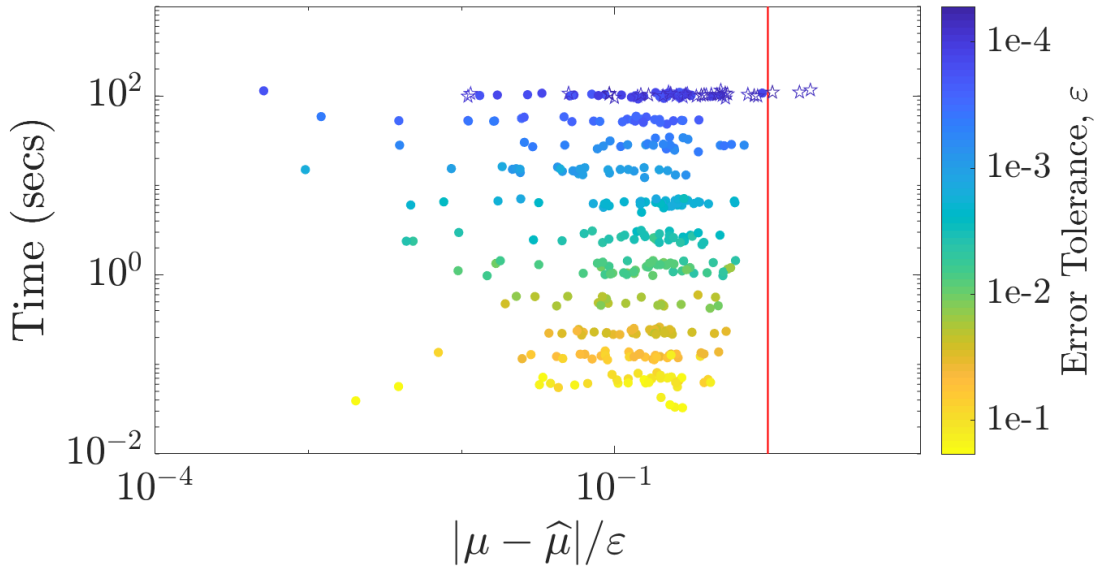

Figure 7.18. `cubBayesNet_g`: Option pricing using the GCV stopping criterion. The hollow stars indicate the algorithm has not met the error threshold ϵ even with using maximum n.

bound is not as conservative as it is for the multivariate Gaussian and Keister examples. A possible reason is that the latter integrands are significantly smoother than the covariance kernel. This is a matter for further investigation.

Most noticeable aspect from the plots of `cubBayesNet_g` is how closer the error bounds are to the true error. This shows that the `cubBayesNet_g`'s estimation of expected error in the stopping criterion is very accurate. Similar to `cubBayesLattice_g`, it missed meeting the given error threshold for the option pricing example, as marked by the hollow stars, for $\varepsilon = 10^{-4}$. The algorithm reached max allowed number of samples, $n = 2^{20}$ due to the complexity of the integrand.

7.6 Comparison with `cubMC_g`, `cubLattice_g` and `cubSobol_g`

GAIL library provides variety of numerical integration algorithms based on different theoretical foundations, We would like to compare how our algorithms perform relatively to these. We consider three GAIL algorithms 1) `cubMC_g`, a simple Monte-Carlo method for multi-dimensional integration, 2) `cubLattice_g`, a quasi-Monte-Carlo method using Lattice points, and 3) `cubSobol_g`, a quasi-Monte-Carlo method using Sobol points.

7.6.1 Keister integral. The Table 7.1 summarizes the performance of the methods MC, Lattice, Sobol, BayesLat, and BayesSob—which refer to the GAIL cubatures, `cubMC_g`, `cubLattice_g`, `cubSobol_g`, `cubBayesLattice_g`, `cubBayesNet_g`, respectively for estimating Keister integral defined in (7.1). We conducted two simulations with $d = 3$ and 8. In the case of $d = 3$, all five methods succeeded completely, meaning, the absolute error is less than given tolerance, i.e., $|\mu - \hat{\mu}| \leq \varepsilon$, where $\hat{\mu}$ is a cubature's approximated value. The fastest method was `cubBayesLattice_g`. In the case of $d = 8$, `cubSobol_g` achieved 100% success rate and was the fastest. But `cubBayesLattice_g` was competitive and had the smallest average absolute error.

`cubBayesNet_g` used lowest number of samples but was slower than `cubSobol_g`.

Table 7.1. Comparison of average performance of cubatures for estimating the integral (7.1) for 1000 independent runs. These results can be conditionally reproduced with the script, `KeisterCubatureExampleBayes.m`, in GAIL.

$d = 3,\ \varepsilon = 0.005$					
Method	MC	Lattice	Sobol	BayesLat	BayesSobol
Absolute Error	0.001 100	0.000 510	0.000 520	0.000 430	0.000 560
Tolerance Met	100%	100%	100%	100%	100%
n	2 500 000	4100	3900	1000	1900
Time (seconds)	0.1800	0.0069	0.0054	0.0029	0.0700

$d = 8,\ \varepsilon = 0.050$					
Method	MC	Lattice	Sobol	BayesLat	BayesSobol
Absolute Error	0.012 000	0.015 000	0.007 300	0.001 800	0.008 300
Tolerance Met	100%	99%	100%	100%	100%
n	7 400 000	15 000	16 000	66 000	8200
Time (seconds)	1.2000	0.0220	0.0160	0.2100	0.3500

7.6.2 Multivariate Gaussian. The Table 7.2 summarizes the performance of the methods MC, Lattice, Sobol, BayesLat, and BayesSob for estimating the multi-dimensional Gaussian probability $\mathbf{X} \sim \mathbf{N}(\mu, \mathbf{\Sigma})$. This experiment demonstrates our algorithm's ability to handle high-dimensional integral.

We conducted two simulations with different Σ and estimation intervals $(\mathbf{a}, \mathbf{b})$ but fixed $\mu = 0$ and required error threshold, $\varepsilon = 10^{-3}$. In the first case, all five methods succeeded completely. The fastest method was `cubBayesLattice_g` but `cubBayesNet_g` used the lowest number of samples. In the second case also, all five methods succeeded, but `cubLattice_g` was the fastest. The `cubBayesNet_g` was competitive and had the smallest average absolute error using lowest number of samples.

The `cubBayesLattice_g` achieved the next lowest average error but was slower than `cubSobol_g`.

Table 7.2. Comparison of average performance of cubatures for estimating the $d = 20$ Multivariate Normal (2.26) for 1000 independent runs with $\varepsilon = 10^{-3}$. These results can be conditionally reproduced with the script, `MVNCubatureExampleBayes.m`, in GAIL.

	$\Sigma = \mathsf{I}_d,\ \boldsymbol{b} = -\boldsymbol{a} = (3.5, \cdots, 3.5)$				
Method	MC	Lattice	Sobol	BayesLat	BayesSobol
Absolute Error	2.20E−16	2.70E−14	2.70E−14	2.20E−16	2.20E−16
Tolerance Met	100%	100%	100%	100%	100%
n	10 000	1000	1000	1000	260
Time (seconds)	0.0410	0.0820	0.0710	0.0650	0.0790

	$\Sigma = 0.4\ \mathsf{I}_d + 0.6\ \mathbf{1}\mathbf{1}^T,\ \boldsymbol{a} = (-\infty, \cdots, -\infty),\ \boldsymbol{b} = \sqrt{d}(U_1, \cdots, U_d)$				
Method	MC	Lattice	Sobol	BayesLat	BayesSobol
Absolute Error	2.30E−4	2.10E−4	4.40E−4	1.00E−4	4.80E−5
Tolerance Met	100%	100%	100%	100%	100%
n	10 000	1000	1000	1000	260
Time (seconds)	0.0350	0.0120	0.0140	0.0150	0.0300

7.7 Shape Parameter Fine-tuning

Allowing the kernel shape parameter to vary for each dimension could improve the accuracy of numerical integration when the integrand under consideration has only very low effective dimension as in the Option Pricing example we demonstrated. We demonstrate this advantage by integrating a function that is not symmetric across dimensions,

$$f(\boldsymbol{x}) = \sum_{j=1}^{d} v_j \sin(2\pi x_j^2) \tag{7.2}$$

which has known integral

$$\int_{[0,1)^d} f(\boldsymbol{x}) = \frac{1}{2}\,\texttt{fresnels}(d) \sum_{j=1}^{d} v_j$$

where `fresnels` is the Fresnel Sine integral,

$$\texttt{fresnels}(z) = \int_0^z \sin\left(\frac{\pi t^2}{2}\right) \mathrm{d}t.$$

Table 7.3. Comparison of average performance of Bayesian Cubature with common shape parameter vs dimension specific shape parameter for estimating the $d = 3$ Fresnel Sine integral. These results can be conditionally reproduced with the script, `demoMultiTheta.m`, in GAIL.

	Fresnel Sine Integral in $d = 3$	
Method	`OneTheta`	`MultiTheta`
Absolute Error	0.000 23	0.063 00
n	4100	260
Time (seconds)	0.0270	0.0230

The results are summarized from the two different approaches in Table 7.3. The first method, called `OneTheta`, uses common shape parameter across all the dimensions, whereas the second method, called `MultiTheta`, allows the shape parameters to vary across the dimensions. In the `MultiTheta` method, the shape parameter search is multivariate, so the magnitude of shape parameter depends on

the integrand's magnitude in each dimensions. We have chosen an integrand particularly to demonstrate this aspect (7.2) where we used $d = 3$ and the constants $\boldsymbol{\upsilon} = (10^{-4}, 1, 10^{4})$. The choice of magnitude variations in constants $\boldsymbol{\upsilon}$ allows to make the integrand vary significantly across dimensions.

We ran this test for 1000 times. In comparison, both the methods successfully computed the integral all the time but `MultiTheta` was slightly faster. The `MultiTheta` method used less number of samples but the integration error was bigger than the `OneTheta`. For the same number of samples, the `OneTheta` method will be much faster since the shape parameter search is faster. The `MultiTheta` method is useful in scenarios where we want to use smaller size, n, and the integrand varies significantly across dimensions.

CHAPTER 8
CONCLUSION AND FUTURE WORK

8.1 Conclusion

We have developed a fast, automatic Bayesian cubature that estimates the high dimensional integral within a user defined error tolerance that occur in many scientific computing such as finance, machine learning, imaging, etc. The stopping criteria arise from assuming the integrand to be a Gaussian process. In Section 2.2, we developed three criteria: empirical Bayes, full Bayes, and generalized cross-validation. Empirical-Bayes uses maximum-likelihood to optimally choose the parameters, where posterior of the parameters given the integrand values is maximized. Alternatively, full-Bayes assumes non-informative prior on the parameters and then computes posterior distribution of the integral μ, which leads to a t-distribution to obtain the parameters. Generalized cross-validation extends the concept of cross-validation to construct an objective which in turn is maximized.

The computational cost of the automatic Bayesian cubature can be dramatically reduced if the covariance kernel matches the nodes. We have demonstrated two such matches in practice. The first algorithm was based on rank-1 lattice nodes and shift-invariant kernels where the matrix-vector multiplications can be accomplished using the fast Fourier Transform. The second algorithm was based on Sobol' points with first order Walsh kernel where the matrix-vector multiplications can be accomplished using the fast Walsh transform. Three integration problems illustrate the performance of our automatic Bayesian cubature algorithms.

For faster computations one could use fixed order kernels in `cubBayesLattice_g`, but for more advanced usage, we have added a kernel variation in Section 4.3 that allows one to optimally choose the kernel order without the constraint of being

an even integer.

During the numerical experiments, we noticed a computation step that causes inaccuracy due to a cancellation error in the estimation of stopping criterion. We have developed a novel technique in Section 6.1 to overcome this cancellation error using the inherent structure of the shift-invariant kernel used in our algorithm.

In Section 3.5.1, we have analytically computed the gradient of the objective function and the shift invariant kernel to use with steepest descent in kernel parameters search. Quasi-Monte Carlo cubature methods are efficient [47] even if the dimension is high given that the effective dimension is low. To take advantage of low effective dimension, one should not fix the kernel shape parameter across all the dimensions. In this situation, steepest descent methods come in handy as one searches for parameters in multi-dimensions.

8.2 Future Work

We demonstrated the capability of our new Bayesian cubature algorithms to successfully compute the integrals faster within the user defined error tolerances. But there are possibilities for improvements and new areas of applications. Some of the improvement ideas are listed here:

- Higher order digital sequences and digital shift invariant kernels [43] [48]: We could improve the computation speed of `cubBayesNet_g` for smoother integrands using higher order digital sequences and matching kernels, which have the potential of being another match that satisfies the conditions in Section 3. The fast Bayesian transform would correspond to a fast Walsh transform similar to the second algorithm we demonstrated. For such kernels and the first order Walsh kernel we demonstrated, periodicity is not assumed, however, special structure of both the sequences and the kernels are required to take advantage

of integrand smoothness.

- Control variates: Hickernell et.al [14] [49] adapted control variates for Quasi-Monte Carlo. Control variates are commonly used to improve the efficiency of IID Monte Carlo integration. One should be able to adapt our Bayesian cubature to control variates, i.e., assuming

$$f = \mathcal{GP}\left(\beta_0 + \beta_1\, g_1 + \cdots + \beta_p\, g_p,\ s^2 C\right),$$

for some choice of vector of functions $\boldsymbol{g} = \{g_1, \ldots, g_p\}$, where $\boldsymbol{g} : [0,1)^d \to \mathbb{R}^p$ whose integrals are known $\mu_{\boldsymbol{g}} := \int_{[0,1)^d} \boldsymbol{g}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and some parameters $\beta_0, \ldots, \beta_p$ in addition to the s and C, then

$$\mu := \int_{[0,1)^d} f(\boldsymbol{x})\mathrm{d}x = \int_{[0,1)^d} h_\beta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \text{ where } h_\beta(\boldsymbol{x}) := f(\boldsymbol{x}) + \boldsymbol{\beta}^T(\mu_{\boldsymbol{g}} - \boldsymbol{g}(\boldsymbol{x})).$$

Here $\boldsymbol{g}$ are the functions on which the QMC method does a good job of integrating it without error. The goal is to choose an optimal $\boldsymbol{\beta}$ to make

$$\widehat{\mu}_{\boldsymbol{\beta},n} := \frac{1}{n}\sum_{i=0}^{n-1} h_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$$

sufficiently close to μ with the least expense, n, possible. The efficacy of this approach has not yet been explored.

- Steepest descent: The kernels's optimal shape parameter searched using steepest descent with kernels gradient could sometime get into local minima. This needs more understanding and enhancements.

- Gaussian diagnosis: We assumed the integrand to be an instance of a Gaussian process. One could attempt to prove if this is a good assumption using statistical diagnosis for goodness of fit.

- Parallel Algorithm: For more demanding high performance computing applications, where the precision requirements are high, our algorithms will try to use

large number samples leading to longer computation time. One approach to overcome this constraint is to use Parallel computing techniques to speedup the algorithm. Most time consuming parts of our algorithm are shape parameter search and fast Bayesian transform computation. Fast Fourier transform (FFT) and Fast Walsh transform are easily amenable to parallelization. There exist plenty of prior work that can be adapted to work with our algorithms. We use radix-2 FFT. One could use a higher radix FFT to make the computations faster.

Another area of improvement is the parameter search. We explored the steepest descent algorithm but the speedup was not significant. One could explore higher order algorithms such as Newton method, which could find the minima faster. Fast Bayesian transforms are repeatedly computed in every step of the parameter search if it can be avoided by interpolation or other techniques, this could significantly speedup the algorithm.

One could also use GPU to run the whole code of our Bayesian Cubature algorithms or just the FFT/FWHT part to get a easier speedup.

BIBLIOGRAPHY

[1] R. Jagadeeswaran and F. J. Hickernell, "Fast automatic Bayesian cubature using lattice sampling," *Statist. Comp.*, vol. 44, pp. 2559–2583, 2019.

[2] F. J. Hickernell and R. Jagadeeswaran, "Comment on "Probabilistic integration: A role in statistical computation?"," *Statist. Sci*, vol. 34, pp. 23–28, 2019.

[3] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, ser. Applications of Mathematics. New York: Springer-Verlag, 2004, vol. 53.

[4] A. Keller, "Quasi-Monte Carlo image synthesis in a nutshell," in *Monte Carlo and Quasi-Monte Carlo Methods 2012*, ser. Springer Proceedings in Mathematics and Statistics, J. Dick, F. Y. Kuo, G. W. Peters, and I. H. Sloan, Eds., vol. 65. Springer Berlin Heidelberg, 2013, pp. 213–249.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[6] F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic, "Probabilistic integration: A role in statistical computation?" *Statist. Sci.*, 2018+, to appear.

[7] P. Diaconis, "Bayesian numerical analysis," in *Statistical Decision Theory and Related Topics IV, Papers from the 4th Purdue Symp., West Lafayette, Indiana 1986*, S. S. Gupta and J. O. Berger, Eds. Springer-Verlag, New York, 1988, vol. 1, pp. 163–175.

[8] A. O'Hagan, "Bayes-Hermite quadrature," *J. Statist. Plann. Inference*, vol. 29, pp. 245–260, 1991.

[9] K. Ritter, *Average-Case Analysis of Numerical Problems*, ser. Lecture Notes in Mathematics. Berlin: Springer-Verlag, 2000, vol. 1733.

[10] C. E. Rasmussen and Z. Ghahramani, "Bayesian Monte Carlo," in *Advances in Neural Information Processing Systems*, S. Thrun, L. K. Saul, and K. Obermayer, Eds. MIT Press, 2003, vol. 15, pp. 489–496.

[11] F. J. Hickernell, "The trio identity for quasi-Monte Carlo error analysis," in *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Stanford, USA, August 2016*, ser. Springer Proceedings in Mathematics and Statistics, P. Glynn and A. Owen, Eds. Springer-Verlag, Berlin, 2018, pp. 13–37, arXiv:1702.01487.

[12] F. J. Hickernell and Ll. A. Jiménez Rugama, "Reliable adaptive cubature using digital sequences," in *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, ser. Springer Proceedings in Mathematics and Statistics, R. Cools and D. Nuyens, Eds., vol. 163. Springer-Verlag, Berlin, 2016, pp. 367–383, arXiv:1410.8615 [math.NA].

[13] Ll. A. Jiménez Rugama and F. J. Hickernell, "Adaptive multidimensional integration based on rank-1 lattices," in *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, ser. Springer Proceedings in Mathematics and Statistics, R. Cools and D. Nuyens, Eds., vol. 163. Springer-Verlag, Berlin, 2016, pp. 407–422, arXiv:1411.1966.

[14] F. J. Hickernell, Ll. A. Jiménez Rugama, and D. Li, "Adaptive quasi-Monte Carlo methods for cubature," in *Contemporary Computational Mathematics — a celebration of the 80th birthday of Ian Sloan*, J. Dick, F. Y. Kuo, and H. Woźniakowski, Eds. Springer-Verlag, 2018, pp. 597–619.

[15] A. OHagan, "Bayes-hermite quadrature," *Journal of Statistical Planning and Inference*, vol. 29(3), p. 245260, 1991.

[16] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press, 2006, (online version at `http://www.gaussianprocess.org/gpml/`).

[17] R. Brent, *Algorithms for Minimization Without Derivatives*. Prentice-Hall, 1973.

[18] G. Forsythe, M. Malcolm, and C. Moler, *Computer methods for mathematical computations*. Prentice-Hall, 1976.

[19] K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. G. Wilson, "Scalable log determinants for gaussian process kernel learning," *NIPS*, 2017, in press.

[20] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, pp. 307–403, 1979.

[21] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–223, 1979.

[22] G. Wahba, *Spline Models for Observational Data*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM, 1990, vol. 59.

[23] A. Genz, "Comparison of methods for the computation of multivariate normal probabilities," *Computing Science and Statistics*, vol. 25, pp. 400–405, 1993.

[24] J. Dick, F. Kuo, and I. H. Sloan, "High dimensional integration — the Quasi-Monte Carlo way," *Acta Numer.*, vol. 22, pp. 133–288, 2013.

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

[26] N. J. Higham, *Functions of matrices: theory and computation*. SIAM, 2008.

[27] F. J. Hickernell and H. Niederreiter, "The existence of good extensible rank-1 lattices," *J. Complexity*, vol. 19, pp. 286–300, 2003.

[28] F. J. Hickernell, "Quadrature error bounds with applications to lattice rules," *SIAM J. Numer. Anal.*, vol. 33, pp. 1995–2016, 1996, corrected printing of Sections 3-6 in ibid., **34** (1997), 853–866.

[29] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, C. W. Clark, and A. B. O. Dalhuis, "Digital library of mathematical functions," 2018. [Online]. Available: http://dlmf.nist.gov/

[30] S.-C. T. Choi, Y. Ding, F. J. Hickernell, L. Jiang, Ll. A. Jiménez Rugama, D. Li, R. Jagadeeswaran, X. Tong, K. Zhang, Y. Zhang, and X. Zhou, "GAIL: Guaranteed Automatic Integration Library (versions 1.0–2.2)," MATLAB software, 2013–2017. [Online]. Available: http://gailgithub.github.io/GAIL_Dev/

[31] D. Nuyens. [Online]. Available: https://people.cs.kuleuven.be/~dirk.nuyens/qmc-generators/

[32] A. Sidi, "Further extension of a class of periodizing variable transformations for numerical integration," *J. Comput. Appl. Math.*, vol. 221, pp. 132–149, 2008.

[33] I. M. Sobol', "The distribution of points in a cube and the approximate evaluation of integrals," *U.S.S.R. Comput. Math. and Math. Phys.*, vol. 7, pp. 86–112, 1967.

[34] H. Niederreiter, "Constructions of (t, m, s)-nets and (t, s)-sequences," *Finite Fields Appl.*, vol. 11, pp. 578–600, 2005.

[35] J. F. Baldeaux, "Higher order nets and sequences," Ph.D. dissertation, The School of Mathematics and Statistics at The University of New South Wales, June 2010.

[36] F. J. Hickernell and R. X. Yue, "The mean square discrepancy of scrambled (t, s)-sequences," *SIAM J. Numer. Anal.*, vol. 38, pp. 1089–1112, 2000.

[37] A. B. Owen, "Randomly permuted (t, m, s)-nets and (t, s)-sequences," pp. 299–317.

[38] J. Matoušek, "On the L_2-discrepancy for anchored boxes," *J. Complexity*, vol. 14, pp. 527–556, 1998.

[39] I. M. Sobol', "Uniformly distributed sequences with an additional uniformity property," *Zh. Vychisl. Mat. i Mat. Fiz.*, vol. 16, pp. 1332–1337, 1976.

[40] F. Y. Kuo and D. Nuyens, "Application of quasi-monte carlo methods to elliptic pdes with random diffusion coefficients a survey of analysis and implementation," *Foundations of Computational Mathematics*, vol. 16(6), pp. 1631–1696, 2016.

[41] D. Nuyens. [Online]. Available: https://people.cs.kuleuven.be/~dirk.nuyens/

[42] H. S. Hong and F. J. Hickernell, "Algorithm 823: Implementing scrambled digital nets," *ACM Trans. Math. Software*, vol. 29, pp. 95–109, 2003.

[43] D. Nuyens, "The construction of good lattice rules and polynomial lattice rules," 08 2013.

[44] P. Bratley and B. L. Fox, "Algorithm 659: Implementing Sobol's quasirandom sequence generator," *ACM Trans. Math. Software*, vol. 14, pp. 88–100, 1988.

[45] J. Dick, "Walsh spaces containing smooth functions an quasi-monte carlo rules of arbitrary high order," *SIAM J. Numer. Anal.*, vol. 46, no. 1519–1553, 2008.

[46] B. D. Keister, "Multidimensional quadrature algorithms," *Computers in Physics*, vol. 10, pp. 119–122, 1996.

[47] I. H. Sloan and H. Woźniakowski, "When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?" *J. Complexity*, vol. 14, pp. 1–33, 1998.

[48] J. Baldeaux, J. Dick, G. Leobacher, D. Nuyens, and F. Pillichshammer, "Efficient calculation of the worst-case error and (fast) component-by-component construction of higher order polynomial lattice rules," *Numerical Algorithms*, vol. 59, pp. 403–431, Mar. 2012.

[49] D. Li, "Reliable quasi-Monte Carlo with control variates," Master's thesis, Illinois Institute of Technology, 2016.

[50] R. Cools and D. Nuyens, Eds., *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, ser. Springer Proceedings in Mathematics and Statistics, vol. 163. Springer-Verlag, Berlin, 2016.