Karvonen, Toni; Kanagawa, Motonobu; Särkkä, Simo

# On the positivity and magnitudes of Bayesian quadrature weights

*Published in:*
STATISTICS AND COMPUTING

*DOI:*
10.1007/s11222-019-09901-0

Published: 04/10/2019

*Document Version*
Publisher's PDF, also known as Version of record

Statistics and Computing
https://doi.org/10.1007/s11222-019-09901-0

# On the positivity and magnitudes of Bayesian quadrature weights

**Toni Karvonen**[1] · **Motonobu Kanagawa**[2,3] · **Simo Särkkä**[1]



**Abstract**
This article reviews and studies the properties of Bayesian quadrature weights, which strongly affect stability and robustness of the quadrature rule. Specifically, we investigate conditions that are needed to guarantee that the weights are positive or to bound their magnitudes. First, it is shown that the weights are positive in the univariate case if the design points locally minimise the posterior integral variance and the covariance kernel is totally positive (e.g. Gaussian and Hardy kernels). This suggests that gradient-based optimisation of design points may be effective in constructing stable and robust Bayesian quadrature rules. Secondly, we show that magnitudes of the weights admit an upper bound in terms of the fill distance and separation radius if the RKHS of the kernel is a Sobolev space (e.g. Matérn kernels), suggesting that quasi-uniform points should be used. A number of numerical examples demonstrate that significant generalisations and improvements appear to be possible, manifesting the need for further research.

**Keywords** Bayesian quadrature · Probabilistic numerics · Gaussian processes · Chebyshev systems · Stability

## 1 Introduction

This article is concerned with *Bayesian quadrature* (O'Hagan 1991; Rasmussen and Ghahramani 2002; Briol et al. 2019), a probabilistic approach to numerical integration and an example of a *probabilistic numerical method* (Larkin 1972; Hennig et al. 2015; Cockayne et al. 2019). Let $\Omega$ be a subset of $\mathbb{R}^d$, $d \geq 1$, and $\nu$ a Borel probability measure on $\Omega$. Given an *integrand* $f\colon \Omega \to \mathbb{R}$, the task is to approximate the integral

$$I_\nu(f) := \int_\Omega f \,\mathrm{d}\nu,$$

the solution of which is assumed not to be available in closed form. In Bayesian quadrature, a user specifies a prior distribution over the integrand as a Gaussian process $f_{\mathrm{GP}} \sim \mathcal{GP}(0, k)$ by choosing a positive-definite covariance kernel $k\colon \Omega \times \Omega \to \mathbb{R}$, so as to faithfully represent their knowledge about the integrand, such as its smoothness. The user then evaluates the true integrand at chosen design points $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \Omega$. By regarding the pairs $\mathcal{D} := \{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^n$ thus obtained as "observed data", the posterior distribution $I_\nu(f_{\mathrm{GP}}) \mid \mathcal{D}$ becomes a Gaussian random variable. This posterior distribution is useful for uncertainty quantification and decision making in subsequent tasks; this is one factor that makes Bayesian quadrature a promising approach in modern scientific computation, where quantification of discretisation errors is of great importance (Briol et al. 2019; Oates et al. 2017).

In Bayesian quadrature, the mean of the posterior over the integral is used as a quadrature estimate. The mean given is as a weighted average of function values:

$$\mathbb{E}\big[I_\nu(f_{\mathrm{GP}}) \mid \mathcal{D}\big] = \sum_{i=1}^n w_{X,i}^{\mathrm{BQ}} f(\boldsymbol{x}_i) \approx \int_\Omega f \,\mathrm{d}\nu,$$

where $w_{X,1}^{\mathrm{BQ}}, \ldots, w_{X,n}^{\mathrm{BQ}} \in \mathbb{R}$ are the weights computed with the kernel $k$, design points $X$ and the measure $\nu$ (see Sect. 2.1 for details). This form is similar to (quasi) Monte Carlo methods, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are (quasi) random points

✉ Toni Karvonen
toni.karvonen@aalto.fi

Motonobu Kanagawa
motonobu.kanagawa@gmail.com

Simo Särkkä
simo.sarkka@aalto.fi

1 Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland

2 University of Tübingen, Tübingen, Germany

3 Max Planck Institute for Intelligent Systems, Tübingen, Germany

Published online: 04 October 2019

from a suitable proposal distribution and $w_1, \ldots, w_n$ are the associated importance weights, positive by definition. This similarity naturally leads to the following question: Are the weights $w_{X,1}^{\mathrm{BQ}}, \ldots, w_{X,n}^{\mathrm{BQ}}$ of Bayesian quadrature positive? These weights are derived with no explicit positivity constraint, so in general some of them can be negative, which is observed in Huszár and Duvenaud (2012, Section 3.1.1). Therefore, the question can be stated as: *Under which conditions on the points and the kernel are the weights guaranteed to be positive?*

This question is important both conceptually and practically. On the conceptual side, positive weights are more natural, given that the weighted sample $(w_i, \boldsymbol{x}_i)_{i=1}^n$ can be interpreted as an approximation of the positive probability measure $\nu$; in fact, the Bayesian quadrature weights provide the best approximation of the representer of $\nu$ in the reproducing kernel Hilbert space (RKHS) of the covariance kernel, provided that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are fixed (see Sect. 2.2). Thus, if the weights are positive, then each weight $w_i$ can be interpreted as representing the "importance" of the associated point $\boldsymbol{x}_i$ for approximating $\nu$. This interpretation may be more acceptable to users familiar with Monte Carlo methods, encouraging them to adopt Bayesian quadrature.

On the practical side, quadrature rules with positive weights enjoy the advantage of being numerically more stable against errors in integrand evaluations. In fact, besides Monte Carlo methods, many other practically successful or in some sense optimal rules have positive weights. Some important examples include Gaussian (Gautschi 2004, Section 1.4.2) and Clenshaw–Curtis quadrature (Clenshaw and Curtis 1960) and their tensor product extensions. Other domains besides subsets of $\mathbb{R}^d$ have also received their share of attention. For instance, positive-weight rules on the sphere are constructed in Mhaskar et al. (2001) and interesting results connecting fill distance and positivity of the weights of quadrature rules on compact Riemannian manifolds appear in Breger et al. (2018). It is also known that, in some typical function classes, such as Sobolev spaces, optimal rates of convergence can be achieved by considering only positive-weight quadrature rules; see for instance Novak (1999, Section 1) and references therein. Therefore, if one can find conditions under which Bayesian quadrature weights are positive, then these conditions may be used as guidelines in construction of numerically stable Bayesian quadrature rules.

This article reviews existing, and derives new, results on properties of the Bayesian quadrature weights, focusing in particular on their *positivity* and *magnitude*. One of our principal aims is to stimulate new research on quadrature weights in the context of probabilistic numerics. While convergence rates of Bayesian quadrature rules have been studied extensively in recent years (Briol et al. 2019; Kanagawa et al. 2016, 2019), analysis of the weights themselves has not attracted much attention. On the other hand, the earliest work by Larkin (1970), Richter-Dyn (1971a) and Barrar and Loeb (1976) [see Oettershagen (2017) for a recent review] done in the 1970s on kernel-based quadrature already revealed certain interesting properties of the Bayesian quadrature weights. These results seem not well-known in the statistics and machine learning community. Moreover, there are some useful results from the literature on scattered data approximation (De Marchi and Schaback 2010), which can be used to analyse the properties of Bayesian quadrature weights. The basics of Bayesian quadrature are reviewed in Sect. 2 while the main contents, including simulation results, of the article are presented in Sects. 3 and 4 .

In Sect. 3, we present results concerning positivity of the Bayesian quadrature weights. We discuss results on the number of the weights that must be positive, focusing on the univariate case and *totally positive* kernels (Definition 2). Corollary 1, the main result of this section, states that all the weights are positive if the design points are locally optimal. A practically relevant consequence of this result is that it may imply that the weights are positive if the design points are obtained by gradient descent, which is guaranteed to provide locally optimal points [see e.g. Lee et al. (2016)].

Section 4 focuses on results on the magnitudes of the weights. More specifically, we discuss the behaviour of the sum of absolute weights, $\sum_{i=1}^n \left| w_{X,i}^{\mathrm{BQ}} \right|$, that strongly affects stability and robustness of Bayesian quadrature. If this quantity is small, the quadrature rule is robust against misspecification of the Gaussian process prior (Kanagawa et al. 2019) and errors in integrand evaluations (Förster 1993) and kernel means (Sommariva and Vianello 2006a, pp. 298–300). This quantity is also related to the numerical stability of the quadrature rule. Using a result on stability of kernel interpolants by De Marchi and Schaback (2010), we derive an upper bound on the sum of absolute weights for some typical cases where the Gaussian process has finite degree of smoothness and the RKHS induced by the covariance kernel is norm-equivalent to a Sobolev space.

## 2 Bayesian quadrature

This section defines a Bayesian quadrature rule as the integral of the posterior of Gaussian process used to model the integrand. We also discuss the equivalent characterisation of this quadrature rule as the worst-case optimal integration rule in the RKHS $\mathcal{H}(k)$ induced by the covariance kernel $k$ of the Gaussian process.

### 2.1 Basics of Bayesian quadrature

In standard Bayesian quadrature (O'Hagan 1991; Minka 2000; Briol et al. 2019), the deterministic integrand $f : \Omega \to \mathbb{R}$ is modelled as a Gaussian process. The integrand is

assigned a zero-mean Gaussian process prior $f_{\text{GP}} \sim \mathcal{GP}(0, k)$ with a positive-definite covariance kernel $k$. This is to say that for any $n \in \mathbb{N}$ and any distinct points $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \Omega$ we have $(f_{\text{GP}}(\boldsymbol{x}_1), \ldots, f_{\text{GP}}(\boldsymbol{x}_n)) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{K}_X)$, with

$$[\boldsymbol{K}_X]_{ij} := \mathrm{Cov}\big[f_{\text{GP}}(\boldsymbol{x}_j),\, f_{\text{GP}}(\boldsymbol{x}_i)\big] = k(\boldsymbol{x}_j, \boldsymbol{x}_i)$$

the $n \times n$ positive-definite (and hence invertible) *kernel matrix*. Conditioning on the *data* $\mathcal{D} = \{(\boldsymbol{x}_i,\, f(\boldsymbol{x}_i))\}_{i=1}^n$, consisting of evaluations $\boldsymbol{f}_X := (f(\boldsymbol{x}_i))_{i=1}^n \in \mathbb{R}^n$ of $f$ at points $X$, yields a Gaussian posterior process with the mean

$$\begin{aligned} \mu_{X,f}(\boldsymbol{x}) &:= \mathbb{E}\big[f_{\text{GP}}(\boldsymbol{x}) \mid \mathcal{D}\big] \\ &= \boldsymbol{k}_X(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{f}_X \end{aligned} \tag{1}$$

and covariance

$$\begin{aligned} \sigma_X^2(\boldsymbol{x}, \boldsymbol{x}') &:= \mathrm{Cov}\big[f_{\text{GP}}(\boldsymbol{x}),\, f_{\text{GP}}(\boldsymbol{x}') \mid \mathcal{D}\big] \\ &= k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_X(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x}'), \end{aligned}$$

where the $n$-vector $\boldsymbol{k}_X(\boldsymbol{x})$ has the elements $[\boldsymbol{k}_X(\boldsymbol{x})]_i = k(\boldsymbol{x}, \boldsymbol{x}_i)$. Note that the posterior covariance only depends on the points, not on the integrand, and that the posterior mean *interpolates* the data (i.e. $\mu_{X,f}(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$). Accordingly, the posterior mean often goes by the name *kernel interpolant* or, if the kernel is isotropic, *radial basis function interpolant*.

Due to the linearity of the integration operator, the posterior of the integral becomes a Gaussian distribution $I(f_{\text{GP}}) \mid \mathcal{D} \sim \mathcal{N}(I_X^{\text{BQ}}(f), \mathbb{V}_X^{\text{BQ}})$ with the mean and variance

$$\begin{aligned} I_X^{\text{BQ}}(f) &:= \mathbb{E}\big[I_\nu(f_{\text{GP}}) \mid \mathcal{D}\big] \\ &= \int_\Omega \mathbb{E}\big[f_{\text{GP}}(\boldsymbol{x}) \mid \mathcal{D}\big] \mathrm{d}\nu(\boldsymbol{x}) \\ &= \boldsymbol{k}_{\nu,X}^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{f}_X, \\ \mathbb{V}_X^{\text{BQ}} &:= \mathrm{Var}[I(f_{\text{GP}}) \mid \mathcal{D}] \\ &= \int_\Omega \int_\Omega \mathrm{Cov}\big[f_{\text{GP}}(\boldsymbol{x}), f_{\text{GP}}(\boldsymbol{x}') \mid \mathcal{D}\big] \mathrm{d}\nu(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x}') \\ &= I_\nu(k_\nu) - \boldsymbol{k}_{\nu,X}^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X}, \end{aligned} \tag{2}$$

where $k_\nu(\boldsymbol{x}) := \int_\Omega k(\cdot, \boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x})$ is the *kernel mean* (Smola et al. 2007), $\boldsymbol{k}_{\nu,X} \in \mathbb{R}^n$ with $[\boldsymbol{k}_{\nu,X}]_i = k_\nu(\boldsymbol{x}_i)$ and

$$I_\nu(k_\nu) = \int_\Omega k_\nu(\boldsymbol{x}) \mathrm{d}\nu(x) = \int_\Omega \int_\Omega k(\boldsymbol{x}, \boldsymbol{x}') \mathrm{d}\nu(\boldsymbol{x}') \mathrm{d}\nu(\boldsymbol{x}).$$

The integral mean $I_X^{\text{BQ}}(f)$ is used to approximate the true intractable integral $I_\nu(f)$ while the variance $\mathbb{V}_X^{\text{BQ}}$ is supposed to quantify epistemic uncertainty, due to partial information being used (i.e. a finite number of function evaluations) inherent to this approximation.

The integral mean $I_X^{\text{BQ}}(f)$ indeed takes the form a quadrature rule, a weighted sum of function evaluations:

$$I_X^{\text{BQ}}(f) = (\boldsymbol{w}_X^{\text{BQ}})^{\mathsf{T}} \boldsymbol{f}_X = \sum_{i=1}^n w_{X,i}^{\text{BQ}} f(\boldsymbol{x}_i),$$

where $w_{X,1}^{\text{BQ}}, \ldots, w_{X,n}^{\text{BQ}}$ are the *Bayesian quadrature weights* given by

$$\boldsymbol{w}_X^{\text{BQ}} := (w_{X,i}^{\text{BQ}})_{i=1}^n := \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X} \in \mathbb{R}^n. \tag{3}$$

The purpose of this article is to analyse the properties of these weights.

A particular property of a Bayesian quadrature rule is that the $n$ kernel translates $k_{\boldsymbol{x}_i} := k(\cdot, \boldsymbol{x}_i)$ are integrated exactly:

$$I_X^{\text{BQ}}(k_{\boldsymbol{x}_i}) = I_\nu(k_{\boldsymbol{x}_i}) = k_\nu(\boldsymbol{x}_i) \;\text{ for each }\; i = 1, \ldots, n, \tag{4}$$

which is derived from the fact that the $j$th equation of the linear system $\boldsymbol{K}_X \boldsymbol{w}_X^{\text{BQ}} = \boldsymbol{k}_{\nu,X}$ defining the weights is

$$\sum_{i=1}^n k(\boldsymbol{x}_j, \boldsymbol{x}_i) w_{X,i}^{\text{BQ}} = k_\nu(\boldsymbol{x}_j).$$

The left-hand side is precisely $I_X^{\text{BQ}}(k_{\boldsymbol{x}_j})$ while on the right-hand side we have $k_\nu(\boldsymbol{x}_j) = I_\nu(k_{\boldsymbol{x}_j})$. Note also that the integral variance is the integration error of the kernel mean:

$$\begin{aligned} \mathbb{V}_X^{\text{BQ}} &= I_\nu(k_\nu) - \boldsymbol{k}_{\nu,X}^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X} \\ &= I_\nu(k_\nu) - (\boldsymbol{w}_X^{\text{BQ}})^{\mathsf{T}} \boldsymbol{k}_{\nu,X} \\ &= I_\nu(k_\nu) - I_X^{\text{BQ}}(k_\nu). \end{aligned}$$

Occasionally, it is instructive to interpret the weights as integrals of the *Lagrange cardinal functions* $\boldsymbol{u}_X = (u_{X,i})_{i=1}^n$ [see e.g. Wendland (2005, Chapter 11)]. These functions are defined as $\boldsymbol{u}_X(\boldsymbol{x}) = \boldsymbol{K}_X^{-1} \boldsymbol{k}_X(\boldsymbol{x})$, from which it follows that

$$\mu_{X,f}(\boldsymbol{x}) = \boldsymbol{u}_X(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{f}_X = \sum_{i=1}^n f(\boldsymbol{x}_i) u_{X,i}(\boldsymbol{x}). \tag{5}$$

Consequently, the cardinality property

$$u_{X,i}(\boldsymbol{x}_j) = \delta_{ij} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

is satisfied, as can be verified by considering the interpolant $\mu_{X,g_i}$ to any function $g_i$ such that $g_i(\boldsymbol{x}_j) = \delta_{ij}$. Since the integral mean is merely the integral of $\mu_{X,f}$, we have from (5) that

$$I_X^{\text{BQ}}(f) = \sum_{i=1}^n f(\boldsymbol{x}_i) I_\nu(u_{X,i}).$$

That is, the $i$th Bayesian quadrature weight is the integral of the $i$th Lagrange cardinal function: $w_{X,i}^{\mathrm{BQ}} = I_\nu(u_{X,i})$.

### 2.2 Reproducing kernel Hilbert spaces

An alternative interpretation of Bayesian quadrature weights is that they are, for the given points, the worst-case optimal weights in the reproducing kernel Hilbert space $\mathcal{H}(k)$ induced by the covariance kernel $k$. The material of this section is contained in, for example, Briol et al. (2019, Section 2), Oettershagen (2017, Section 3.2) and Karvonen and Särkkä (2018a, Section 2). For a comprehensive introduction to RKHSs, see the monograph of Berlinet and Thomas-Agnan (2011).

The RKHS induced by $k$ is the unique Hilbert space of functions characterised by (i) the *reproducing property* $\langle k_{\boldsymbol{x}}, f\rangle_{\mathcal{H}(k)} = f(\boldsymbol{x})$ for every $f \in \mathcal{H}(k)$ and $\boldsymbol{x} \in \Omega$ and (ii) the fact that $k_{\boldsymbol{x}} \in \mathcal{H}(k)$ for every $\boldsymbol{x} \in \Omega$. The *worst-case error* in $\mathcal{H}(k)$ of a quadrature rule with points $X$ and weights $\boldsymbol{w} \in \mathbb{R}^n$ is

$$e_{\mathcal{H}(k)}(X, \boldsymbol{w})^2 := \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \left| \int_\Omega f \,\mathrm{d}\nu - \sum_{i=1}^n w_i f(\boldsymbol{x}_i) \right| = I_\nu(k_\nu) - 2\boldsymbol{w}^\mathsf{T} \boldsymbol{k}_{\nu,X} + \boldsymbol{w}^\mathsf{T} \boldsymbol{K}_X \boldsymbol{w}.$$

It can be then shown that the Bayesian quadrature weights $\boldsymbol{w}_X^{\mathrm{BQ}}$ are the unique minimiser of the worst-case error among all possible weights for these points:

$$\boldsymbol{w}_X^{\mathrm{BQ}} = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^n} e_{\mathcal{H}(k)}(X, \boldsymbol{w})$$

and

$$\mathbb{V}_X^{\mathrm{BQ}} = e_{\mathcal{H}(k)}(X, \boldsymbol{w}_X^{\mathrm{BQ}})^2. \tag{6}$$

Furthermore, the worst-case error can be written as the RKHS error in approximating the integration representer $k_\nu$ that satisfies $I_\nu(f) = \langle k_\nu, f\rangle_{\mathcal{H}(k)}$ for all $f \in \mathcal{H}(k)$:

$$e_{\mathcal{H}(k)}(X, \boldsymbol{w}_X^{\mathrm{BQ}}) = \left\| k_\nu - k_Q \right\|_{\mathcal{H}(k)}, \quad k_Q := \sum_{i=1}^n w_{X,i}^{\mathrm{BQ}} k_{\boldsymbol{x}_i}.$$

From this representation and the Cauchy–Schwarz inequality it follows that

$$\begin{aligned} \left| I_\nu(f) - I_X^{\mathrm{BQ}}(f) \right| &= \left| \langle k_\nu - k_Q, f\rangle_{\mathcal{H}(k)} \right| \\ &\leq \|f\|_{\mathcal{H}(k)} \left\| k_\nu - k_Q \right\|_{\mathcal{H}(k)} \\ &= \|f\|_{\mathcal{H}(k)} \, e_{\mathcal{H}(k)}(X, \boldsymbol{w}_X^{\mathrm{BQ}}). \end{aligned}$$

For analysis of convergence of Bayesian quadrature rules as $n \to \infty$, it is therefore sufficient to analyse how the worst-case error (i.e. integral variance) behaves—as long as the integrand indeed lives in $\mathcal{H}(k)$. Convergence will be discussed in Sect. 4.

## 3 Positivity

This section reviews existing results on the positivity of the weights of Bayesian quadrature that can be derived in one dimension when the covariance kernel is *totally positive*. This assumption, given in Definition 2, is stronger than positive-definiteness but is satisfied by, for example, the Gaussian kernel. For most of the section, we assume that $d = 1$ and $\Omega = [a, b]$ for $a < b$. Furthermore, the measure $\nu$ is typically assumed to admit a density function with respect to the Lebesgue measure,[1] an assumption that implies $I_\nu(f) > 0$ if $f(x) > 0$ for almost every $x \in \Omega$.

Positivity of the weights was actively investigated during the 1970s (Richter 1970; Richter-Dyn 1971a, b; Barrar et al. 1974; Barrar and Loeb 1976), and these results have been recently refined and collected by Oettershagen (2017, Section 4). To simplify presentation, some of the results in this section are given in a slightly less general form than possible. Two of the most important results are

- **Theorem** 1: At least one half of the weights of *any* Bayesian quadrature rule are positive.
- **Corollary** 1: All the weights are positive when the points are selected so that the integral posterior variance in (2) is locally minimised in the sense that each of its $n$ partial derivatives with respect to the integration points vanishes (Definition 3).

The latter of these results is particularly interesting since (i) it implies that points selected using a gradient descent algorithm may have positive weights and (ii) the resulting Bayesian quadrature rule is a positive linear functional and hence potentially well-suited for integration of functions that are known to be positive—a problem for which a number of transformation-based methods have been developed recently (Osborne et al. 2012; Gunter et al. 2014; Chai and Garnett 2018).

As no multivariate extension of the theory used to prove the aforementioned results appears to have been developed, we do not provide any general theoretical results on the weights in higher dimensions. However, some special cases based on, for example, tensor products are discussed in Sects. 3.7 and 3.9 and two numerical examples are used to provide some evidence for the conjectures that multivariate versions of Theorem 1 and Corollary 1 hold.

[1] This can be usually relaxed to $I_\nu$ being a positive linear functional: $I_\nu(f) > 0$ whenever $f$ is almost everywhere positive.

It will turn out that optimal Bayesian quadrature rules are analogous to classical Gaussian quadrature rules in the sense that, in addition to being exact for kernel interpolants [recall (4)], they also exactly integrate Hermite interpolants (see Sect. 3.2.2). We thus begin by reviewing the argument used to establish positivity of the Gaussian quadrature weights.

### 3.1 Gaussian quadrature

Under the assumption that $\nu$ admits a density[2] there exist unique weights $w_1, \ldots, w_n$ and points $x_1, \ldots, x_n \in [a, b]$ such that

$$\sum_{i=1}^{n} w_i P(x_i) = \int_a^b P(x) \mathrm{d}\nu(x) \tag{7}$$

for every polynomial $P$ of degree at most $2n-1$ (Gautschi 2004, Chapter 1). This quadrature rule is known as a *Gaussian quadrature rule* (for the measure $\nu$). One can show the positivity of the weights of a Gaussian rule as follows.

**Proposition 1** *Assume that $\nu$ admits a Lebesgue density. Then the weights $w_1, \ldots, w_n$ of the Gaussian quadrature (7) are positive.*

***Proof*** For each $i = 1, \ldots, n$ there exists a unique polynomial $L_i$ of degree $n-1$ such that $L_i(x_j) = \delta_{ij}$. This property is shared by the function $G_i := L_i^2 \geq 0$ that, being of degree $2n-2$, is also integrated exactly by the Gaussian rule. Because $G_i$ is almost everywhere positive, it follows from the assumption that $\nu$ admits a density that

$$0 < \int_a^b G_i(x) \mathrm{d}\nu(x) = \sum_{i=1}^{n} w_j G_i(x_j) = w_i.$$

The positivity of the weights is thus concluded. □

This proof may appear to be based on the closedness of the set of polynomials under exponentiation. Closer analysis reveals a structure that can be later generalised.

To describe this, recall that one of the basic properties of polynomials is that a polynomial $P$ of degree $n$ can have at most $n$ zeroes, when counting multiplicities [for some properties of polynomials and interpolation with them, see e.g. Atkinson (1989, Chapter 3)]. This is to say that, if for some points $x_1, \ldots, x_m$ it holds that

$$P^{(j_i)}(x_i) := \frac{\mathrm{d}^{j_i}}{\mathrm{d}x^{j_i}} P(x)\Big|_{x=x_i} = 0$$

for $j_i = 0, \ldots, q_i - 1$, with $q_i$ being the *multiplicity* of the zero $x_i$ of $P$, then $\sum_{i=1}^{m} q_i \leq n$. This fact on zeroes of polynomials can be used to supply a proof of positivity of the Gaussian quadrature weights that does not explicitly use of the fact that square of a function is non-negative. By the chain rule, the derivative of $G_i$ vanishes at each $x_j$ such that $j \neq i$. That is, $G_i$ has a double zero at each of these $n-1$ points (i.e. $G_i(x_j) = 0$ and $G_i^{(1)}(x_j) = 0$), for the total of $2n-2$ zeroes. Being a polynomial of degree $2n-2$, $G_i$ cannot have any other zeroes besides these. Since all the zeroes of $G_i$ are double, it cannot hence have any sign changes. This is because, in general, a function $g$ that satisfies $g(x) = g^{(1)}(x) = 0$ but $g^{(2)}(x) \neq 0$ at a point $x$ cannot change its sign at $x$, since its derivative changes sign at the point. From $G_i(x_i) = 1 > 0$ it then follows that $G_i$ is almost everywhere positive.

### 3.2 Chebyshev systems and generalised Gaussian quadrature

The argument presented above works almost as such when the polynomials are replaced with generalised polynomials and the Gaussian quadrature rule with a generalised Gaussian quadrature rule. Much of the following material is covered by the introductory chapters of the monograph by Karlin and Studden (1966). In the following $C^m([a, b])$ stands for the set of functions that are $m$ times continuously differentiable on the open interval $(a, b)$.

**Definition 1** (*Chebyshev system*) A collection of functions $\{\phi_i\}_{i=1}^m \subset C^{m-1}([a, b])$ constitutes an (extended) *Chebyshev system* if any non-trivial linear combination of the functions, called a *generalised polynomial*, has at most $m-1$ zeroes, counting multiplicities.

***Remark 1*** Some of the results we later present, such as Proposition 3, are valid even when a less restrictive definition, that does not require differentiability of $\phi_i$, of a Chebyshev system is used. Of course, in this case the definition is not given in terms of multiple zeroes. The above definition is used here to simplify presentation. The simplest relaxation is to require that $\{\phi_i\}_{i=1}^m$ are merely continuous and that no linear combination can vanish at more than $m-1$ points.

By selecting $\phi_i(x) = x^{i-1}$, we see that polynomials are an example of a Chebyshev system. Perhaps the simplest example of a non-trivial Chebyshev system is given by the following example.

***Example 1*** Let $\phi_i(x) = \mathrm{e}^x x^{i-1}$ for $i = 1, \ldots, m$. Then $\{\phi_i\}_{i=1}^m$ constitute a Chebyshev system. To verify this, observe that any linear combination $\phi$ of $\phi_1, \ldots, \phi_m$ is of the form $\phi(x) = \mathrm{e}^x P(x)$ for a polynomial $P$ of degree at most $m-1$ and that the $j$th derivative of this function takes the form

$$\phi^{(j)}(x) = \mathrm{e}^x \left[ P(x) + c_1 P^{(1)}(x) + \cdots + c_j P^{(j)}(x) \right] \tag{8}$$

[2] This can be generalised to the cumulative distribution function having infinitely many points of increase.

for certain integer coefficients $c_1, \ldots, c_j$. We observe that $\phi(x_0) = 0$ for a point $x_0$ if and only if $P(x_0) = 0$. If also $\phi^{(1)}(x_0) = 0$, then it follows from (8) that $P^{(1)}(x_0) = 0$, and, generally, that $\phi^{(i)}(x_0) = 0$ for $i = 0, \ldots, j$ if and only if $P^{(i)}(x_0) = 0$. That is, the zeroes of $\phi$ are precisely those of $P$ and, consequently, the functions $\phi_i$ constitute a Chebyshev system.

#### 3.2.1 Interpolation using a Chebyshev system

A crucial property of generalised polynomials is that unique interpolants can be constructed using them, as we next show. For any Chebyshev system $\{\phi_i\}_{i=1}^n$ and a set of distinct points $X = \{x_1, \ldots, x_n\} \subset [a, b]$, we know that there cannot exist $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \neq \mathbf{0}$ such that

$$\sum_{i=1}^{n} \alpha_i \phi_i(x_j) = 0 \quad \text{for every} \quad j = 1, \ldots, n$$

since $\alpha_1\phi_1 + \cdots + \alpha_n\phi_n$ can have at most $n - 1$ zeroes. Equivalently, the only solution $\boldsymbol{\beta} \in \mathbb{R}^n$ to the linear system $\boldsymbol{V}_X^\mathsf{T}\boldsymbol{\beta} = \mathbf{0}$ defined by the $n \times n$ matrix $[\boldsymbol{V}_X]_{ij} = \phi_i(x_j)$ is $\boldsymbol{\beta} = \mathbf{0}$. That is, $\boldsymbol{V}_X$ is invertible.

For any data $\{(x_i, f(x_i))\}_{i=1}^n$, the above fact guarantees the existence and uniqueness of an interpolant $s_{X,f}$ such that (i) $s_{X,f}$ is in $\text{span}\{\phi_1, \ldots, \phi_n\}$ and (ii) $s_{X,f}(x_j) = f(x_j)$ for each $j = 1, \ldots, n$. These two requirements imply that

$$s_{X,f}(x_j) = \sum_{i=1}^{n} \alpha_i \phi_i(x_j) = f(x_j)$$

for and some $\boldsymbol{\alpha} \in \mathbb{R}^n$ and every $j = 1, \ldots, n$. In matrix form, these $n$ equations are equivalent to $\boldsymbol{V}_X^\mathsf{T}\boldsymbol{\alpha} = \boldsymbol{f}_X$. Hence $\boldsymbol{\alpha} = \boldsymbol{V}_X^{-\mathsf{T}}\boldsymbol{f}_X$ and the interpolant is

$$s_{X,f}(x) = \boldsymbol{\phi}(x)^\mathsf{T}\boldsymbol{\alpha} = \boldsymbol{\phi}(x)^\mathsf{T}\boldsymbol{V}_X^{-\mathsf{T}}\boldsymbol{f}_X \tag{9}$$

for $[\boldsymbol{\phi}(x)]_i = \phi_i(x)$ an $n$-vector.

#### 3.2.2 Hermite interpolants

A *Hermite interpolant* $s_{X,\boldsymbol{q},f}$ is based on data containing also derivative values [Atkinson (1989, Section 3.6) for polynomial and Fasshauer (2007, Chapter 36) for kernel-based Hermite interpolation]. In this setting, the point set $X$ contains $m$ points and $\boldsymbol{q} \in \mathbb{N}_0^m$ is a vector of multiplicities such that $\sum_{i=1}^m q_i = n$. The data to be interpolated are

$$\{(x_i, f^{(j_i)}(x_i)) : i = 1, \ldots, m \text{ and } j_i = 0, \ldots, q_i - 1\}.$$

That is, the interpolant is to satisfy

$$s_{X,\boldsymbol{q},f}^{(j_i)}(x_i) = f^{(j_i)}(x_i)$$

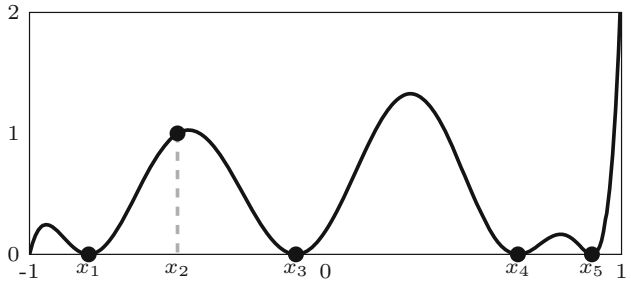


**Fig. 1** Example of a Hermite interpolant $F_i$ used in proving positivity of the weights of generalised Gaussian quadrature rule. This figure uses the Chebyshev system formed by $\phi_i(x) = \mathrm{e}^x x^{i-1}$

for each $i = 1, \ldots, m$ and $j_i = 0, \ldots, q_i - 1$. Note that the interpolant $s_{X,f}$ is a Hermite interpolant with $m = n$ and $q_1 = \cdots = q_n = 1$. If the interpolant is to lie in $\text{span}\{\phi_1, \ldots, \phi_n\}$, we must have, for some $\alpha_1, \ldots, \alpha_n$,

$$s_{X,\boldsymbol{q},f}^{(j_i)}(x_i) = \sum_{l=1}^{n} \alpha_l \phi_l^{(j_i)}(x_i) = f^{(j_i)}(x_i).$$

Again, these $n$ equations define a linear system that is invertible because $\{\phi_i\}_{i=1}^n$ constitute a Chebyshev system. The Hermite interpolant can be written in the form (9) with $\boldsymbol{V}_X$ replaced with a version involving also derivatives of $\phi_i$ [see e.g. Oettershagen (2017, Section 2.3.1)].

#### 3.2.3 Generalised Gaussian quadrature

A *generalised Gaussian quadrature rule* is a quadrature rule that uses $n$ points to integrate exactly all functions in the span of $\{\phi_i\}_{i=1}^{2n}$ constituting a Chebyshev system:

$$\sum_{i=1}^{n} w_i \phi(x_i) = \int_a^b \phi(x) \mathrm{d}\nu(x) \tag{10}$$

for every $\phi \in \text{span}\{\phi_1, \ldots, \phi_{2n}\}$. The existence and uniqueness of the points and weights is guaranteed under fairly general assumptions (Barrow 1978). We prove positivity of the weights by constructing a function $F_i \in \text{span}\{\phi_1, \ldots, \phi_{2n}\}$ analogous to $G_i$ in Sect. 3.1.

**Proposition 2** *Assume that $\nu$ admits a Lebesgue density. Then the weights $w_1, \ldots, w_n$ of the generalised Gaussian quadrature rule (10) are positive.*

***Proof*** Let $F_i$ be the Hermite interpolant to the data

$$f(a) = 0, \quad f(x_i) = 1, \quad f(x_j) = f^{(1)}(x_j) = 0 \text{ for } j \neq i.$$

An example is depicted in Fig. 1. As there are $2n$ data points, $F_i$ indeed exists since $\{\phi_i\}_{i=1}^{2n}$ are a Chebyshev system. Moreover, $F_i$ has $2n - 1$ zeroes. Because all its zeroes occurring

on $(a, b)$ are double, $F_i$ cannot have sign changes. Since $F_i(x_i) = 1 > 0$, we conclude $F_i$ is almost everywhere positive. Consequently, $w_i = I_\nu(F_i) > 0$. □

Next we turn our attention to kernels whose translates and their derivatives constitute Chebyshev systems.

### 3.3 Totally positive kernels

We are now ready to begin considering kernels and Bayesian quadrature. A concept related to Chebyshev systems is that of totally positive kernels whose theory is covered by the monograph of Karlin (1968). For a sufficiently differentiable kernel, define the derivatives

$$k_y^{(j)}(x) := k^{(j)}(x, y) := \frac{\partial^j}{\partial z^j} k(x, z)\Big|_{z=y}. \tag{11}$$

If the derivative

$$\frac{\partial^{2j}}{\partial x^j \partial y^j} k(x, y)$$

exists and is continuous for every $j \leq m$, the kernel is said to be $m$ times continuously differentiable, which we denote by writing $k \in C^m([a, b]^2)$. In this case, $f \in C^m([a, b])$ if $f \in \mathcal{H}(k)$ and the kernel derivatives (11) act as representers for differentiation (i.e. $\langle f, k^{(j)}(\cdot, x)\rangle_{\mathcal{H}(k)} = f^{(j)}(x)$ for $f \in \mathcal{H}(k)$ and $j \leq m$); see Corollary 4.36 and its proof in Steinwart and Christmann (2008)

**Definition 2** (*Totally positive kernel*) A kernel $k \in C^\infty([a, b]^2)$ is (extended) *totally positive of order* $q \in \mathbb{N}$ if the collection

$$\left\{k_{x_i}^{(j_i)} : i = 1, \ldots, m \text{ and } j_i = 0, \ldots, q_i - 1\right\}$$

constitutes a Chebyshev system for any $m \in \mathbb{N}$, any distinct $x_1, \ldots, x_m \in \Omega$ and any multiplicities $q_1, \ldots, q_m \leq q$ of these points.

The class of totally positive kernels is smaller than that of positive-definite kernels. For the simplest case of $q = 1$ and $m = n$ the total positivity condition is that the kernel translates $k_{x_1}, \ldots, k_{x_n}$ constitute a Chebyshev system. This implies that the $n \times n$ matrix $[\boldsymbol{K}_{Y,X}] := k(y_j, x_i)$, which is just the matrix $\boldsymbol{V}_Y$ considered in Sect. 3.2 for the Chebyshev system $\phi_i = k_{x_i}$, is invertible for any $Y = \{y_1, \ldots, y_n\} \subset [a, b]$. Positive-definiteness of $k$ only guarantees that $\boldsymbol{K}_{Y,X}$ is invertible when $Y = X$.

Basic examples of totally positive kernels are the Gaussian kernel

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \tag{12}$$

with length-scale $\ell > 0$ and the Hardy kernel $k(x, x') = r^2/(r^2 - xx')$ for $r > 0$. Both of these kernels are totally positive of any order. There is also a convenient result that guarantees total positivity (Burbea 1976, Proposition 3): $k$ is totally positive if there are positive constants $a_m$ and a positive increasing function $v \in C^\infty([a, b])$ such that

$$k(x, x') = \sum_{m=0}^{\infty} a_m v(x)^m v(x')^m$$

for all $x, x' \in \Omega$. More examples are collected in Karlin (1968) and Burbea (1976).

### 3.4 General result on weights

The following special case of the theory developed in Karlin and Studden (1966, Chapter 2) appears in, for instance, Richter-Dyn (1971a, Lemma 2). Its proof is a generalisation of the proof for the case $m = 2n$ that is discussed in Sect. 3.2.

**Proposition 3** *Suppose that* $\{\phi_i\}_{i=1}^m \subset C^{m-1}([a, b])$ *constitute a Chebyshev system, that* $\nu$ *admits a Lebesgue density and that* $Q(f) := \sum_{i=1}^n w_i f(x_i)$ *for* $x_1, \ldots, x_m \in \Omega$ *is a quadrature rule such that* $Q(\phi_i) = I_\nu(\phi_i)$ *for each* $i = 1, \ldots, m$. *Then at least* $\lfloor (m + 1)/2 \rfloor$ *of the weights* $w_1, \ldots, w_n$ *are positive.*

An immediate consequence of this proposition is that a Bayesian quadrature rule based on a totally positive kernel has at least one half of its weights positive.

**Theorem 1** *Suppose that the kernel* $k \in C^\infty([a, b]^2)$ *is totally positive of order 1. Then, for any points, at least* $\lfloor (n + 1)/2 \rfloor$ *of the Bayesian quadrature weights* $w_{X,1}^{\mathrm{BQ}}, \ldots, w_{X,n}^{\mathrm{BQ}}$ *are positive.*

***Proof*** Since the kernel is totally positive of order 1, the translates $\{k_{x_i}\}_{i=1}^n$ constitute a Chebyshev system. The exactness condition (4) holds for each of these functions. The claim follows by setting $m = n$ in Proposition 3. □

### 3.5 Weights for locally optimal points

Recall the definition of the Bayesian quadrature variance:

$$\mathbb{V}_X^{\mathrm{BQ}} = I_\nu(k_\nu) - \sum_{i=1}^n w_{X,i}^{\mathrm{BQ}} k_\nu(x_i) = I_\nu(k_\nu) - \boldsymbol{k}_{\nu,X}^{\mathsf{T}} \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X}.$$

The variance can be considered a function $X \mapsto \mathbb{V}_X^{\mathrm{BQ}}$ defined on the simplex

$$\mathcal{S}^n := \{\boldsymbol{z} \in [a, b]^n : a < z_1 < \cdots < z_n < b\} \subset [a, b]^n.$$

We introduce the following definition of locally optimal points. For this purpose, define the function

$$E(Z) := \mathbb{V}_Z^{\mathrm{BQ}} \quad \text{for} \quad Z = (z_1, \ldots, z_n) \in \mathcal{S}^n$$

and its partial derivatives

$$E_j(X) := \frac{\partial}{\partial z_j} E(Z)\Big|_{Z=X}.$$

**Definition 3** Let $m \leq n$. A Bayesian quadrature rule with points $X \subset [a, b]$ is *locally m-optimal* if $X \in \mathcal{S}^n$ and there is an index set $\mathcal{I}_m^* \subset \{1, \ldots, n\}$ of $m$ indices such that

$$E_j(X) = \frac{\partial}{\partial z_j} \mathbb{V}_Z^{\mathrm{BQ}}\Big|_{Z=X} = 0 \ \text{ for every } \ j \in \mathcal{I}_m^*. \tag{13}$$

A locally $n$-optimal rule is called *locally optimal*. The point set of a locally $m$-optimal Bayesian quadrature rule is also called locally $m$-optimal.

When the kernel is totally positive of any order, it has been shown that any local minimiser of $\mathbb{V}_X^{\mathrm{BQ}}$ is locally optimal in the sense of above definition. That is, no point in a point set that locally minimises the variance can be located on the boundary of the integration interval nor can any two points in the set coalesce.[3] These results, the origins of which can be traced to the 1970s (Barrar et al. 1974; Barrar and Loeb 1976; Bojanov 1979), have been recently collated by Oettershagen (2017, Corollary 5.13).

A locally $m$-optimal Bayesian quadrature rule is, in addition to the kernel translates at $X$, exact for translate derivatives at $x_j$ with $j \in \mathcal{I}_m^*$ [it is worth noting that Bayesian quadrature rules with derivative evaluations have been recently considered in Prüher and Särkkä (2016) and Wu et al. (2018)]. When $m = n$, this is analogous to the interpretation of classical Gaussian quadrature rules as integrated Hermite interpolants (Richter-Dyn 1971b). This result first appeared in Larkin (1970). Its proof is typically based on considering the RKHS representation

$$\mathbb{V}_X^{\mathrm{BQ}} = \left\| k_\nu - \sum_{i=1}^n w_{X,i}^{\mathrm{BQ}} k_{x_i} \right\|_{\mathcal{H}(k)}^2$$

of the variance; see Richter-Dyn (1971a, Section 3) or Oettershagen (2017, Section 5.1.3). We present a mainly linear algebraic proof.

[3] Coalescence is possible because $\mathbb{V}_X^{\mathrm{BQ}}$ is in fact a continuous function of $X$ defined on the whole of $\Omega^n$, not merely on $\mathcal{S}^n$ (Oettershagen 2017, Proposition 5.5). Coalescence of some of the points would result in a quadrature rule that uses also evaluations of derivatives of the integrand.

**Proposition 4** *Let $m \leq n$. Suppose that the $n$-point set $X \in \mathcal{S}^n$ is locally $m$-optimal. If the kernel $k$ is once continuously differentiable, then*

$$\begin{aligned} I_X^{\mathrm{BQ}}(k_x) &= I_\nu(k_x) && \text{for} \quad x \in X, \\ I_X^{\mathrm{BQ}}(k_{x_j}^{(1)}) &= I_\nu(k_{x_j}^{(1)}) \ \text{ or } \ w_{X,j}^{\mathrm{BQ}} = 0 && \text{for} \quad j \in \mathcal{I}_m^*, \end{aligned} \tag{14}$$

*where $k_x^{(1)}$ is the kernel derivative defined in (11).*

***Proof*** By definition of local $m$-optimality, the partial derivatives

$$E_j(X) = \frac{\partial}{\partial z_j} E(Z)\Big|_{Z=X}$$

must vanish for each $j \in \mathcal{I}_m^*$. Let $\partial_i \boldsymbol{g}(X) \in \mathbb{R}^n$ stand for the $i$th partial derivative of a vector-valued function $\boldsymbol{g} \colon \mathbb{R}^n \to \mathbb{R}^n$ evaluated at $X$. From the explicit expression (2) for the variance we compute

$$\begin{aligned} E_j(X) = & -2(\partial_j \boldsymbol{k}_{\nu,X}^{\mathsf{T}}) \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X} \\ & + \boldsymbol{k}_{\nu,X}^{\mathsf{T}} \boldsymbol{K}_X^{-1} (\partial_j \boldsymbol{K}_X) \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X} \\ = & -2(\partial_j \boldsymbol{k}_{\nu,X}^{\mathsf{T}}) \boldsymbol{w}_X^{\mathrm{BQ}} + (\boldsymbol{w}_X^{\mathrm{BQ}})^{\mathsf{T}} (\partial_j \boldsymbol{K}_X) \boldsymbol{w}_X^{\mathrm{BQ}}, \end{aligned}$$

where the inverse matrix derivative formula

$$\frac{\mathrm{d}}{\mathrm{d}x} \boldsymbol{A}(x)^{-1} = -\boldsymbol{A}(x)^{-1} \left[ \frac{\mathrm{d}}{\mathrm{d}x} \boldsymbol{A}(x) \right] \boldsymbol{A}(x)^{-1}$$

and the weight expression $\boldsymbol{w}_X^{\mathrm{BQ}} = \boldsymbol{K}_X^{-1} \boldsymbol{k}_{\nu,X}$ have been used. The two partial derivatives appearing in the equation for $E_j(X)$ can be explicitly computed. First, only the $j$th element of $\boldsymbol{k}_{\nu,X}$ depends on $x_j$. Thus,

$$[\partial_j \boldsymbol{k}_{\nu,X}]_i = \frac{\partial}{\partial x_j} \int_a^b k(x, x_i) \mathrm{d}\nu(x) = I_\nu(k_{x_j}^{(1)}) \delta_{ij}.$$

Secondly, only the $j$th row and column of $\boldsymbol{K}_X$ have dependency on $x_j$. For $l \neq j$ we have

$$[\partial_j \boldsymbol{K}_X]_{lj} = [\partial_j \boldsymbol{K}_X]_{jl} = \frac{\partial}{\partial z} k(x_l, z)\Big|_{z=x_j} = k_{x_j}^{(1)}(x_l),$$

where the first equality is consequence of symmetry of the kernel. The diagonal element is a total derivative:

$$\begin{aligned} [\partial_j \boldsymbol{K}_X]_{jj} = \frac{\mathrm{d}}{\mathrm{d}x_j} k(x_j, x_j) &= 2 \frac{\partial}{\partial z} k(x_j, z)\Big|_{z=x_j} \\ &= 2 k_{x_j}^{(1)}(x_j). \end{aligned}$$

Therefore, $\partial_j \boldsymbol{K}_X$ is a zero matrix except for the $j$th row and column that are

$$\left[ k_{x_j}^{(1)}(x_1) \cdots k_{x_j}^{(1)}(x_{j-1}) \ 2k_{x_j}^{(1)}(x_j) \ k_{x_j}^{(1)}(x_{j+1}) \cdots k_{x_j}^{(1)}(x_n) \right]$$

and its transpose, respectively. Hence

$$E_j(X) = -2w^{\text{BQ}}_{X,j} I_\nu(k^{(1)}_{x_j}) + \sum_{i=1}^{n}\sum_{l=1}^{n} w^{\text{BQ}}_{X,i} w^{\text{BQ}}_{X,l} [\partial_j \boldsymbol{K}_X]_{il}$$
$$= -2w^{\text{BQ}}_{X,j} I_\nu(k^{(1)}_{x_j}) + 2w^{\text{BQ}}_{X,j} \sum_{i=1}^{n} w^{\text{BQ}}_{X,i} k^{(1)}_{x_j}(x_i)$$
$$= -2w^{\text{BQ}}_{X,j} \big[ I_\nu(k^{(1)}_{x_j}) - I^{\text{BQ}}_X(k^{(1)}_{x_j}) \big].$$

If $w^{\text{BQ}}_{X,j} \neq 0$, then $E_j(X) = 0$ so that the form of $E_j$ above implies that $I^{\text{BQ}}_X(k^{(1)}_{x_j}) = I_\nu(k^{(1)}_{x_j})$. This concludes the proof. □

***Remark 2*** Proposition 4 admits an obvious multivariate extension (Gavrilov 1998, теорема 2): when $d > 1$, the $md$ partial derivative representers

$$\frac{\partial}{\partial z_j} k(\cdot, z)\Big|_{z=\boldsymbol{x}_i}$$

for $j = 1, \ldots, d$ and $i \in \mathcal{I}^*_m$ are integrated exactly by a locally $m$-optimal Bayesian quadrature rule, defined by requiring a gradient version of (13). See also Gavrilov (2007). However, there appear to exist no generalisations of Chebyshev systems and Proposition 3 to higher dimensions.

**Theorem 2** *Let $k \in C^\infty([a,b]^2)$ be a totally positive kernel of order 2 and $m \leq n$. Suppose that the point set $X \in \mathcal{S}^n$ is locally $m$-optimal with an index set $\mathcal{I}^*_m \subset \{1, \ldots, n\}$ and that the weights associated with $q \leq m$ indices in $\mathcal{I}^*_m$ are non-zero. Then at least $\lfloor (n + 2m - q + 1)/2 \rfloor$ of the weights are non-negative, and $q$ must satisfy $2m - n \leq q$.*

***Proof*** By (14), the Bayesian quadrature rule in the statement is exact for $n$ kernel translates and $q$ of their derivatives. By the total positivity of the kernel, the collection of these $n+q$ functions constitutes a Chebyshev system. By Proposition 3, at least $\lfloor (n + q + 1)/2 \rfloor$ of the weights are positive. Since the weights associated with $m - q$ indices in $\mathcal{I}^*_m$ are zero, it follows that at least $\lfloor (n+q+1)/2 \rfloor + m - q = \lfloor (n + 2m - q + 1)/2 \rfloor$ of the weights are non-negative. The lower-bound for $q$ follows because $\lfloor (n + 2m - q + 1)/2 \rfloor \leq n$ implies that $n + 2m - q + 1 \leq 2n + 1$. □

The main result of this section follows by setting $m = n$ in the preceding theorem and observing that this implies $q = n$, which means that there can be no zero weights.

**Corollary 1** *If $k \in C^\infty([a,b]^2)$ is totally positive of order 2 and $X \in \mathcal{S}^n$ is locally optimal, then all the Bayesian quadrature weights $w^{\text{BQ}}_{X,1}, \ldots, w^{\text{BQ}}_{X,n}$ are positive.*

***Remark 3*** A key consequence of Corollary 1 is the following: If $w^{\text{BQ}}_{X,1}, \ldots, w^{\text{BQ}}_{X,n}$ contain negative values, then the design points $X$ are *not* locally optimal. In other words, in this case there is still room for improvement by optimising these points using, for example, gradient descent. In this way, the signs of the weights can provide information about the quality of the design point set.

A positive-weight quadrature rule is a positive linear functional (i.e. every positive function is mapped to a positive real). A locally optimal Bayesian quadrature rule may therefore be appropriate for numerical integration of functions that are a priori known to be positive, such as likelihood functions. Theoretical comparison to warped models (Osborne et al. 2012; Gunter et al. 2014; Chai and Garnett 2018) that encode positivity of the integrand by placing the GP prior on, for example, square root of the integrand would be an interesting topic of research.

### 3.6 Greedily selected points

The optimal points discussed in the preceding section cannot be constructed efficiently. See Oettershagen (2017, Section 5.2) for what appears to be the most advanced published algorithm. In practice, points selected by greedy minimisation of the integral variance are often used. This approach is known as *sequential Bayesian quadrature* (Cook and Clayton 1998; Huszár and Duvenaud 2012). Assuming for a moment that $d$ is arbitrary and an $n$-point set $X_n \subset \Omega$ has been already generated, sequential Bayesian quadrature proceeds by selecting a new point $\boldsymbol{x}_{n+1} \in \Omega$ by minimising the integral variance:

$$\boldsymbol{x}_{n+1} = \arg\min_{\boldsymbol{x} \in \Omega} \mathbb{V}^{\text{BQ}}_{X_n \cup \{\boldsymbol{x}\}}.$$

In higher dimensions, there is little that we are able to say about qualitative properties of the resulting quadrature rules. However, when $d = 1$ we can invoke Theorem 2 since $X_n \cup \{x_{n+1}\}$ is locally 1-optimal.

**Proposition 5** *Suppose that $k \in C^\infty([a,b]^2)$ is totally positive of order 2. If $X_n \cup x_{n+1} \in \mathcal{S}^n$, then at least $\lfloor (n+3)/2 \rfloor$ of the weights of a $n+1$ point sequential Bayesian quadrature rule are positive.*

### 3.7 Other kernels and point sets

A number of combinations of kernels and point sets, that are not covered by the theory above, have been shown, either theoretically or experimentally, to yield positive Bayesian quadrature weights:

– The GP posterior mean for the Brownian motion kernel $k(x, x') = \min(x, x')$ on $[0, 1]$ is a piecewise linear interpolant. As this implies that the Lagrange cardinal functions $u_{X,i}$ are non-negative, it follows from the identity $w^{\text{BQ}}_{X,i} = I_\nu(u_{X,i})$ that the weights are positive. See

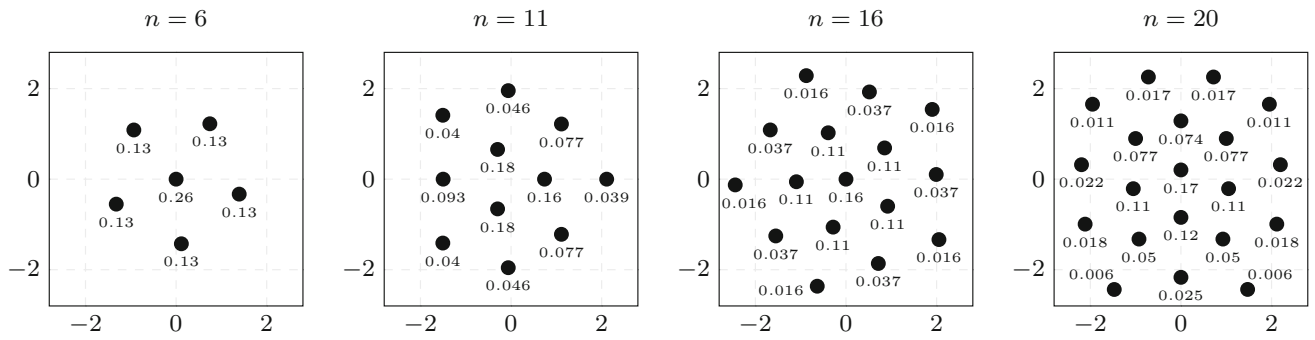


**Fig. 2** Locally optimal Bayesian quadrature point sets for the Gaussian measure and kernel on $\mathbb{R}^2$. The corresponding weights are written in grey. The sums of weights are 0.91 ($n = 6$), 0.978 ($n = 11$), 0.9975 ($n = 16$) and 1.011 ($n = 20$)

Diaconis (1988) and Ritter (2000, Lemma 8 in Section 3.2, Chapter 2) for more discussion.

- Suitably selected priors give rise to Bayesian quadrature rules whose posterior mean coincides with a classical rule, such a Gaussian quadrature (Karvonen and Särkkä 2017; Karvonen et al. 2018b). Analysis of the weights and their positivity naturally reduces to that of the reproduced classical rule.
- There is convincing numerical evidence that the weights are positive if the nodes for the Gaussian kernel and measure on $\mathbb{R}$ are selected by suitable scaling the classical Gauss–Hermite nodes (Karvonen and Särkkä 2019).
- Uniform weighting (i.e. $w_{X,i}^{\mathrm{BQ}} = 1/n$) can be achieved when certain quasi-Monte Carlo point sets and shift-invariant kernels are used (Jagadeeswaran and Hickernell 2019).

### 3.8 Upper bound on the sum of weights

We summarise below a simple yet generic result that has an important consequence on the stability of Bayesian quadrature in Sect. 4.

**Lemma 1** *Let $\Omega \subset \mathbb{R}^d$. If the Bayesian quadrature weights $w_{X,1}^{\mathrm{BQ}}, \ldots, w_{X,n}^{\mathrm{BQ}}$ are non-negative, then we have*

$$\sum_{i=1}^{n} w_{X,i}^{\mathrm{BQ}} \leq \frac{\sup_{\boldsymbol{x} \in \Omega} I_\nu(k_{\boldsymbol{x}})}{\inf_{\boldsymbol{x}, \boldsymbol{x}' \in \Omega} k(\boldsymbol{x}, \boldsymbol{x}')}.$$

***Proof*** The claim immediately follows from the property (4) that $\sum_{i=1}^{n} w_{X,i}^{\mathrm{BQ}} k_{\boldsymbol{x}_j}(\boldsymbol{x}_i) = I_\nu(k_{\boldsymbol{x}_j})$ for each $j = 1, \ldots, n$. □

Combined with Corollary 1, we get a bound on the sum of absolute weights $\sum_{i=1}^{n} |w_{X_n,i}^{\mathrm{BQ}}|$, which is the main topic of discussion in Sect. 4.

**Corollary 2** *Let $\Omega = [a, b] \subset \mathbb{R}$. If $k \in C^\infty([a, b]^2)$ is totally positive of order 2 and design points $X \in \mathcal{S}^n$ are locally optimal, then we have*

$$\sum_{i=1}^{n} |w_{X,i}^{\mathrm{BQ}}| = \sum_{i=1}^{n} w_{X,i}^{\mathrm{BQ}} \leq \frac{\sup_{x \in [a,b]} I_\nu(k_x)}{\inf_{x, x' \in [a,b]} k(x, x')}.$$

Most importantly, Corollary 2 is applicable to the Gaussian kernel, for which the upper bound is finite. This result will be discussed in Sect. 4.4 in more detail. One may see supporting evidence in Fig. 2, where the sum of weights seems to converge to a value around 1.

### 3.9 Higher dimensions

As far as we are aware of, there are no extensions of the theory of Chebyshev systems to higher dimensions. Consequently, it is not possible to say much about positivity of the weights when $d > 1$. Some simple cases can be analysed, however.

Let $\Omega_1 = [a, b]$, $\nu_1$ be a measure on $\Omega_1$, $\Omega = \Omega_1^d \subset \mathbb{R}^d$ and $\nu = \nu_1^d$. That is, $\Omega = \Omega_1 \times \cdots \times \Omega_1$ and $\mathrm{d}\nu(\boldsymbol{x}) = \mathrm{d}\nu_1(x_1) \times \cdots \times \mathrm{d}\nu_1(x_d)$, where there are $d$ terms in the products. Suppose that

(i) the point set $X$ is now a Cartesian product of one-dimensional sets $X_1 = \{x_1^1, \ldots, x_n^1\} \subset \Omega_1$: $X = X_1^d$;
(ii) the kernel is of product form: $k(\boldsymbol{x}, \boldsymbol{x}') = \prod_{i=1}^{d} k_1(x_i, x_i')$ for some kernel $k_1$ on $\Omega_1$.

A quadrature rule using Cartesian product points is called a *tensor product rule*. For such points, the Bayesian quadrature weights $\boldsymbol{w}_X^{\mathrm{BQ}}$ are products of the one-dimensional weights $\boldsymbol{w}_{X_1}^{\mathrm{BQ}}$: the weight for the point $(x_{i(1)}, \ldots, x_{i(d)}) \in X$ is $\prod_{j=1}^{d} w_{X_1,i(j)}^{\mathrm{BQ}}$ (Oettershagen 2017, Section 2.4). In particular, if $k_1$ is totally positive and $X_1$ is a locally optimal set of points, then all the $n^d$ weights $\boldsymbol{w}_X^{\mathrm{BQ}}$ are positive.[4] Analysis of more flexible sparse grid and symmetry-based methods (Karvonen and Särkkä 2018a) might yield more interesting results.

[4] Note that a tensor product rule based on an optimal one-dimensional point set need not be locally optimal for $\Omega$, $\nu$ and $k$.

We conclude this section with two numerical examples. Both of them involve the standard Gaussian measure

$$\mathrm{d}\nu(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right)\mathrm{d}\boldsymbol{x}$$

on $\Omega = \mathbb{R}^d$ and the Gaussian kernel (12).

*Locally optimal points* First, we investigated positivity of weights for locally optimal points. We set $\ell = 1$ and $d = 2$ and used a gradient-based quasi-Newton optimisation method (MATLAB's `fminunc`) to find points that locally minimise the integral variance for $n = 2, \ldots, 20$. Optimisation was initialised with a set of random points. The point set output by the optimiser was then randomly perturbated and optimisation repeated for 20 times, each time initialising with the point set giving the smallest Bayesian quadrature variance so far. The weights were always computed directly from (3). However, to improve numerical stability, the kernel matrix $\boldsymbol{K}_X$ was replaced by $\boldsymbol{K}_X + 10^{-6}\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix, during point optimisation. Some point sets generated using the same algorithm have appeared in Särkkä (2016, Section IV) [for other examples of optimal points in dimension two, see O'Hagan (1992) and Minka (2000) and, in particular, Oettershagen (2017, Chapter 6)]. The point sets we obtained appear sensible and all of them are associated with positive weights; four sets and their weights are depicted in Fig. 2. For $n = 20$, the maximal value of a partial derivative of $\mathbb{V}_X^{\mathrm{BQ}}$ at the computed points was $9 \times 10^{-10}$.

*Random points* Secondly, we investigated the validity of Theorem 1 in higher dimensions. We set $\ell = 1.5$ and $d = 4$ and counted the number of positive weights for $n = 2, \ldots, 1000$ when each $n$-point set is generated by drawing Monte Carlo samples from $\nu$. Random samples are often used in Bayesian quadrature (Rasmussen and Ghahramani 2002; Briol et al. 2019, 2017) and they also function as a suitable test case where structurality of point sets has little role in constraining behaviour of some subsets of the weights as happens when product or symmetric point designs is used. Figure 3 shows the proportion of positive weights; it appears that at least half of the weights for randomly drawn points are always positive. This supports the obvious conjectural extension to higher dimensions of Theorem 1.

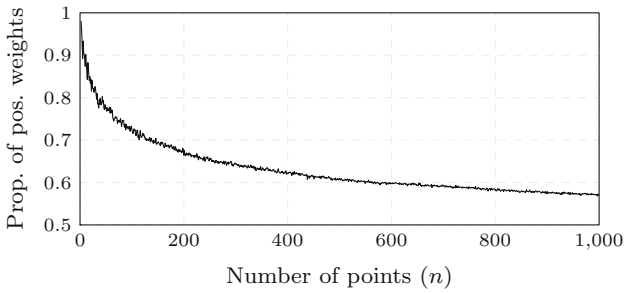


**Fig. 3** Proportion of positive weights for the Gaussian kernel and $n$ points drawn from the standard Gaussian distribution on $\mathbb{R}^4$. The results have been averaged over 50 independent Monte Carlo runs. Among all runs the minimal proportion encountered was exactly $1/2$

## 4 Magnitudes of weights and the stability

This section studies the magnitudes of the weights in a Bayesian quadrature rule and discusses how they are related to stability and robustness of the quadrature rule. We are in particular interested in the following quantity, which we call the Bayesian quadrature *stability constant*:

$$\Lambda_{X_n}^{\mathrm{BQ}} := \sum_{i=1}^{n} \left| w_{X_n,i}^{\mathrm{BQ}} \right|. \tag{15}$$

To make dependency on $n$ more explicit, the quadrature point set is denoted by $X_n$ instead of $X$ in this section. The terminology is motivated by the close connection of $\Lambda_{X_n}^{\mathrm{BQ}}$ to the *Lebesgue constant* $\Lambda_{X_n}$, a quantity that characterises the stability of an interpolant. For kernel interpolants, the Lebesgue constant is

$$\Lambda_{X_n} := \sup_{\boldsymbol{x}\in\Omega} \sum_{i=1}^{n} \left| u_{X_n,i}(\boldsymbol{x}) \right|,$$

where $u_{X_n,i}$ are Lagrange cardinal functions from Sect. 2.1. The connection to (15) arises from the fact that $w_{X_n,i}^{\mathrm{BQ}} = I_\nu(u_{X_n,i})$ for $i = 1, \ldots, n$.

The importance of the stability constant (15) is illustrated by the following argument. Let $\mu_f^*$ be an optimal approximant to the integrand function $f\colon \Omega \to \mathbb{R}$ in the span of $\{k_{\boldsymbol{x}_i}\}_{i=1}^n$ in the sense that

$$\mu_f^* \in \operatorname*{arg\,min}_{\mu_f \in \mathrm{span}\{k_{\boldsymbol{x}_i}\}_{i=1}^n} \left\| f - \mu_f \right\|_\infty,$$

where $\left\| f - \mu_f \right\|_\infty := \sup_{\boldsymbol{x}\in\Omega} \left| f(\boldsymbol{x}) - \mu_f(\boldsymbol{x}) \right|$ is the uniform norm. Note that $\mu_f^*$ does not in general interpolate $f$ at $X_n$ nor coincide with the Gaussian process posterior mean $\mu_{X,f}$. Then

$$\begin{aligned}
&\left| I_\nu(f) - I_{X_n}^{\mathrm{BQ}}(f) \right| \\
&\quad \le \left| I_\nu(f) - I_\nu(\mu_f^*) \right| + \left| I_\nu(\mu_f^*) - I_{X_n}^{\mathrm{BQ}}(f) \right| \\
&\quad = \left| I_\nu(f) - I_\nu(\mu_f^*) \right| + \left| I_{X_n}^{\mathrm{BQ}}(\mu_f^*) - I_{X_n}^{\mathrm{BQ}}(f) \right| \\
&\quad \le \left\| f - \mu_f^* \right\|_\infty + \sum_{i=1}^{n} \left| w_{X_n,i}^{\mathrm{BQ}} \right| \left| \mu_f^*(\boldsymbol{x}_i) - f(\boldsymbol{x}_i) \right| \\
&\quad \le (1 + \Lambda_{X_n}^{\mathrm{BQ}}) \left\| f - \mu_f^* \right\|_\infty,
\end{aligned}$$

where we have used the fact that $I^{\mathrm{BQ}}_{X_n}(g) = I_\nu(g)$ if $g \in \mathrm{span}\{k_{\boldsymbol{x}_i}\}_{i=1}^n$. That is, the approximation error by a Bayesian quadrature rule can be related to that by the best uniform approximant via the stability constant. The stability constant also controls the error introduced by inaccurate function evaluations. Suppose that the function evaluations contain errors (which may be numerical or stochastic), denoted by $\epsilon_i$ and modelled as indendent zero-mean random variables with variance $\sigma^2$. Then the mean-square error (where the expectation is w.r.t. $\epsilon_1, \ldots, \epsilon_n$) of Bayesian quadrature is given by

$$\begin{aligned}
&\mathbb{E}\left[\left(I_\nu(f) - \sum_{i=1}^n w^{\mathrm{BQ}}_{X_n,i}[f(\boldsymbol{x}_i) + \epsilon_i]\right)^2\right] \\
&= \left(I_\nu(f) - \sum_{i=1}^n w^{\mathrm{BQ}}_{X_n,i} f(\boldsymbol{x}_i)\right)^2 + \sigma^2 \sum_{i=1}^n (w^{\mathrm{BQ}}_{X_n,i})^2 \\
&\leq \left(I_\nu(f) - \sum_{i=1}^n w^{\mathrm{BQ}}_{X_n,i} f(\boldsymbol{x}_i)\right)^2 + \sigma^2 \left(\sum_{i=1}^n \left|w^{\mathrm{BQ}}_{X_n,i}\right|\right)^2.
\end{aligned}$$

This implies a small stability constant (15) suppresses the additional error caused by the perturbations $\epsilon_i$. A third motivating example will be given in Sect. 4.2, after introducing necessary notation.

It is clear from Lemma 1 that if the weights are positive for every $n$, the stability constant remains uniformly bounded. However, the results on positivity in the preceding section are valid only when $d = 1$ and the kernel is totally positive. This section uses a different technique to analyse the stability constant. The results are based on those in De Marchi and Schaback (2010), which are applicable to kernels that induce Sobolev-equivalent RKHSs (e.g. Matérn kernels). Accordingly, we mainly focus on such kernels in this section. We begin by reviewing basic properties of Sobolev spaces in Sect. 4.1 and convergence results for Bayesian quadrature in Sect. 4.2. The main results, Theorem 5 and Corollary 3, on the magnitudes of quadrature weights and the stability constant appear in Sect. 4.3. We discuss a relevant stability issue, known as the Runge phenomenon, for infinitely smooth kernels such as the Gaussian kernel in Sect. 4.4. Finally, simulation results in Sect. 4.5 demonstrate that the obtained upper bound is conservative; there is much room for improving the results.

*Notation and basic definitions* The Fourier transform $\hat{f}$ of a Lebesgue integrable $f : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\hat{f}(\boldsymbol{\xi}) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathrm{e}^{-\sqrt{-1}\, \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{x}} \mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\xi} \in \mathbb{R}^d.$$

Two normed vector spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ are *norm-equivalent* if $\mathcal{F}_1 = \mathcal{F}_2$ as a set and there exist constants $C_1, C_2 > 0$ such that

$$C_1 \|f\|_{\mathcal{F}_2} \leq \|f\|_{\mathcal{F}_1} \leq C_2 \|f\|_{\mathcal{F}_2} \quad \text{for all} \quad f \in \mathcal{F}_1.$$

### 4.1 Kernels inducing Sobolev-equivalent RKHSs

Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a continuous and integrable positive-definite function with Fourier transform satisfying

$$c_1(1 + \|\boldsymbol{\xi}\|^2)^{-r} \leq \hat{\Phi}(\boldsymbol{\xi}) \leq c_2(1 + \|\boldsymbol{\xi}\|^2)^{-r} \tag{16}$$

for $r > d/2$, some positive constants $c_1$ and $c_2$, and for all $\boldsymbol{\xi} \in \mathbb{R}^d$. In this section, we consider shift-invariant kernels on $\mathbb{R}^d$ of the form $k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x} - \boldsymbol{x}')$. For instance, a Matérn kernel [Rasmussen and Williams, 2006, Section 4.2.1]

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

with smoothness parameter $\rho := r - d/2$ and length-scale parameter $\ell > 0$ satisfies (16).[5] Here $\mathrm{K}_\rho$ is the modified Bessel function of the second kind of order $\rho$. Another notable class of kernels satisfying (16) are Wendland kernels (Wendland 2005, Theorem 10.35).

By Wendland (2005, Corollary 10.13), the RKHS $\mathcal{H}(k)$ of any kernel $k$ satisfying (16) is norm-equivalent to the *Sobolev space* $H^r(\mathbb{R}^d)$ of order $r > d/2$ on $\mathbb{R}^d$, which is a Hilbert space consisting of square-integrable and continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$\|f\|^2_{H^r(\mathbb{R}^d)} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \left(1 + \|\boldsymbol{\xi}\|^2\right)^r \left|\hat{f}(\boldsymbol{\xi})\right|^2 \mathrm{d}\boldsymbol{\xi} < \infty.$$

As can be seen from this expression, $r$ quantifies the smoothness of functions in $H^r(\mathbb{R}^d)$: as $r$ increases, function in $H^r(\mathbb{R}^d)$ become smoother.

The Sobolev space $H^r(\Omega)$ on a general measurable domain $\Omega \subset \mathbb{R}^d$ can be defined as the restriction of $H^r(\mathbb{R}^d)$ onto $\Omega$. The kernel $k$ satisfying (16), when seen as a kernel on $\Omega$, then induces an RKHS that is norm-equivalent to $H^r(\Omega)$ (Wendland 2005, Theorems 10.12, 10.46 and 10.47).[6]

[5] Note that the smoothness parametrisation $\rho = r$ is often used. With this parametrisation $k_\rho$ would satisfy (16) with the exponent $-(r + d/2)$ and its RKHS would be norm-equivalent to $H^{r+d/2}(\mathbb{R}^d)$.

[6] The reader may ask whether $\Omega$ needs to have a Lipschitz boundary for this norm-equivalence, but this assumption is indeed not needed. The assumption that $\Omega$ has a Lipschitz boundary is required when using *Stein's extension theorem* (Stein 1970, p. 181) for Sobolev spaces defined using weak derivatives [see the proof of Wendland (2005, Corollary 10.48)]. On the other hand, we consider here a Sobolev space defined in terms of the Fourier transform, and the norm-equivalence follows from the extension and restriction theorems for a generic RKHS (Wendland 2005, Theorems 10.46 and 10.47) and the expression of the RKHS norm in terms of Fourier transforms (Wendland 2005, Theorem 10.12).

### 4.2 Convergence for Sobolev-equivalent kernels

Recall from Sect. 2.2 that integration error by a Bayesian quadrature rule for functions in $\mathcal{H}(k)$ satisfies

$$\left|I_\nu(f) - I^{\text{BQ}}_{X_n}(f)\right| \leq \|f\|_{\mathcal{H}(k)}\, e_{\mathcal{H}(k)}(X_n, \boldsymbol{w}^{\text{BQ}}_X),$$

so that in convergence analysis only the behaviour of the worst-case error needs to be considered. If the RKHS is norm-equivalent to a Sobolev space, rates of convergence for Bayesian quadrature can be established. These results follow from Arcangéli et al. (2007, Corollary 4.1). See Wendland (2005, Corollary 11.33) or Wendland and Rieger (2005, Proposition 3.6) for earlier and slightly more restricted results that require $\lfloor r \rfloor > d/2$ and Kanagawa et al. (2019, Proposition 4) for a version specifically for numerical integration. Some assumptions, satisfied by all domains of interest to us, are needed; see for instance Kanagawa et al. (2019, Section 3) for precise definitions.

**Assumption 3** The set $\Omega \subset \mathbb{R}^d$ is a bounded open set that satisfies an interior cone condition and has a Lipschitz boundary.

This assumption essentially says that the boundary of $\Omega$ is sufficiently regular (Lipschitz boundary) and that there is no "pinch point" on the boundary of $\Omega$ (interior cone condition). Convergence results are expressed in terms of the *fill-distance*

$$h_{X_n,\Omega} := \sup_{\boldsymbol{x}\in\Omega} \min_{i=1,\ldots,n} \|\boldsymbol{x} - \boldsymbol{x}_i\|$$

that quantifies the size of the largest "hole" in an $n$-point set $X_n$. We use $\lesssim$ to denote an inequality that is valid up to a constant independent of $n$, number of points, and $f$, the integrand. That is, for generic sequences of functionals $g_n$ and $h_n$, $g_n(f) \lesssim h_n(f)$ means that there is a constant $C > 0$ such that $g_n(f) \leq C h_n(f)$ for all $n \in \mathbb{N}$ and any $f$ in a specified class of functions.

**Theorem 4** *Suppose that (i) $\Omega$ satisfies Assumption 3 (ii) that the measure $\nu$ has a bounded (Lebesgue) density function and that (iii) the kernel $k$ satisfies (16) for a constant $r$ such that $r > d/2$. Then*

$$\left|I_\nu(f) - I^{\text{BQ}}_{X_n}(f)\right| \lesssim \|f\|_{H^r(\Omega)}\, h^r_{X_n,\Omega}$$

*for any $f \in H^r(\Omega)$ when the fill-distance is sufficiently small.*

The following simple result is an immediate consequence of this theorem.

**Proposition 6** *Suppose that the assumptions of Theorem 4 are satisfied. Then $\left|1 - \sum_{i=1}^n w^{\text{BQ}}_{X_n,i}\right| \lesssim h^r_{X_n,\Omega}$ when the fill-distance is sufficiently small.*

***Proof*** Under the assumptions, constant functions are in $H^r(\Omega)$. Setting $f \equiv 1$ in (17) verifies the claim. □

Note that the same argument can be used whenever a general rate of convergence for functions in an RKHS is known and constant functions are contained in the RKHS. However, this is not always the case; for example, the RKHS of the Gaussian kernel (12) does not contain polynomials (Minh 2010, Theorem 2).

Rates explicitly dependent on the number of points are achieved for point sets that are *quasi-uniform*, which is to say that

$$\tilde{c}_1 q_{X_n} \leq h_{X_n,\Omega} \leq \tilde{c}_2 q_{X_n}$$

for some constants $\tilde{c}_1, \tilde{c}_2 > 0$ independent of $n$. Here

$$q_X := \frac{1}{2} \min_{i\neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$$

is the *separation distance*. In dimension $d$ quasi-uniform sets satisfy $h_{X_n,\Omega} = \mathcal{O}(n^{-1/d})$ as $n \to \infty$ (e.g. regular product grids). In Theorem 4 we thus obtain the rate

$$\left|I_\nu(f) - I^{\text{BQ}}_{X_n}(f)\right| \lesssim n^{-r/d} \tag{17}$$

for $f \in H^r(\Omega)$ when the point sets are quasi-uniform and $n$ is sufficiently large.

Of course, it is the stability constant $\Lambda^{\text{BQ}}_{X_n} = \sum_{i=1}^n \left|w^{\text{BQ}}_{X_n,i}\right|$ that we analyse next whose behaviour is typically more consequential. However, the above proposition may be occasionally interesting if one desires to interpret Bayesian quadrature as a weighted Dirac approximation $\nu_{\text{BQ}} := \sum_{i=1}^n w^{\text{BQ}}_{X_n,i}\delta_{\boldsymbol{x}_i} \approx \nu$ of a probability measure (i.e. $\nu_{\text{BQ}}(\Omega) \approx 1$). Note that there is also a simple way to ensure summing up to one of the weights by inclusion of a non-zero prior mean function for the Gaussian process prior; see O'Hagan (1991) and Karvonen et al. (2018b, Section 2.3).

Finally, we provide a third example that highlights the importance of analysing the stability constant. Kanagawa et al. (2019, Section 4.1) [see also Kanagawa et al. (2016)] studied convergence rates of kernel-based quadrature rules in Sobolev spaces when the integrand is potentially *rougher* (i.e. $f \in H^s(\Omega)$ for some $s \leq r$) than assumed. If $s < r$, the integrand $f$ may not belong to the Sobolev space $H^r(\Omega)$ that is assumed by the user when constructing the quadrature rule; therefore, this is a *misspecifed* setting. Under certain conditions, they showed (Kanagawa et al. 2019, Corollary 7) that if $\Lambda^{\text{BQ}}_{X_n} \lesssim n^c$ for a constant $c \geq 0$, then

$$\left|I_\nu(f) - I^{\text{BQ}}_{X_n}(f)\right| \lesssim n^{-s/d + c(r-s)/r}, \tag{18}$$

when $X_n$ are quasi-uniform.

The condition $\Lambda_{X_n}^{\mathrm{BQ}} \lesssim n^c$ means that the stability constant $\Lambda_{X_n}^{\mathrm{BQ}}$ should not grow quickly as $n$ increases. The bound (18) shows that the error in the misspecified setting becomes small if $c$ is small. This implies that if the stability constant $\Lambda_{X_n}^{\mathrm{BQ}}$ does not increase quickly, then the quadrature rule becomes robust against the misspecification of a prior. This provides a third motivation for understanding the behaviour of $\Lambda_{X_n}^{\mathrm{BQ}}$.

### 4.3 Upper bounds for absolute weights

We now analyse magnitudes of individual weights and the stability constant (15). We first derive an upper bound on the magnitude of each weight $w_{X_n,i}^{\mathrm{BQ}}$. The proof of this result is based on an upper bound on the $L^2(\Omega)$ norm of Lagrange functions derived in De Marchi and Schaback (2010).

**Theorem 5** *Suppose that (i) $\Omega$ satisfies Assumption 3, that (ii) the measure $\nu$ has a bounded (Lebesgue) density function and that (iii) the kernel $k$ satisfies (16) for a constant $r$ such that $r > d/2$. Then*

$$\left| w_{X_n,i}^{\mathrm{BQ}} \right| \lesssim \left( \frac{h_{X_n,\Omega}}{q_{X_n}} \right)^{r-d/2} h_{X_n,\Omega}^{d/2} \tag{19}$$

*for all $i = 1, \ldots, n$, provided that $h_{X_n,\Omega}$ is sufficiently small. When $X_n$ are quasi-uniform, this becomes*

$$\left| w_{X_n,i}^{\mathrm{BQ}} \right| \lesssim n^{-1/2} \tag{20}$$

*for $n$ large enough.*

***Proof*** It is proved in De Marchi and Schaback (2010, Theorem 1) that each of the Lagrange functions $u_{X_n,i}$ admits the bound

$$\left( \int_\Omega u_{X_n,i}(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} \right)^{1/2} \lesssim \left( \frac{h_{X_n,\Omega}}{q_{X_n}} \right)^{r-d/2} h_{X_n,\Omega}^{d/2}, \tag{21}$$

provided that $h_{X_n,\Omega}$ is sufficiently small. Let $\|\nu\|_\infty < \infty$ stand for the supremum of the density function of $\nu$. Then it follows from $w_{X_n,i}^{\mathrm{BQ}} = I_\nu(u_{X_n,i})$ that

$$\begin{aligned} |w_{X_n,i}^{\mathrm{BQ}}| &\leq \int |u_{X_n,i}(\boldsymbol{x})| d\nu(\boldsymbol{x}) \\ &\leq \left( \int_\Omega u_{X_n,i}(\boldsymbol{x})^2 \mathrm{d}\nu(\boldsymbol{x}) \right)^{1/2} \\ &\leq \|\nu\|_\infty \left( \int_\Omega u_{X_n,i}(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} \right)^{1/2}. \end{aligned}$$

Inequality (19) now follows from (21). When $X_n$ are quasi-uniform, the ratio $h_{X_n,\Omega}/q_{X_n}$ remains bounded and $h_{X_n,\Omega}$ behaves like $n^{-1/d}$. □

An important consequence of Theorem 5 is that the magnitudes of quadrature weights decrease uniformly to zero as $n$ increases if the design points are quasi-uniform and $\nu$ has a density. In other words, none of the design points will have a constant weight that does not decay. This is similar to importance sampling, where the weights decay uniformly at rate $1/n$. As a direct corollary of Theorem 5 we obtain bounds on the stability constant $\Lambda_{X_n}^{\mathrm{BQ}}$.

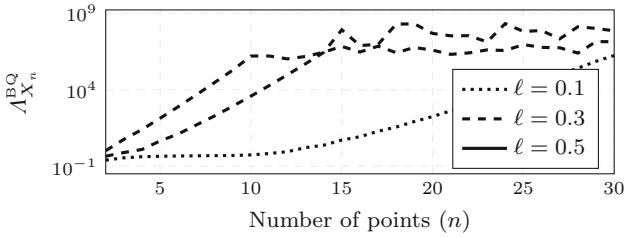


**Fig. 4** Bayesian quadrature stability constants for the Gaussian kernel (12) with different length-scales, the uniform measure on [0, 1] and $n$ points uniformly placed on this interval (end points not included). The levelling off appears to be caused by loss of numerical precision

**Corollary 3** *Under the assumptions of Theorem 5 and provided that $h_{X_n,\Omega}$ is sufficiently small we have*

$$\Lambda_{X_n}^{\mathrm{BQ}} \lesssim n \left( \frac{h_{X_n,\Omega}}{q_{X_n}} \right)^{r-d/2} h_{X_n,\Omega}^{d/2}. \tag{22}$$

*When $X_n$ are quasi-uniform and $n$ sufficiently large this becomes*

$$\Lambda_{X_n}^{\mathrm{BQ}} \lesssim \sqrt{n}. \tag{23}$$

While the bounds of Corollary 3 are somewhat conservative (as will be demonstrated in Sect. 4.5), they are still useful in understanding the factors affecting stability and robustness of Bayesian quadrature. That is, inequality (22) shows that the stability constant can be made small if the ratio $h_{X_n,\Omega}/q_{X_n}$ is kept small; this is possible if the point set is sufficiently uniform.

Another important observation concerns the exponent $r - d/2$ of the ratio $h_{X_n,\Omega}/q_{X_n}$: if the smoothness $r$ of the kernel is large, then the stability constant may also become large if the points are not quasi-uniform. This is true because $h_{X_n,\Omega}/q_{X_n} \geq 1$ for any configuration of $X_n$, as can be seen easily from the definitions of $q_{X_n}$ and $h_{X_n,\Omega}$. This observation implies that the use of a smoother kernel may lead to higher numerical instability. Accordingly, we next discuss stability of infinitely smooth kernels and the Runge phenomenon that manifests itself in this setting.

### 4.4 On infinitely smooth kernels

While the theoretical results of this section only concern kernels of finite smoothness, we make a few remarks on the stability of Bayesian quadrature when using infinitely smooth

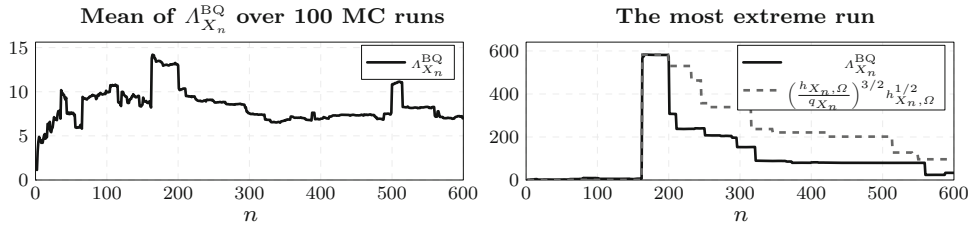


**Fig. 5** The Bayesian quadrature stability constant for a Matérn kernel, the uniform measure on $[0, 1]$ and $n$ points drawn from the uniform distribution. Left: $\Lambda^{\mathrm{BQ}}_{X_n}$ averaged over 100 independent Monte Carlo runs. Right: the run where most extreme behaviour, in terms of $\Lambda^{\mathrm{BQ}}_{X_n}$ attaining maximal value, was observed. Plotted are both $\Lambda^{\mathrm{BQ}}_{X_n}$ and a scaled version of $(h_{X_n,\Omega}/q_{X_n})^{3/2} h^{1/2}_{X_n,\Omega}$ (its true maximum was roughly $2.2\times10^7$) that is expected to control the stability constant. Note that the theoretical upper bound (22) contains an additional multiplication by $n$

kernels, such as the Gaussian kernel. When using such a kernel, Bayesian quadrature rules suffer from the famous *Runge phenomenon*: if equispaced points are used, then Lebesgue constants and the stability constants grow rapidly; see Oettershagen (2017, Section 4.3), Platte and Driscoll (2005) and Platte et al. (2011). This effect is demonstrated in Fig. 4, and can be seen also in Sommariva and Vianello (2006b, Table 1).

A key point is that Runge phenomenon typically occurs when the design points are quasi-uniform (e.g. equispaced). This means that quasi-uniformity of the points does not ensure stability of Bayesian quadrature when the kernel is infinitely smooth. Care has to be taken if a numerically stable Bayesian quadrature rule is to be constructed with such a kernel. One possibility is to use locally optimal design points from Sect. 3.5. Corollary 2 then guarantees uniform boundedness of the stability constant, at least when $d = 1$.

### 4.5 A numerical example

Numerical examples of the behaviour of kernel Lebesgue constants can be found in De Marchi and Schaback (2008), where it was observed that the theoretical bounds similar to (23) are conservative: the Lebesgue constant appears to remain uniformly bounded. Bayesian quadrature weights are no different. We experimented with the Matérn kernel

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

with length-scale $\ell = 0.5$ and the uniform measure on the interval $[0, 1]$. When uniformly spaced points were used, all weights remained positive and their sum quickly converged to one when $n$ was increased. In contrast, Corollary 3 provides the, up to a constant, upper bound $\sqrt{n}$ that is in this case clearly very conservative. When points were drawn from the uniform distribution on $[0, 1]$, more interesting behaviour was observed (Fig. 5). As expected, the magnitude of $\Lambda^{\mathrm{BQ}}_{X_n}$ was closely related to the ratio $h_{X_n,\Omega}/q_{X_n}$. Nevertheless, increase in $n$ did not generally correspond to increase in $\Lambda^{\mathrm{BQ}}_{X_n}$.

Note that the results of Sect. 3 do not explain why the weights became positive in this experiment, because Matérn kernels do not appear to be totally positive *even if the differentiability requirements were to be relaxed and only single zeroes counted* (recall Remark 1). We have numerically observed that selecting $n > \rho + 1/2$ and point sets such that $\max X < \min Y$ makes the matrix $\boldsymbol{K}_{Y,X}$ discussed in Sect. 3.3 singular for the Matérn kernel $k_\rho$. This implies that there is a non-trivial linear combination of the $n$ Matérn translates at $X$ that vanishes at more than $n - 1$ points. When $\rho = 1/2$ and $\ell = 1$ (so that $k_\rho(x, x') = \mathrm{e}^{-|x-y|}$), an analytical counterexample can be constructed by setting $X = \{x_1, x_2\}$ and $Y = \{x_1+h, x_2+h\}$ with $h > x_2-x_1$. Then

$$\boldsymbol{K}_{Y,X} = \mathrm{e}^{-h} \begin{bmatrix} 1 & \mathrm{e}^{-(x_2-x_1)} \\ \mathrm{e}^{-(x_1-x_2)} & 1 \end{bmatrix},$$

which is not invertible because multiplying the first row by $\mathrm{e}^{-(x_1-x_2)}$ yields the second row. Therefore, the positivity of quadrature weights for Matérns and other kernels with finite smoothness requires a further research.

**Acknowledgements** Open access funding provided by Aalto University. TK was supported by the Aalto ELEC Doctoral School. MK acknowledges support by the European Research Council (StG Project PANAMA). SS was supported by the Academy of Finland project 313708. This material was developed, in part, at the *Prob Num 2018* workshop hosted by the Lloyd's Register Foundation programme on Data-Centric Engineering at the Alan Turing Institute, UK, and supported by the National Science Foundation, USA, under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the above-named funding bodies and research institutions.

## References


Arcangéli, R., de Silanes, M.C.L., Torrnes, J.J.: An extension of a bound for functions in Sobolev spaces, with applications to $(m, s)$-spline interpolation and smoothing. Numer. Math. **108**(2), 181–211 (2007)

Atkinson, K.E.: An Introduction to Numerical Analysis, 2nd edn. Wiley, Amsterdam (1989)

Barrar, R.B., Loeb, H.L.: Multiple zeroes and applications to optimal linear functionals. Numer. Math. **25**(3), 251–262 (1976)

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

Barrow, D.L.: On multiple node Gaussian quadrature formulae. Math. Comput. **32**(142), 431–439 (1978)

Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer, New York (2011)

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

Breger, A., Ehler, M., Gräf, M.: Points on manifolds with asymptotically optimal covering radius. J. Complex. **48**, 1–14 (2018)

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

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

Burbea, J.: Total positivity of certain reproducing kernels. Pac. J. Math. **67**(1), 101–130 (1976)

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

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

Cockayne, J., Oates, C. J., Sullivan, T., Girolami, M.: Bayesian probabilistic numerical methods. SIAM Rev. arxiv:1702.03673 (2019)

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

De Marchi, S., Schaback, R.: Stability constants for kernel-based interpolation processes. Technical Report 59/08, Universita degli Studi di Verona (2008)

De Marchi, S., Schaback, R.: Stability of kernel-based interpolation. Adv. Comput. Math. **32**(2), 155–161 (2010)

Diaconis, P.: Bayesian numerical analysis. In: Gupta, S.S., Berger, J.O. (eds.) Statistical Decision Theory and Related Topics IV, vol. 1, pp. 163–175. Springer-Verlag, New York (1988)

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

Förster, K.J.: Variance in quadrature—a survey. In: Brass, H., Hämmerlin, G. (eds.) Numerical Integration IV, vol. 112, pp. 91–110. Birkhäuser, Basel (1993)

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

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

Gavrilov, A.V.: On optimal quadrature formulas. J. Appl. Ind. Math. **1**(2), 190–192 (2007)

Gunter, T., Osborne, M.A., Garnett, R., Hennig, P., Roberts, S.J.: Sampling for inference in probabilistic models with fast Bayesian quadrature. Adv. Neural Inf. Process. Syst. **27**, 2789–2797 (2014)

Hennig, P., Osborne, M.A., Girolami, M.: Probabilistic numerics and uncertainty in computations. Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci. **471**(2179), 20150142 (2015)

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

Jagadeeswaran, R., Hickernell, F. J.: Fast automatic Bayesian cubature using lattice sampling. Stat. Comput. (2019). https://doi.org/10.1007/s11222-019-09895-9

Kanagawa, M., Sriperumbudur, B.K., Fukumizu, K.: Convergence guarantees for kernel-based quadrature rules in misspecified settings. Adv. Neural Inf. Process. Syst. **29**, 3288–3296 (2016)

Kanagawa, M., Sriperumbudur, B.K., Fukumizu, K.: Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. Found. Comput. Math. (2019). https://doi.org/10.1007/s10208-018-09407-7

Karlin, S.: Total Positivity, vol. 1. Stanford University Press, Palo Alto (1968)

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

Karvonen, T., Särkkä, S.: Classical quadrature rules via Gaussian processes. In: 27th IEEE International Workshop on Machine Learning for Signal Processing (2017)

Karvonen, T., Särkkä, S.: Fully symmetric kernel quadrature. SIAM J. Sci. Comput. **40**(2), A697–A720 (2018)

Karvonen, T., Särkkä, S.: Gaussian kernel quadrature at scaled Gauss–Hermite nodes. Bit Numer Math (2019). https://doi.org/10.1007/s10543-019-00758-3

Karvonen, T., Oates, C.J., Särkkä, S.: A Bayes-Sard cubature method. Adv. Neural Inf. Process. Syst. **31**, 5882–5893 (2018)

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

Larkin, F.M.: Gaussian measure in Hilbert space and applications in numerical analysis. Rocky Mt. J. Math. **2**(3), 379–421 (1972)

Lee, J. D., Simchowitz, M., Jordan, M. I., Recht, B.: Gradient descent only converges to minimizers. In: 29th Annual Conference on Learning Theory, pp. 1246–1257 (2016)

Mhaskar, H.N., Narcowich, F.J., Ward, J.D.: Spherical Marcinkiewicz–Zygmund inequalities and positive quadrature. Math. Comput. **70**(235), 1113–1130 (2001)

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

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

Novak, E.: Intractability results for positive quadrature formulas and extremal problems for trigonometric polynomials. J. Complex. **15**(3), 299–316 (1999)

Oates, C.J., Niederer, S., Lee, A., Briol, F.-X., Girolami, M.: Probabilistic models for integration error in the assessment of functional cardiac models. Adv. Neural Inf. Process. Syst. **30**, 109–117 (2017)

Oettershagen, J.: Construction of optimal cubature algorithms with applications to econometrics and uncertainty quantification. Ph.D. thesis, Institut für Numerische Simulation, Universität Bonn (2017)

O'Hagan, A.: Bayes–Hermite quadrature. J. Stat. Plann. Inference **29**(3), 245–260 (1991)

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

Osborne, M., Garnett, R., Ghahramani, Z., Duvenaud, D.K., Roberts, S.J., Rasmussen, C.E.: Active learning of model evidence using Bayesian quadrature. Adv. Neural Inf. Process. Syst. **25**, 46–54 (2012)

Platte, R.B., Driscoll, T.B.: Polynomials and potential theory for Gaussian radial basis function interpolation. SIAM J. Numer. Anal. **43**(2), 750–766 (2005)

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

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

Rasmussen, C.E., Ghahramani, Z.: Bayesian Monte Carlo. Adv. Neural Inf. Process. Syst. **15**, 505–512 (2002)

Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)

Richter, N.: Properties of minimal integration rules. SIAM J. Numer. Anal. **7**(1), 67–79 (1970)

Richter-Dyn, N.: Properties of minimal integration rules. II. SIAM J. Numer. Anal. **8**(3), 497–508 (1971a)

Richter-Dyn, N.: Minimal interpolation and approximation in Hilbert spaces. SIAM J. Numer. Anal. **8**(3), 583–597 (1971b)

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

Särkkä, S., Hartikainen, J., Svensson, L., Sandblom, F.: On the relation between Gaussian process quadratures and sigma-point methods. J. Adv. Inf. Fusion **11**(1), 31–46 (2016)

Smola, A., Gretton, A., Song, L., Schölkopf, B.: A Hilbert space embedding for distributions. In: International Conference on Algorithmic Learning Theory, pp. 13–31. Springer (2007)

Sommariva, A., Vianello, M.: Numerical cubature on scattered data by radial basis functions. Computing **76**(3–4), 295–310 (2006a)

Sommariva, A., Vianello, M.: Meshless cubature by Green's formula. Appl. Math. Comput. **183**(2), 1098–1107 (2006b)

Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton (1970)

Steinwart, I., Christmann, A.: Support Vector Machines. Information Science and Statistics. Springer, New York (2008)

Wendland, H.: Scattered Data Approximation. Number 28 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2005)

Wendland, H., Rieger, C.: Approximate interpolation with applications to selecting smoothing parameters. Numer. Math. **101**(4), 729–748 (2005)

Wu, A., Aoi, M. C., Pillow, J. W.: Exploiting gradients and Hessians in Bayesian optimization and Bayesian quadrature. Preprint. arXiv:1704.00060 (2018)