Classical quadrature rules via Gaussian processes

Toni Karvonen and Simo Särkkä

This article has been accepted for publication in *Proceedings of 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, to be held in Tokyo, Japan, on September 25–28, 2017. When citing this article, use the following reference:

> T. Karvonen and S. Särkkä, "Classical quadrature rules via Gaussian processes," in *27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017. Accepted for publication.

2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

CLASSICAL QUADRATURE RULES VIA GAUSSIAN PROCESSES

Toni Karvonen and Simo Särkkä

Aalto University
Department of Electrical Engineering and Automation
Rakentajanaukio 2 C, Espoo, Finland

ABSTRACT

In an extension to some previous work on the topic, we show how all classical polynomial-based quadrature rules can be interpreted as Bayesian quadrature rules if the covariance kernel is selected suitably. As the resulting Bayesian quadrature rules have zero posterior integral variance, the results of this article are mostly of theoretical interest in clarifying the relationship between the two different approaches to numerical integration.

***Index Terms*—** Bayesian quadrature, kernel quadrature, classical quadrature, Gaussian process regression

1. INTRODUCTION

Bayesian quadrature (BQ) [1, 2], non-probabilistic origins of which go at least back to the work of Larkin in the 1970s [3, 4], is an exciting alternative to classical quadrature rules for numerical computation of intractable integrals. In contrast to the the classical approach of constructing an integration rule so that low-degree polynomials are integrated exactly, Bayesian quadrature is based on treating the integrand as a Gaussian process (GP; prompting the alternative term Gaussian process quadrature) of which a number of "observations" are available. The observations induce a Gaussian posterior distribution, the mean of which is a quadrature rule, on the integral. What makes Bayesian quadrature particularly interesting is that, following the paradigm of probabilistic numerics [5, 6, 7], the posterior integral variance can be used in modelling uncertainty inherent to the quadrature approximation.

There are interesting connections between the classical and probabilistic approaches to numerical integration (and approximation in general). In their famous work, Kimeldorf and Wahba [8] established a connection between stochastic processes and interpolation by certain types of splines. Diaconis [9] reviews how many basic quadrature rules, the trapezoidal rule among them, can be obtained using (integrated) Brownian motion priors. Lately, there has been renewed interest in interpreting many classical numerical methods both for numerical integration [10] and for solving differential equations [11, 12] in probabilistic framework.

Särkkä *et al.* [10] showed that many quadrature and cubature rules[1] frequently used in Kalman filtering for non-linear systems can be interpreted as Bayesian integration rules if the covariance kernel of the underlying GP prior is of certain polynomial form. In this article we prove that all classical (in a sense that is made clear in Section 2.1) quadrature rules admit such an interpretation and make precise what is required of the degree of this polynomial kernel, defined in Equation (5). In particular, Gaussian quadrature rules are unique optimal Bayesian quadrature rules for appropriate choice of degree of the polynomial kernel. Our main result is Theorem 4 in Section 3 and some simple numerical examples are presented in Section 5. We also sketch some generalisations for cubature rules in Section 4 and point out an interesting connection to Mysovskikh's theorem [13] that features prominently in construction of polynomially optimal cubature rules.

It is to be noted that the results of this article are of no help to practitioners wishing to replicate classical methods using stochastic processes in order to model uncertainty arising from the numerical approximation. The reason is that for those of the kernels introduced in this article that result in the BQ coinciding with a classical rule the integral posterior variance is zero. The same defect is present already in [10].

2. CLASSICAL AND BAYESIAN QUADRATURE

Let Ω be a subset of the real line, μ a probability measure on Ω, and $f\colon \Omega \to \mathbb{R}$ a measurable function. A quadrature rule Q is an approximation to the integral $\mu(f) \coloneqq \int_\Omega f \,\mathrm{d}\mu$ of the form

$$Q(f) \coloneqq \sum_{i=1}^{n} w_i f(x_i) \approx \int_\Omega f \,\mathrm{d}\mu, \tag{1}$$

where $x_1, \ldots, x_n \in \Omega$ are the *nodes* (*points*, *sigma-points*) and $w_1, \ldots, w_n \in \mathbb{R}$ the *weights*. Effectively, the measure μ and quadrature rule Q are functionals that map μ-measurable functions to reals. This is reflected in our notation.

This work was supported by Aalto ELEC Doctoral School and Academy of Finland projects 266940, 2734755, and 295080.

[1] In this article, *quadrature* always refers to numerical integration rules for univariate functions and *cubature* for multivariate functions.

978-1-5090-6341-3/17/$31.00

2.1. Classical quadrature

The classical approach to numerical integration is to select the nodes and weights so that the quadrature approximation (1) is exact whenever the integrand is a low-degree polynomial [14, 15]. Such a quadrature rule is called a *classical quadrature rule* (this is formulated precisely in Definition 2).

Definition 1. A quadrature rule Q is of *degree* m if it is exact for all polynomials of degree at most m and inexact for at least one polynomial of degree $m+1$.

The n polynomial exactness conditions $Q(x^i) = \mu(x^i)$, $i = 0, \dots, n-1$, correspond to the linear system of equations

$$\begin{pmatrix} x_1^0 & \cdots & x_n^0 \\ \vdots & \ddots & \vdots \\ x_1^{n-1} & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \mu(x^0) \\ \vdots \\ \mu(x^{n-1}) \end{pmatrix} \tag{2}$$

for the weights. If the nodes are distinct, the Vandermonde matrix on the left-hand side is non-singular. Hence, for any distinct nodes there exists a unique classical quadrature rule of degree at least $n-1$.

Definition 2. An n-point quadrature rule is called a *classical quadrature rule* if its degree is at least $n-1$. That is, its weights are the unique solution to the linear system of equations (2).

One can do much better in terms of polynomial exactness. Gaussian quadrature rules use n points to integrate polynomials up to degree $2n-1$ exactly and are optimal in this respect. See [16, Section 1.4] for the theory and precise conditions on μ (that we assume hold throughout this article) behind the following theorem. There exists a unique (up to constant coefficients) sequence $\psi_0, \psi_1, \dots$ of *orthogonal polynomials* that satisfy

$$\int_\Omega \psi_i \psi_j \,\mathrm{d}\mu = c_i \delta_{ij}, \;\; c_i \neq 0,$$

and $\deg \psi_i = i$. A particular property of these polynomials is that $\mu(\psi_i) = 0$ for $i > 0$ because $\varphi_i = \varphi_0 \varphi_i / \sqrt{c_0}$.

Theorem 3 (Gaussian quadrature)**.** *There exists a unique n-point classical quadrature rule of degree $2n-1$. The nodes of this rule are the roots of the nth orthogonal polynomial ψ_n and its weights are positive.*

Degree of an n-point classical quadrature rule is not restricted to $n-1$ or $2n-1$. A rule of degree $2n-m-1$, $1 \leq m \leq n$, can be constructed by selecting the nodes to be the roots of the polynomial

$$\psi_n + \gamma_{n-1}\psi_{n-1} + \cdots + \gamma_{n-m}\psi_{n-m}$$

for any non-zero constants $\gamma_{n-1}, \dots, \gamma_{n-m}$. This follows easily for example from [16, Theorem 1.45].

2.2. Bayesian quadrature

In *Bayesian quadrature*, the integrand $f\colon \Omega \to \mathbb{R}$ is assigned a Gaussian process prior [17]. What this means is that for any finite collection $x_1, \dots, x_n \in \Omega$ of points the joint distribution of $f(x_1), \dots, f(x_n)$ is $\mathcal{N}(\mathbf{0}, \mathbf{K})$, where $[\mathbf{K}]_{ij} = k(x_i, x_j)$ is the covariance (or kernel) matrix defined by the choice of the symmetric covariance kernel $k\colon \Omega \times \Omega \to \mathbb{R}$ (not assumed positive-definite in general).

If the number of points n is such that the matrix $\mathbf{K}$ is positive-definite, the data $\mathcal{D}$, consisting of the points $\mathcal{X} = \{x_1, \dots, x_n\}$ and the function evaluations (or "observations" in GP jargon) $\mathbf{y} = (f(x_1), \dots, f(x_n))$, induces a Gaussian posterior process $f \mid \mathcal{D}$ with the mean and covariance

$$\begin{aligned} \mathbb{E}[f(x) \mid \mathcal{D}] &= \mathbf{y}^\mathsf{T} \mathbf{K}^{-1} k(\mathcal{X}, x), \\ \operatorname{Cov}[f(x), f(x') \mid \mathcal{D}] &= k(x, x') - k(\mathcal{X}, x)^\mathsf{T} \mathbf{K}^{-1} k(\mathcal{X}, x'), \end{aligned}$$

where $[k(\mathcal{X}, x)]_i = k(x_i, x)$. It then follows [1, 18, 2] that the integral $\mu(f) = \int_\Omega f \,\mathrm{d}\mu$ has a Gaussian distribution with the mean and variance

$$\begin{aligned} Q_{\text{BQ}}(f) &\coloneqq \mathbb{E}[\mu(f) \mid \mathcal{D}] = \mathbf{y}^\mathsf{T} \mathbf{K}^{-1} k_\mu(\mathcal{X}), \\ V_{\text{BQ}} &\coloneqq \operatorname{Var}[\mu(f) \mid \mathcal{D}] = \mu(k_\mu) - k_\mu(\mathcal{X})^\mathsf{T} \mathbf{K}^{-1} k_\mu(\mathcal{X}), \end{aligned}$$

where $k_\mu \coloneqq \int_\Omega k(\cdot, x) \,\mathrm{d}\mu(x)$ is the *kernel mean*. The integral posterior mean Q_{BQ} is called the *Bayesian quadrature rule* and is indeed of the form (1):

$$Q_{\text{BQ}}(f) = \sum_{i=1}^n [\mathbf{K}^{-1} k_\mu(\mathcal{X})]_i f(x_i) =: \sum_{i=1}^n w_i^{\text{BQ}} f(x_i).$$

The posterior variance, that is independent of the integrand, is actually the Bayesian quadrature approximation error for the kernel mean:

$$V_{\text{BQ}} = \mu(k_\mu) - Q_{\text{BQ}}(k_\mu). \tag{3}$$

This can be seen by recognising that

$$k_\mu(\mathcal{X})^\mathsf{T} \mathbf{K}^{-1} k_\mu(\mathcal{X}) = k_\mu(\mathcal{X})^\mathsf{T} \mathbf{w}_{\text{BQ}}$$

in the expression for the variance.

Minimisation of the integral posterior variance V_{BQ} provides a natural way of selecting the "best" Bayesian quadrature nodes. An *optimal Bayesian quadrature rule* is a Bayesian quadrature rule whose nodes $\mathcal{X}^{\text{opt}}$ globally minimise V_{BQ}:

$$\mathcal{X}^{\text{opt}} = \operatorname*{arg\,min}_{\mathcal{X} \in \Omega^n} V_{\text{BQ}} = \operatorname*{arg\,min}_{\mathcal{X} \in \Omega^n} \left[\mu(k_\mu) - k_\mu(\mathcal{X})^\mathsf{T} \mathbf{K}^{-1} k_\mu(\mathcal{X})\right].$$

2.3. Reproducing kernel Hilbert spaces

Every positive-definite kernel (and some other kernels too, as discussed in Section 3) induces a unique *reproducing kernel Hilbert space* [19, 20] (RKHS) $\mathcal{H}$ of functions from Ω to $\mathbb{R}$ characterised by the properties i) $k(\cdot, x) \in \mathcal{H}$ for every $x \in \Omega$

and ii) $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ for every $x \in \Omega$ (the *reproducing property*).

The *worst-case error* (WCE) $e(Q)$ of a quadrature rule Q is defined as

$$e(Q) \coloneqq \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mu(f) - Q(f)|. \tag{4}$$

It turns out (see e.g. [2, 21] for a more thorough review) that the weights of a Bayesian quadrature rule with given nodes are such that the WCE is minimised among all quadrature rules $Q_{\mathcal{X}}$ with a fixed set of nodes $\mathcal{X} = \{x_1, \ldots, x_n\}$. That is,

$$\begin{aligned}(w_1^{\text{BQ}}, \ldots, w_n^{\text{BQ}}) &= \operatorname*{arg\,min}_{w_1,\ldots,w_n \in \mathbb{R}} e(Q_{\mathcal{X}}) \\ &= \operatorname*{arg\,min}_{w_1,\ldots,w_n \in \mathbb{R}} \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \int_\Omega f \,\mathrm{d}\mu - \sum_{i=1}^n w_i f(x_i) \right|\end{aligned}$$

and it also follows that the (root) integral posterior variance of a Bayesian quadrature rule is equal to the WCE of the rule: $\sqrt{V_{\text{BQ}}} = e(Q_{\text{BQ}})$.

3. CLASSICAL QUADRATURE RULES VIA GAUSSIAN PROCESSES

Let $p \geq 0$ and $\varphi_0, \ldots, \varphi_{p-1}$ with $\deg \varphi_i = i$ be linearly independent polynomials that form a basis of $\mathbb{P}_{p-1}$, the space of polynomials of degree at most $p-1$. We define k^p, the *polynomial kernel* of *degree* p (not to be confused with some other kernels with similar names in the literature), as

$$k^p(x, x') = \sum_{i=0}^{p-1} \varphi_i(x)\varphi_i(x'). \tag{5}$$

Given $n \leq p$ distinct points $x_1, \ldots, x_n$, the kernel matrix can be written as

$$\mathbf{K} = \mathbf{\Phi}^\mathsf{T}\mathbf{\Phi} + \sum_{i=n}^{p-1} \varphi_i(\mathcal{X})\varphi_i(\mathcal{X})^\mathsf{T},$$

where $[\varphi_i(\mathcal{X})]_j = \varphi_i(x_j)$ and $[\mathbf{\Phi}]_{ij} = \varphi_{i-1}(x_j)$ is a (generalised) Vandermonde matrix. Because $\mathbf{\Phi}$ is non-singular by the assumption on linear independency of φ_i, the product $\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}$ is positive-definite and, as rest of the terms are positive-semidefinite, the whole kernel matrix is positive-definite. However, note that k^p is not a positive-definite kernel in the strict sense of the definition because the kernel matrix is positive-definite only if the number of points does not exceed p.

The finite-dimensional RKHS induced by k^p is $\mathbb{P}_{p-1}$. It is easy to determine the inner product of this Hilbert space. For

$$f = \sum_{i=0}^{p-1} f_i\varphi_i \in \operatorname{span}(\varphi_0, \ldots, \varphi_{p-1}) = \mathbb{P}_{p-1},$$

$$g = \sum_{i=0}^{p-1} g_i\varphi_i \in \mathbb{P}_{p-1},$$

define the inner product $\langle f, g\rangle = \sum_{i=0}^{p-1} f_i g_i$. Then, since $k^p(\cdot, x) \in \mathbb{P}_{p-1}$ for any $x \in \Omega$, we have

$$\langle f, k^p(\cdot, x)\rangle = \sum_{i=0}^{p-1} f_i\varphi_i(x) = f(x),$$

which is exactly the reproducing property and consequently $\langle\cdot,\cdot\rangle_{\mathcal{H}} = \langle\cdot,\cdot\rangle$. Note that φ_i are then orthonormal. Furthermore, if these polynomials are orthonormal also under the regular $L^2(\mu)$ inner product $\langle f, g\rangle_{L^2(\mu)} = \int_\Omega fg \,\mathrm{d}\mu$, this inner product is that of the RKHS:

$$\begin{aligned}\langle f, k^p(\cdot, x)\rangle_{L^2(\mu)} &= \sum_{i=0}^{p-1}\sum_{j=0}^{p-1} f_j\varphi_i(x) \int_\Omega \varphi_i\varphi_j \,\mathrm{d}\mu \\ &= \sum_{i=0}^{p-1} f_i\varphi_i(x) \int_\Omega \varphi_i^2 \,\mathrm{d}\mu \\ &= f(x).\end{aligned}$$

We say that a Bayesian quadrature rule *coincides with* a classical quadrature rule if the rules have the same nodes and weights. We can now state and prove the main result of this article that establishes a necessary and sufficient condition for Bayesian quadrature with the polynomial kernel (5) constructed out of orthogonal polynomials to coincide with a classical quadrature rule. We are confident that the orthogonality assumption is not necessary but have not been able to furnish a proof that does not make use of this fact.

Theorem 4. *Let $\varphi_0, \ldots, \varphi_{p-1}$ be the orthogonal polynomials. Then the Bayesian quadrature rule with the kernel k^p coincides with a classical quadrature rule of degree $m-1$ if and only if $n \leq p \leq m$. For such a rule, $V_{\text{BQ}} = 0$.*

Proof. Suppose first that $p \leq m$. Let $x_1, \ldots, x_n$ be the nodes of a classical quadrature rule of degree $m-1$. Because $\mathcal{H} = \mathbb{P}_{p-1} \subset \mathbb{P}_{m-1}$, this classical rule has zero worst-case error. But, as the weights of the Bayesian quadrature rule minimise the WCE for given nodes and the weights of a classical quadrature rule are uniquely determined by the nodes, it follows that these two rules must coincide.

Suppose then that $p > m$ and that the BQ rule coincides with the classical one. From Equation (4) we have

$$\begin{aligned}\sqrt{V_{\text{BQ}}} &= \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mu(f) - Q_{\text{BQ}}(f)| \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \sum_{i=0}^{p-1} f_i \big[\mu(\varphi_i) - Q_{\text{BQ}}(\varphi_i)\big] \right| \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \sum_{i=m}^{p-1} f_i \big[\mu(\varphi_i) - Q_{\text{BQ}}(\varphi_i)\big] \right| \\ &\geq \big|\mu(\varphi_m) - Q_{\text{BQ}}(\varphi_m)\big| \\ &> 0,\end{aligned}$$

where $f = \sum_{i=0}^{p-1} f_i \varphi_i$, and the strict inequality follows from that fact that $\mu(\varphi_m) - Q_{\text{BQ}}(\varphi_m) \neq 0$ by the definition of degree of a classical quadrature rule and the assumption on the BQ rule coinciding with the classical one. However, Equation (3) implies

$$V_{\text{BQ}} = \sum_{i=m}^{p-1} \mu(\varphi_i)\big[\mu(\varphi_i) - Q_{\text{BQ}}(\varphi_i)\big],$$

but this must be zero because a consequence of orthogonality of φ_i (as noted in Section 2.1) is that $\mu(\varphi_i) = 0$ for $i > 1$. Thus the BQ rule must be different from the classical one. □

When $p < m = 2n$, there are in general multiple optimal Bayesian quadrature rules. For $p = m = 2n$, uniqueness of Gaussian quadrature rules results in the following corollary.

Corollary 5. *When φ_i are the orthogonal polynomials there is a unique n-point optimal Bayesian quadrature rule for the kernel k^{2n}. This rule is the Gaussian quadrature rule for the measure μ.*

4. ON MULTIVARIATE GENERALISATIONS

It is possible to generalise Theorem 4 to *cubature rules* that are just multivariate versions of quadrature rules:

$$Q(f) \coloneqq \sum_{i=1}^{n} w_i f(\mathbf{x}_i) \approx \int_\Omega f \,\mathrm{d}\mu,$$

where $\Omega \subset \mathbb{R}^d$, μ is a probability measure on Ω, f is multivariate function from Ω to $\mathbb{R}$, and the nodes $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \Omega$ are vectors. Classical and Bayesian cubature rules can be defined completely analogously to the univariate versions reviewed in Section 2. Due to the lack of uniqueness results for approximation with multivariate polynomials (see the Mairhuber–Curtis theorem in e.g. [22, Section 2.1]) the multivariate version of Theorem 4 that can be straightforwardly obtained is weaker.

The space of d-variate polynomials of degree at most m is denoted by $\mathbb{P}_m^d$. The dimension N_m^d of this space is

$$N_m^d \coloneqq \dim \mathbb{P}_m^d = \binom{m+d}{m} = \frac{(m+d)!}{m!\,d!}. \tag{6}$$

Analogously to the univariate case, a classical cubature rule is said to be of degree m if it is exact for all polynomials in $\mathbb{P}_m^d$. Unfortunately, in higher dimensions not all node sets result in unique weights. A classical cubature rule is said to be *interpolatory* if its weights are uniquely determined by the nodes (see [15, Section 6.1]). A rule of degree m must satisfy $Q(\varphi_i) = \mu(\varphi_i)$ for all $\varphi_1, \ldots, \varphi_{N_m^d}$ that form a basis of $\mathbb{P}_m^d$. That is, the nodes $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of an interpolatory rule are such that the linear system of equations

$$\begin{pmatrix} \varphi_1(\mathbf{x}_1) & \cdots & \varphi_1(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \varphi_{N_m^d}(\mathbf{x}_1) & \cdots & \varphi_{N_m^d}(\mathbf{x}_n) \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \mu(\varphi_1) \\ \vdots \\ \mu(\varphi_{N_m^d}) \end{pmatrix}$$

has a unique solution. In contrast to the univariate case, this system is, as evident from Equation (6), indeed overdetermined whenever $d > 1$ even if $m = n - 1$. We can now formulate a somewhat weaker but multivariate version of Theorem 4. The proof is essentially the same and thus not presented. Note that orthogonality of φ_i is not required due to weaker nature of the result.

Theorem 6. *Consider an interpolatory classical cubature rule of degree m and a Bayesian one with the same nodes. If the kernel is*

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{N_p^d} \varphi_i(\mathbf{x})\varphi_i(\mathbf{x}'), \tag{7}$$

where $p = m$ and the polynomials φ_i form a basis of $\mathbb{P}_p^d$, the Bayesian and the classical rule coincide. The integral posterior variance of this Bayesian cubature rule is zero.

This theorem extends the results of Särkkä *et al.* [10, Section IV] who considered the special cases of certain fully symmetric rules of McNamee and Stenger [23] (see [21] for general probabilistic versions of these rules) and Gauss–Hermite product rules. Much more could probably be proved but we do not pursue the issue further here. Instead, we end this section be pointing out an interesting connection to multivariate versions of Gaussian quadrature rules.

The following theorem, originally due to Mysovskikh [13] (see also [15, Section 7.2]), provides an intriguing connection between polynomially optimal cubature rules (in effect, multivariate versions of the Gaussian rules of Theorem 3) and multivariate polynomial kernels of the form (7). Multivariate orthogonal polynomials are defined completely analogously to the univariate ones, though one should keep in mind that they are not unique in higher dimensions.

Theorem 7. *Let the multivariate polynomials $\varphi_1, \ldots, \varphi_{N_m^d}$ form an orthonormal basis of $\mathbb{P}_m^d$. The nodes $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are those of a classical cubature rule that is exact for $\mathbb{P}_{2m}^d$ if and only if the kernel matrix $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for the kernel*

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{N_m^d} \varphi_i(\mathbf{x})\varphi_i(\mathbf{x}')$$

is diagonal and the diagonal elements are non-zero.

5. SOME NUMERICAL EXAMPLES

In this section we illustrate the theoretical results of Section 3 as well as the counterintuitive fact that a Gaussian process with non-zero posterior variance can produce a Bayesian quadrature rule with zero posterior integral variance. We work on the interval $[-1, 1]$ and use the uniform probability measure. That is, $\Omega = [-1, 1]$ and $\mathrm{d}\mu = \frac{1}{2}\,\mathrm{d}x$. We construct the kernel k^p, defined in Equation (5), from normalised Legendre polynomials that are orthonormal for this selection of Ω and μ. The

Legendre polynomials L_i, $i \geq 0$, can be defined via Rodrigues' formula as

$$\mathrm{L}_i(x) = \frac{1}{2^i i!} \frac{\mathrm{d}^i}{\mathrm{d}x^i} \left[(x^2 - 1)^i \right] \tag{8}$$

and they satisfy

$$\frac{1}{2} \int_{-1}^{1} \mathrm{L}_i(x) \mathrm{L}_j(x) \, \mathrm{d}x = \frac{1}{2i+1} \delta_{ij}.$$

For Gaussian process regression and Bayesian quadrature, we set $n = 4$ and select the four nodes as the roots of L_4 so that the corresponding classical rule is Gaussian (recall Theorem 3). Then Theorem 4 guarantees that, for $p = 4, \ldots, 8$, the kernels k^p with $\varphi_i = \mathrm{L}_i$ yield Bayesian quadrature rules that coincide with the Gaussian one. The posterior processes for $p = 5, \ldots, 8$ (for $p = 4$ the posterior variance would vanish) are depicted in Figure 1 for some rather arbitrarily picked data values (i.e. integrand evaluations). What is interesting and most likely slightly counterintuitive is that the posterior processes have non-zero variance but, nevertheless, the Bayesian quadrature integral posterior variances are zero in accordance with Theorem 4.

6. CONCLUSIONS AND DISCUSSION

We have shown in Section 3 that any classical quadrature rule can be interpreted as a Bayesian one if the kernel is an orthogonal polynomial kernel of the form (5) with suitable degree p. Perhaps the most interesting consequence is the correspondence of Gaussian quadrature rules to optimal Bayesian ones for the selection $p = 2n$. We have also sketched a multivariate extension of this result and pointed out some interesting connections to polynomially optimal classical cubature rules. Because the resulting probabilistic quadrature rules have zero posterior integral variance, they cannot be used in modelling uncertainty of numerical approximations and are therefore of little practical interest.

We envision that at least the following four slight generalisations and extensions are possible:

— The requirement of orthogonality in Theorem 4 and Corollary 5 can most likely be dispensed of.

— It is probably possible to do more in the multivariate setting than what is done in Section 4.

— Most of the results can be retained even if the functions φ_i in the definition of the kernel k^p are replaced with other linearly independent functions than polynomials.

— If also derivative evaluations are used (see [24] for this in BQ setting and [22, Chapter 16] for a more general account of the topic), connections to Gauss–Turán rules [25] (see also [16, Section 3.1.3]) should arise.

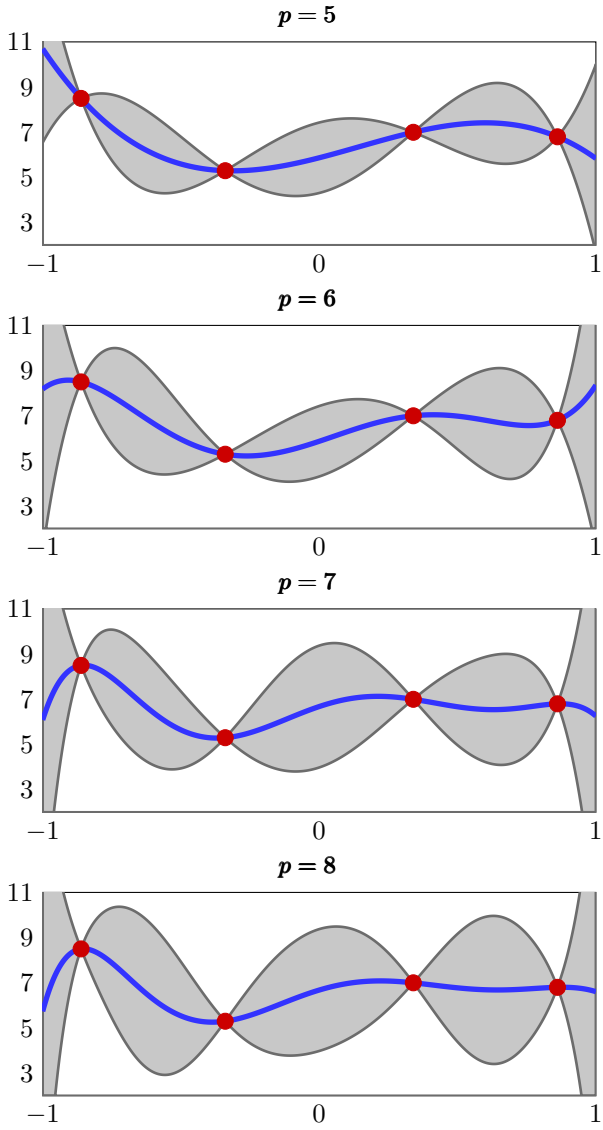

Fig. 1: Posterior processes corresponding to GP priors with the kernel k^p for $p = 5, \ldots, 8$. The polynomials φ_i are the Legendre polynomials (8). The data consists of the evaluations 8.5, 5.3, 7, and 6.8 at the roots of L_4 that are roughly ± 0.8611 and ± 0.34. Posterior mean is blue, the data points red, and the shaded area around the mean is the 95% confidence interval of the posterior.

7. REFERENCES

[1] A. O'Hagan, "Bayes–Hermite quadrature," *Journal of Statistical Planning and Inference*, vol. 29, no. 3, pp. 245–260, 1991.

[2] F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic, "Probabilistic integration: A role for statisticians in numerical analysis?," 2016, arXiv:1512.00933.

[3] F. M. Larkin, "Optimal approximation in Hilbert spaces with reproducing kernel functions," *Mathematics of Computation*, vol. 24, no. 112, pp. 911–921, 1970.

[4] F. M. Larkin, "Gaussian measure in Hilbert space and applications in numerical analysis," *Rocky Mountain Journal of Mathematics*, vol. 2, no. 3, pp. 379–422, 1972.

[5] A. O'Hagan, "Some Bayesian numerical analysis," *Bayesian Statistics*, vol. 4, pp. 345–363, 1992.

[6] P. Hennig, M. A. Osborne, and M. Girolami, "Probabilistic numerics and uncertainty in computations," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 471, no. 2179, 2015.

[7] J. Cockayne, C. Oates, T. Sullivan, and M. Girolami, "Bayesian probabilistic numerical methods," 2017, arXiv:1702.03673.

[8] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[9] P. Diaconis, "Bayesian numerical analysis," in *Statistical Decision Theory and Related Topics IV*, pp. 163–175. 1988.

[10] S. Särkkä, J. Hartikainen, L. Svensson, and F. Sandblom, "On the relation between Gaussian process quadratures and sigma-point methods," *Journal of Advances in Information Fusion*, vol. 11, no. 1, pp. 31–46, 2016.

[11] M. Schober, D. Duvenaud, and P. Hennig, "Probabilistic ODE solvers with Runge-Kutta means," in *Advances in Neural Information Processing Systems*, 2014, vol. 27, pp. 739–747.

[12] O. Teymur, K. Zygalakis, and B. Calderhead, "Probabilistic linear multistep methods," in *Advances in Neural Information Processing Systems*, 2016, vol. 29, pp. 4321–4328.

[13] I. P. Mysovskikh, "On the construction of cubature formulas with fewest nodes," *Soviet Mathematics Doklady*, vol. 9, pp. 277–280, 1968.

[14] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, 2nd edition, 1984.

[15] R. Cools, "Constructing cubature formulae: the science behind the art," *Acta Numerica*, vol. 6, pp. 1–54, 1997.

[16] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Numerical Mathematics and Scientific Computation. Oxford University Press, 2004.

[17] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[18] T. Minka, "Deriving quadrature rules from Gaussian processes," Tech. Rep., Statistics Department, Carnegie Mellon University, 2000.

[19] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[20] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, 2004.

[21] T. Karvonen and S. Särkkä, "Fully symmetric kernel quadrature," 2017, arXiv:1703.06359.

[22] H. Wendland, *Scattered Data Approximation*, Number 17 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2010.

[23] J. McNamee and F. Stenger, "Construction of fully symmetric numerical integration formulas," *Numerische Mathematik*, vol. 10, no. 4, pp. 327–344, 1967.

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

[25] P. Turán, "On the theory of mechanical quadrature," *Acta Scientiarum Mathematicarum*, vol. 12, pp. 30–37, 1950.