Control Functionals for Quasi-Monte Carlo Integration

Chris. J. Oates
University of Technology Sydney and

Mark Girolami
University of Warwick and
The Alan Turing Institute for Data Science

Abstract

Quasi-Monte Carlo (QMC) methods are being adopted in statistical applications due to the increasingly challenging nature of numerical integrals that are now routinely encountered. For integrands with d-dimensions and derivatives of order α, an optimal QMC rule converges at a best-possible rate $O(N^{-\alpha/d})$. However, in applications the value of α can be unknown and/or a rate-optimal QMC rule can be unavailable. Standard practice is to employ α_L-optimal QMC where the lower bound $\alpha_L \leq \alpha$ is known, but in general this does not exploit the full power of QMC. One solution is to trade-off numerical integration with functional approximation. This strategy is explored herein and shown to be well-suited to modern statistical computation. A challenging application to robotic arm data demonstrates a substantial variance reduction in predictions for mechanical torques.

1 Introduction

Consider a Lebesgue-integrable test function $f : \mathcal{X} \to \mathbb{R}$ defined on a bounded measurable subspace $\mathcal{X} \subseteq \mathbb{R}^d$ ($d \in \mathbb{N}$) with square integrable derivatives of order $\alpha > 0$ in each variable. Our focus is numerical computation of the integral $I[f] := \int_{\mathcal{X}} f(\boldsymbol{x}) d\boldsymbol{x}$. The Quasi-Monte Carlo (QMC) approach is based on an approximation

$$Q[f; \boldsymbol{x}^{1:N}] := \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}^n)$$

where the (possibly random) design points $\boldsymbol{x}^{1:N} = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\} \subset \mathcal{X}$ have low discrepancy; that is, the points are 'well-spaced' in a precise sense defined below. This contrasts with the Monte Carlo (MC) approach whereby the design points are sampled independently from a uniform distribution over $\mathcal{X}$. MC integration achieves a root mean square error (RMSE) convergence rate of $O(N^{-1/2})$ whereas QMC integration can in principle achieve a rate $O(N^{-\alpha/d})$ on specific geometric sequences $\{N_n\}_{n=1}^{\infty}$ [24]. It is known that this rate is best-possible [19] and explicit algorithms to generate design points that attain this rate are now available for many (but not all) values of α [6]. Challenging integration problems are common in contemporary statistics, for example when computing expectations, marginal probability densities or normalising constants, and QMC methods are therefore gaining importance in statistical applications [12, 17, 35].

Contrary to the above theoretical considerations, rate-optimal QMC is often not employed in practice. This is mainly due to three reasons; either (R1) the smoothness parameter α is unknown, (R2) there does not currently exist an explicit QMC rule that is rate-optimal for functions of smoothness α, or (R3) it is simply more convenient to employ a basic QMC rule based on a weaker smoothness assumption $\alpha_L < \alpha$, as implemented in standard software. In each situation there is a gap between theory and practice that, as we show in this paper, can be bridged using functional approximation.

Previous work on variance reduction techniques for QMC includes [1], who considered modified importance sampling strategies, and [14], who considered constructing control variates for QMC. Neither approach improved the asymptotic error rate, though in some cases the QMC error was reduced by a constant factor. Interestingly, [14] reports some quite negative results for control variate strategies in this setting, because the objective being minimised by QMC is not equivalent to the MC variance that is minimised by control variates. [33] demonstrates variance reductions in QMC are possible using additive approximations, though again the asymptotics were unchanged.

This paper studies a general approach to variance re-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

duction for QMC rules, building on kernel methods and recent work in the Monte Carlo setting due to [22, 31]. The mathematics that underpins our work comes from the functional approximation literature. This takes the form of a 'control functional' $\psi : \mathcal{X} \to \mathbb{R}$ that satisfies (i) ψ integrates to zero, (ii) $f - \psi$ is more amenable to QMC methods than f, in a precise sense. The general approach that we explore is to replace the integrand f by $f - \psi$ and target the QMC objective directly. This can lead to accelerated asymptotics. The main contribution of this paper is to explore this strategy in the settings (R1-3) above. Theoretical analysis of convergence rates is provided, along with empirical results and a challenging application to robotics. We begin by presenting some background on QMC theory below, before describing the methodology in more detail.

2 Background

QMC is naturally studied in reproducing kernel Hilbert spaces (RKHS; [8]). Below we draw connections with kernel methods, that are themselves naturally studied in RKHS.

Notation. We work in a Hilbert space H, consisting of measurable functions $f : \mathcal{X} \to \mathbb{R}$. For simplicity of presentation we assume H includes the constant functions. We follow the mainstream QMC literature by taking $\mathcal{X} = [0,1]^d$, equipped with the Euclidean norm $\|\boldsymbol{x}\| := (\sum_{i=1}^d x_i^2)^{1/2}$. Denote the scalar product and norm on H by $\langle \cdot, \cdot \rangle_H$ and $\| \cdot \|_H$ respectively. Suppose further that H is a RKHS with kernel $K : [0,1]^d \times [0,1]^d \to \mathbb{R}$; that is, K satisfies (i) $K(\cdot, \boldsymbol{x}) \in H$ for all $\boldsymbol{x} \in [0,1]^d$ and (ii) $f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x}) \rangle_H$ for all $f \in H$ and all $\boldsymbol{x} \in [0,1]^d$. K is assumed to be non-trivial, i.e. $K \neq 0$.

Quadrature Error Analysis. The quadrature methods that we focus on aim to minimise the 'worst case' integration error which, for design points $\boldsymbol{x}^{1:N}$ and Hilbert space H, is defined to be

$$e_H(\boldsymbol{x}^{1:N}) := \sup_{\|f\|_H \leq 1} \left| Q[f; \boldsymbol{x}^{1:N}] - I[f] \right| \tag{1}$$

where the supremum is taken over all test functions f belonging to the unit ball in H. It follows from linearity that, for any function $f \in H$, the integration error obeys

$$\left| Q[f; \boldsymbol{x}^{1:N}] - I[f] \right| \leq e_H(\boldsymbol{x}^{1:N}) \|f\|_H. \tag{2}$$

The worst case error $e_H(\boldsymbol{x}^{1:N})$ is the usual target of QMC innovation, with $\boldsymbol{x}^{1:N}$ chosen to (approximately, asymptotically) minimise $e_H(\boldsymbol{x}^{1:N})$ [8]. Note that Eqn. 1 is also the 'maximum mean discrepancy' (MMD), as studied extensively in the kernel methods literature [4, 30].

Quadrature is naturally studied in RKHS because there exists a closed-form expression for the worst case error in terms of the kernel K, which facilitates the principled selection of design points [8]:

$$\begin{aligned} e_H(\boldsymbol{x}^{1:N})^2 &= \int\int_{[0,1]^d} K(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \\ &\quad - \frac{2}{N} \sum_{n=1}^N \int_{[0,1]^d} K(\boldsymbol{x}^n, \boldsymbol{y}) d\boldsymbol{y} \\ &\quad + \frac{1}{N^2} \sum_{m,n=1}^N K(\boldsymbol{x}^n, \boldsymbol{x}^m) \end{aligned} \tag{3}$$

The mainstream QMC literature supposes H is a Sobolev space of known order α (defined below). In this setting, $O(N^{-\alpha/d})$ is the best-possible rate for the worst case error when $\boldsymbol{x}^{1:N}$ are chosen deterministically and $O(N^{-\alpha/d-1/2})$ is the best-possible RMSE when $\boldsymbol{x}^{1:N}$ are allowed to be random [19]. We will refer to QMC rules that achieve these optimal rates as 'α-QMC rules'.

This paper focuses on improving performance in the situation where a (sub-optimal) α_L-QMC rule is used to integrate a test function of smoothness $\alpha > \alpha_L$. For reasons (R1-3), this scenario is commonly encountered in statistical applications. In contrast to QMC [8] (and kernel methods that aim to minimise the MMD [2]), the rate constant $\|f\|_H$ is the primary target of our methodology below.

3 Methodology

Control Functionals for QMC. The approach that we pursue in this paper aims to construct a Lebesgue-integrable functional $\psi : [0,1]^d \to \mathbb{R}$ that satisfies

$$I[\psi] = 0. \tag{4}$$

When $\boldsymbol{x}$ has the interpretation of a random variable, $\psi(\boldsymbol{x})$ is classically known as a 'control variate' [14]. When ψ itself is estimated, we follow [22] and refer to the entire mapping ψ as a 'control functional' (CF). In the CF approach to estimation, the test function f is replaced by $f - \psi$; it is hoped that the latter is more amenable to numerical integration. Clearly $I[f - \psi] = I[f]$. In this paper we construct a CF ψ_N based on a tractable approximation f_N to f. (The dependence on N will be explained below.) It is required that the integral $I[f_N]$ is available in closed-form. We then set

$$\psi_N(\boldsymbol{x}) = f_N(\boldsymbol{x}) - I[f_N] \tag{5}$$

so that ψ_N satisfies Eqn. 4. For this to make sense mathematically, it must be the case that $f_N \in H$ and

this informs our method of approximation (the constant function with value $I[f_N]$ belongs to H by assumption). Intuitively, a good CF ψ_N will provide a close approximation to fluctuations of the test function f, so that the functional difference $f - \psi_N$ become increasingly 'flat' and thus more amenable to QMC methods. More precisely, motivated by Eqn. 2 we aim to construct a CF such that $\|f - \psi_N\|_H < \|f\|_H$. This connection with functional approximation offers the possibility to leverage kernel methods for these problems, see e.g. [27, 31].

Control Functional Error Analysis. Consider partitioning $\boldsymbol{x}^{1:N}$ into two sets $\boldsymbol{u}^{1:M}$ and $\boldsymbol{v}^{M+1:N}$ where $1 < M < N$ and $M/N \to c \in (0,1)$ as $N \to \infty$. The first set $\boldsymbol{u}^{1:M}$, possibly non-random, will be used in a preliminary step to construct an approximation $f_M(\cdot; \boldsymbol{u}^{1:M})$ to f. Then the second set $\boldsymbol{v}^{M+1:N}$, possibly random, is used to evaluate the 'CF estimator'

$$\begin{aligned} E[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N}] &:= Q[f - \psi_N(\cdot; \boldsymbol{u}^{1:M}); \boldsymbol{v}^{M+1:N}] \\ &= Q[f - f_M(\cdot; \boldsymbol{u}^{1:M}); \boldsymbol{v}^{M+1:N}] \\ &\quad + I[f_M(\cdot; \boldsymbol{u}^{1:M})]. \end{aligned} \quad (6)$$

We remark that if the points $\boldsymbol{v}^n$ are random and marginally distributed as $U([0,1]^d)$ then $E[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N}]$ will be an unbiased estimator for $I[f]$. Error analysis for the CF estimator is based on the following:

Theorem 1. *Given $f, f_M \in H$, we have*

$$\begin{aligned} &|E[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N}] - I[f]| \\ &\qquad \leq e_H(\boldsymbol{v}^{M+1:N}) \|f - f_M(\cdot; \boldsymbol{u}^{1:M})\|_H. \end{aligned} \quad (7)$$

Proof. Since $f, f_M \in H$ we have that $f - f_M \in H$. The result then follows by applying the fundamental inequality from Eqn. 2 to the function $f - f_M$ and using linearity of the integral operator I. □

Thus the CF methodology produces an estimator $E[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N}]$ that has asymptotically zero error relative to standard QMC estimators, providing that it is possible to construct an approximation f_M to f in such a way that $\|f - f_M(\cdot; \boldsymbol{u}^{1:M})\|_H \to 0$ as $M \to \infty$. The next sections establish convergence rates for functional approximation using kernel methods.

Sobolev Spaces. To achieve consistent approximation $\|f - f_M\|_H \to 0$ it is necessary to impose regularity conditions on H. Sobolev spaces are a general setting in which to formulate such regularity assumptions; our main reference here is [27]. Firstly suppose that $k \in \mathbb{N}_0$, $k > d/2$ and $1 \leq p < \infty$. For a multi-index $\boldsymbol{a} \in \mathbb{N}_0^d$ we write $|\boldsymbol{a}| = a_1 + \cdots + a_d$. Define the '$p$-Sobolev space of order k' to be

$$\begin{aligned} W^{k,p} := \{&f : [0,1]^d \to \mathbb{R} \mid D^{\boldsymbol{a}} f \text{ exists and} \\ &D^{\boldsymbol{a}} f \in L_p([0,1]^d), \forall \boldsymbol{a} \in \mathbb{N}_0^d \text{ with } |\boldsymbol{a}| \leq k\}. \end{aligned}$$

Here $D^{\boldsymbol{a}} f$ denotes the weak (or 'distributional') derivative of f; the reader is referred to the above reference for details. Clearly $W^{k,p}$ is a vector space over $\mathbb{R}$ when addition and (scalar) multiplication are defined point-wise. For the special case $p = 2$ we equip $W^{k,2}$ with the inner product

$$\langle f, g \rangle_k := \sum_{\boldsymbol{a} \in \mathbb{N}_0^d, |\boldsymbol{a}| \leq k} I[D^{\boldsymbol{a}} f D^{\boldsymbol{a}} g]$$

and denote this inner-product space $H^k := (W^{k,2}, \langle \cdot, \cdot \rangle_k)$. Defined in this way, H^k is a Hilbert space of functions whose (weak) derivatives exist up to order k. Moreover H^k can be made into a RKHS with an appropriate choice of kernel (see below). Our results below apply also to Sobolev spaces with non-integer k, but this construction is more technical and we refer the reader to [27] for details.

Approximation in Sobolev Spaces. Our assumptions are naturally stated using Sobolev spaces: Given two Hilbert spaces H, H', defined on the same element set, with norms $\|\cdot\|_H$, $\|\cdot\|_{H'}$, we say that H and H' are 'norm-equivalent', written $H \equiv H'$, whenever there exist positive constants c_1, c_2 such that $c_1 \|f\|_H \leq \|f\|_{H'} \leq c_2 \|f\|_H$ for all $f \in H$.

Assumption 1: $H \equiv H^{\alpha_L}$ where $\alpha_L > d/2$.

Assumption 2: $f \in H^{\alpha}$ where $\alpha \geq \alpha_L$.

Assumption 1 is a technical requirement to ensure the space H (where QMC is performed) admits consistent functional approximation. Assumption 2 ensures that the test function f is 'smooth enough' for α_L-QMC methods to converge at the α_L-rate. This follows from the fact that Sobolev spaces are nested, so that $f \in H^{\alpha} \implies f \in H^{\alpha_L}$.

For consistent approximation of f it is necessary to base our approximation f_M in a space H_* of functions that are 'at least as smooth' as f:

Assumption 3: $H_* \equiv H^{\alpha_U}$ where $\alpha_U \geq \alpha$.

It follows again from the nested property that $f_M \in H^{\alpha_L}$ and thus the functional difference $f - f_M$ exists in H^{α_L}. The Sobolev spaces H_* can be characterised as RKHS via an appropriate reproducing kernel K_*, such as the well-known Matérn kernel.

Finally an approximation f_M to f is constructed based on the points $\boldsymbol{u}^{1:M}$ as follows:

$$f_M(\boldsymbol{x}; \boldsymbol{u}^{1:M}) := \sum_{n=1}^{M} \beta_n K_*(\boldsymbol{x}, \boldsymbol{u}^n) \quad (8)$$

where the weights $\beta_n \in \mathbb{R}$ are defined as the solution to the linear system of interpolation equations

$$f_M(\boldsymbol{u}^n; \boldsymbol{u}^{1:M}) = f(\boldsymbol{u}^n), \quad n = 1, \ldots, M. \tag{9}$$

It is well-known that Eqn. 8 is the unique minimiser of the H_*-norm under all functions in H_* that satisfy the linear system in Eqn. 9 [27]. In practice it may be necessary to regularise the linear system in order to facilitate inversion, but we do not go into details here, see e.g. [27].

We note that $I[f_M]$ will *not* have a closed-form expression when the Matérn kernel is employed and for this technical reason we instead employ tensor products of polynomial kernels (these give rise to Sobolev spaces of mixed dominating smoothness - full details are provided at the end of this section).

Theory: Deterministic Case. We begin by considering the case where the design points $\boldsymbol{v}^{M+1:N}$ are chosen deterministically. Define the 'fill distance'

$$h(\boldsymbol{u}^{1:M}) := \sup_{\boldsymbol{x} \in [0,1]^d} \min_n \|\boldsymbol{x} - \boldsymbol{u}^n\|,$$

the 'separation radius'

$$q(\boldsymbol{u}^{1:M}) := \frac{1}{2} \min_{j \neq k} \|\boldsymbol{u}^j - \boldsymbol{u}^k\|$$

and the 'mesh ratio' $\rho(\boldsymbol{u}^{1:M}) := h(\boldsymbol{u}^{1:M})/q(\boldsymbol{u}^{1:M})$. The set $\boldsymbol{u}^{1:M}$ is called 'quasi-uniform' if $\rho(\boldsymbol{u}^{1:M}) \to 1$ as $M \to \infty$.

Theorem 2. *Under Assumptions 1-3 the CF estimator has error bounded by*

$$|E[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N}] - I[f]| \leq C e_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N}) h(\boldsymbol{u}^{1:M})^{\alpha - \alpha_L} \rho(\boldsymbol{u}^{1:M})^{\alpha_U - \alpha_L} \|f\|_{H^\alpha}$$

where $C > 0$ is a constant that depends on α, α_L and α_U but not on f, $\boldsymbol{v}^{M+1:N}$ and $\boldsymbol{u}^{1:M}$.

Proof. From [27] (Theorem 7.8) we have that the kernel estimator in Eqn. 8 is consistent for the non-parametric regression problem at a rate

$$\|f - f_M(\cdot; \boldsymbol{u}^{1:M})\|_{H^{\alpha_L}} \leq C h(\boldsymbol{u}^{1:M})^{\alpha - \alpha_L} \rho(\boldsymbol{u}^{1:M})^{\alpha_U - \alpha_L} \|f\|_{H^\alpha}$$

where C depends only on $\alpha, \alpha_L, \alpha_U$. Combining this with Eqn. 7 completes the proof. □

For quasi-uniform $\boldsymbol{u}^{1:M}$, there is no asymptotic penalty from employing a kernel K_* that imposes 'too much smoothness' on the approximation f_M, with $\rho \to 1$. In this case $h(\boldsymbol{u}^{1:M}) = O(M^{-1/d})$ and, since M and N are proportional, $h(\boldsymbol{u}^{1:M}) = O(N^{-1/d})$. However the rate constant C will increase when too much smoothness is assumed so that, as a rule of thumb, we should try to select α_U as close as possible to α. Our main result is stated below:

Corollary 1. *When $\boldsymbol{u}^{1:M}$ is quasi-uniform, CFs accelerate α_L-QMC by a factor $O(N^{-(\alpha - \alpha_L)/d})$.*

Remark: The improvement due to CFs appears to be mainly limited to low-dimensional integrals (d small), but in fact CFs can in principle be extended to high-dimensional integrals under additional tractability assumptions, as discussed in Sec. 5.

Remark: Optimising the bound in Theorem 2 enables us to obtain the optimal scaling

$$\frac{M}{N} \to c^* = \frac{\alpha - \alpha_L}{\alpha},$$

see the Supplement for full details.

The overall convergence rate of the CF estimator depends on how the design points $\boldsymbol{v}^{M+1:N}$ are generated. For this there are many QMC methodologies available, each leading to different convergence rates for the worst case error $e_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N})$; see [7] for a recent survey of some of these approaches. Of particular interest in statistical applications is the case of random design points which we discuss below.

Theory: Randomised Case. Modern QMC methods begin with a deterministic set/sequence of design points (e.g. a Halton sequence or a Sobol sequence), then apply a random transformation leading to a low discrepancy set with high probability. Below we consider three types of randomisation; shifting, folding and scrambling.

Shifting: In 'random shift' QMC the design points $\boldsymbol{v}^{M+1:N}$ are translated by a common uniform random vector $\boldsymbol{\Delta} \in [0,1]^d$, so that $\boldsymbol{v}^n \mapsto \boldsymbol{v}^n + \boldsymbol{\Delta}$ for each $n = M+1, \ldots, N$. For convenience we write this 'shifted' set as $\boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}$. Applying Theorem 2 to $\boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}$ and then marginalising over $\boldsymbol{\Delta} \in [0,1]^d$ produces a RMSE bound for the CF estimator:

Corollary 2. *Under Assumptions 1-3 the random shift CF estimator has error bounded by*

$$\sqrt{\mathbb{E}|E[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}] - I[f]|^2} \leq C e^{sh}_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N}) h(\boldsymbol{u}^{1:M})^{\alpha - \alpha_L} \rho(\boldsymbol{u}^{1:M})^{\alpha_U - \alpha_L} \|f\|_{H^\alpha}$$

where

$$(e^{sh}_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N}))^2 := \int_{[0,1]^d} e_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta})^2 d\boldsymbol{\Delta}$$

and $C > 0$ is a constant that does not depend on f, $\boldsymbol{v}^{M+1:N}$ or $\boldsymbol{u}^{1:M}$.

For quasi-uniform $\boldsymbol{u}^{1:M}$, CFs accelerate random shift α_L-QMC by a factor $O(N^{-(\alpha-\alpha_L)/d})$ (compare against Sec. 5.2 of [7]).

Folding: A shifted and 'folded' QMC rule takes the form

$$Q_{\boldsymbol{b}}(f; \boldsymbol{z}^{1:N} + \boldsymbol{\Delta}) := \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{b}(\boldsymbol{z}^n + \boldsymbol{\Delta}))$$

where $\boldsymbol{b}$ is the 'baker's transformation', given by $b_i(\boldsymbol{t}) = 1 - |2t_i - 1|$. This transformation reduces error rates; for example, for $f \in SH^2([0,1]^d)$ (defined below), folding and shifting a uniform lattice $\boldsymbol{z}^{1:N}$ leads to a RMSE $O(N^{-2+\epsilon})$ that is smaller than the RMSE $O(N^{-1+\epsilon})$ for a shifted lattice (p. 59 of [7]). The CF estimator here is

$$E_{\boldsymbol{b}}[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}] \\ := I[f_M(\cdot; \boldsymbol{u}^{1:M})] + Q_{\boldsymbol{b}}[f - f_M(\cdot; \boldsymbol{u}^{1:M}); \boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}].$$

For convenience we denote the shifted and folded design points by $\boldsymbol{b}(\boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta})$. Applying Theorem 2 to $\boldsymbol{b}(\boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta})$ and then marginalising over $\boldsymbol{\Delta} \in [0,1]^d$ produces:

Corollary 3. *Under Assumptions 1-3 the shifted and folded CF estimator has error bounded by*

$$\sqrt{\mathbb{E}|E_{\boldsymbol{b}}[f; \boldsymbol{u}^{1:M}, \boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}] - I[f]|^2} \\ \leq C e^{sh,\boldsymbol{b}}_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N}) h(\boldsymbol{u}^{1:M})^{\alpha-\alpha_L} \rho(\boldsymbol{u}^{1:M})^{\alpha_U-\alpha_L} \|f\|_{H^\alpha}$$

where

$$(e^{sh,\boldsymbol{b}}_{H^{\alpha_L}}(\boldsymbol{v}^{M+1:N}))^2 := \int_{[0,1]^d} e_{H^{\alpha_L}}(\boldsymbol{b}(\boldsymbol{v}^{M+1:N} + \boldsymbol{\Delta}))^2 d\boldsymbol{\Delta}$$

and $C > 0$ is a constant independent of f, $\boldsymbol{v}^{M+1:N}$ and $\boldsymbol{u}^{1:M}$.

Again, for quasi-uniform $\boldsymbol{u}^{1:M}$, CFs accelerate shifted and folded α_L-QMC by a factor $O(N^{-(\alpha-\alpha_L)/d})$ (compare against Sec. 5.9 of [7]).

Scrambling: An explicit α-QMC rule that applies for all integer values of α was recently discovered by [6]. For simplicity focussing on $d = 1$, these random design points achieve α-rates and, moreover, the RMSE is controlled by a norm of the form $\|f\|_{H^\alpha}$. When α is known and is an integer, one may achieve optimal rates and CFs provide no rate improvement. However, when $\alpha \notin \mathbb{N}$, CFs can be used to transform these sub-optimal integrators into optimal integrators.

Choice of Kernel: The QMC+CF methodology has some flexibility in terms of the choice of kernel K_* that is used to construct the approximation f_M. Our main requirements here are: (i) K_* imposes 'enough smoothness' on f_M in order to be able to faithfully approximate f (Assumption 3). Moreover, K_* should be tunable to achieve a pre-specified minimum level of smoothness. Below we make an explicit connection between K_* and the order of the associated 'native' Sobolev space that will allow us to satisfy this requirement. (ii) The functions $K_*(\cdot, \boldsymbol{y})$ can be integrated analytically, so that $I[f_M]$ is available in closed form. This second requirement leads us to consider tensor products of Sobolev spaces, as described below.

To construct analytically integrable functional approximations we consider kernels that are given by polynomials. Wendland's compactly supported functions [34] are defined via the recursion

$$\varphi_{d,k} = \mathcal{I}^k[\varphi_{\lfloor d/2 \rfloor + k + 1}],$$

the base function $\varphi_\ell(r) = (1-r)^\ell_+$ with $x_+ := \max\{0, x\}$, and the integral operator

$$\mathcal{I}[\varphi](r) = \int_r^\infty t\varphi(t) dt$$

($r \geq 0$), so that

$$\varphi_{d,k}(r) = \begin{cases} (1-r)^{\ell+k} p_{d,k}(r), & r \in [0,1] \\ 0, & r > 1 \end{cases}$$

where $\ell = \lfloor d/2 \rfloor + k + 1$ and $p_{d,k}$ is a polynomial of degree k (see e.g. p.87 of [9] for explicit formulae). Then the kernel $K_*(\boldsymbol{x}, \boldsymbol{y}) = \varphi_{d,k}(\|\boldsymbol{x} - \boldsymbol{y}\|)$ has native space $H^{d/2+k+1/2}$ (where the restriction $d > 3$ is in principle required for the special case $k = 0$) (see e.g. p.109 of [9]). With this kernel we can therefore guarantee a minimum level of smoothness. By rescaling, the kernel's support can be changed from the unit ball (as above) to balls of smaller radius. This in turn enforces sparsity on the system of interpolation equations that are the basis of the CF estimator and reduces the computational cost of inverting this linear system.

Wendland's kernel cannot be integrated analytically in $d \geq 2$ dimensions, violating requirement (ii). However we can exploit recent work by [29] that shows the d-dimensional tensor product space $H^k([0,1]) \otimes \cdots \otimes H^k([0,1])$ is norm-equivalent to $SH^k = SH^k([0,1]^d)$, the Sobolev space with dominating mixed smoothness:

$$SH^k := \{f : [0,1]^d \to \mathbb{R} \mid D^{\boldsymbol{a}} f \text{ exists and } \\ D^{\boldsymbol{a}} f \in L_p([0,1]^d), \forall \boldsymbol{a} \in \mathbb{N}_0^d \text{ with } a_i \leq k\}.$$

(The distinction with $H^k([0,1]^d)$ is that the multi-index $\boldsymbol{a}$ is now constrained component-wise, $a_i \leq k$, rather than $|\boldsymbol{a}| \leq k$.) In particular $SH^k([0,1]^d) \subseteq H^k([0,1]^d)$ so that functions in SH^k are at least as smooth as functions in H^k. We therefore propose to employ the product kernel

$$K^{(k)}_*(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^d \varphi_{1,k}(|x_i - y_i|) \qquad (10)$$

whose native space is SH^{k+1}. The integral

$$\int_{[0,1]^d} K_*^{(k)}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x}$$

of tensor products of Wendland functions in Eqn. 10 can now be integrated analytically. This approach provides a convenient mechanism to control the degree of smoothness that we impose on the approximation f_M.

4 Experimental Results

Our methodology provides a variance reduction technique for QMC that is able to accelerate convergence rates, yet is also practical. The first numerical study below is a 'proof-of-principle' designed to validate this specific claim in the empirical setting.

Simulation Study: For objective assessment we exploited the test package proposed by [11]. This package defines 6 function families, each of them characterized by some peculiarity, such as oscillation, discontinuity or corner peaks, with the property that their exact integrals are available. The 'discontinuous' Genz function provides an example where smoothness assumptions on the test function are violated. We used the MATLAB implementation of [11] that is freely available at `http://people.sc.fsu.edu/~jburkardt/m_src/testpack/testpack.html`.

In the experiments below, we focus on the two QMC rules that are most widely used in practice. In the first experiment, the random QMC point set $\boldsymbol{v}^{M+1:N}$ was generated by truncating the Halton sequence, scrambling the digits of the resulting points using the reverse-radix algorithm [16] and applying a uniform random shift. This QMC rule achieves the $\alpha_L = 1$ rate on the subsequence $N_n = 2^n$ when the test function has mixed partial derivatives of first order. To ensure that these QMC rules were implemented faithfully, we restricted attention to the case where $M = N/2$ so that $N - M$ was always a power of two. The training points $\boldsymbol{u}^{1:M}$ were taken to be d-dimensional square lattices in all experiments.

We considered the 6 Genz functions in $d = 1, 2, 3, 4$ dimensions. The performance of QMC with and without CFs was compared, in each case ensuring that the total number of evaluations of the integrand f was equal for all methods. For CFs, the tensor-product Wendland kernel with $k = 1$ was employed (i.e. approximation with functions $f_M \in H^2$, so $\alpha_U = 2$). Results are presented in Fig. 1. (For clarity we chose not present results for MC, since these were inferior to QMC methods in all cases considered.) For the first 5 Genz functions it holds that $f \in H^\alpha$ with $\alpha = 2$ and theory (for the random case) guarantees an acceleration of $O(N^{-1/d})$; this is borne out in experimental results. In the 6th, discontinuous case the QMC+CF method does not out-perform QMC (at least in dimension $d > 1$), as the functional approximation f_M is poor due to violation of our continuity assumption. In all cases the performance of QMC+CF approaches that of QMC as the dimension d increased. In higher dimensions ($d \geq 5$, not shown) the QMC+CF and QMC estimators demonstrated effectively identical performance, in line with theory.

The experiments were then repeated with rougher ($k = 0$) and smoother ($k = 2$) regression kernels. Results in the Supplement (Figs. S3-8) demonstrated a slight improvement in the performance of QMC+CF when $k = 2$, in line with theory, though generally estimates were robust to the choice of regression kernel. To further assess the generality of these conclusions, further experiments were performed using a different QMC rule (truncated Sobol sequence with scrambling due to [18]). Results in the Supplement showed that the same conclusions can be drawn in each case. Taken together, these results demonstrate that CFs can accelerate QMC, at least in low-dimensional settings, and thus complete our 'proof-of-principle'. MATLAB code to reproduce these results is provided.

Application to Robot Arm Data: To demonstrate the benefits of our methodology we consider the problem of estimating the inverse dynamics of a seven degrees-of-freedom robot arm. The task, as described in [25], is to map from a 21-dimensional input space (7 positions, 7 velocities, 7 accelerations) to the corresponding 7 joint torques. Following [25] we present results below on just one of the mappings, from the 21 input variables to the first of the seven torques. The dataset consists of 48,933 input-output pairs, of which $44,484$ were used as a training set and the remaining 4,449 were used as a test set. The inputs were linearly rescaled to have mean zero and unit variance on the training set. The outputs were centred to have mean zero on the training set.

We consider a hierarchical model based on 21-dimensional Gaussian process (GP) regression. Denote by $Y_i \in \mathbb{R}$ a measured response variable at state $\boldsymbol{z}_i \in \mathbb{R}^{21}$, assumed to satisfy $Y_i = g(\boldsymbol{z}_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ are independent for $i = 1, \ldots, n$ and $\sigma > 0$ will be assumed known. In order to use training data $(y_i, \boldsymbol{z}_i)_{i=1}^n$ to make predictions regarding an unseen test point $\boldsymbol{z}_*$, we place a GP prior $g \sim \mathcal{GP}(0, c(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\theta}))$ where $c(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\theta}) = \theta_1 \exp(-\frac{1}{2}\theta_2^{-2}\|\boldsymbol{z} - \boldsymbol{z}'\|_2^2)$. Here $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are hyper-parameters that control how training samples are used to predict the response at a new test point. A fully-Bayesian treatment aims to marginalise over these hyper-parameters and we assign independent priors $\theta_1 \sim \Gamma(\alpha, \beta)$, $\theta_2 \sim \Gamma(\gamma, \delta)$ in the shape/scale

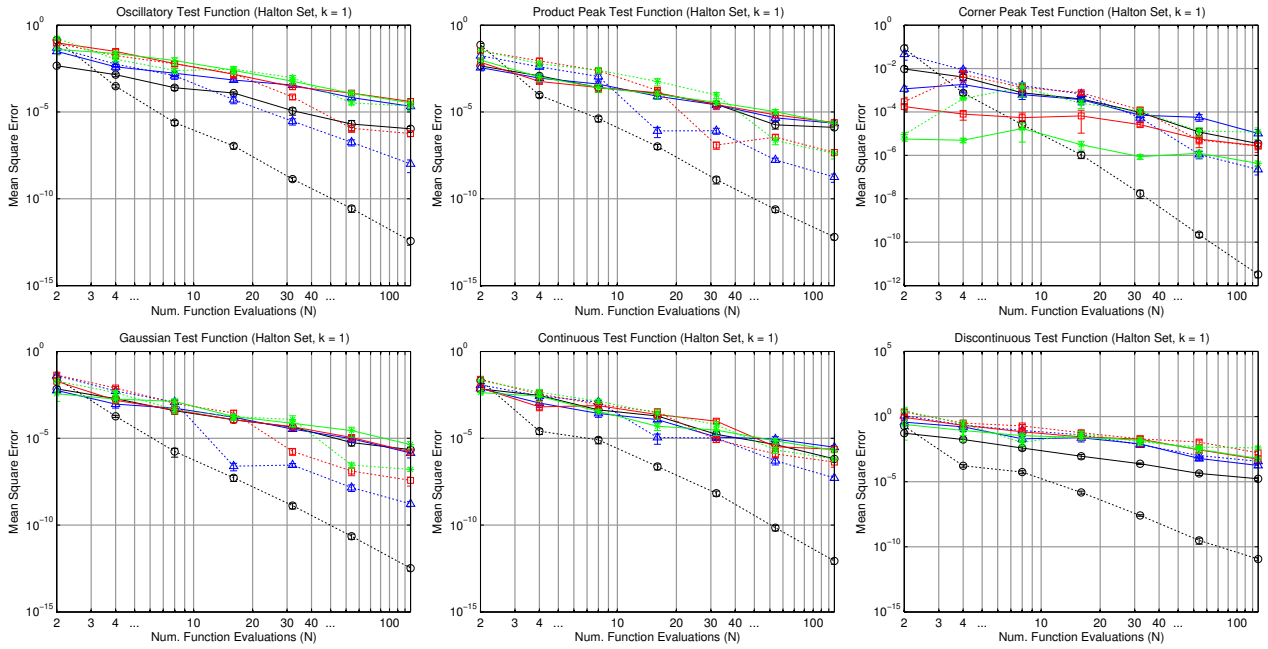

Figure 1: Simulation study (Genz functions): Each panel represents one test function. Solid lines correspond to standard QMC, dashed lines correspond to QMC+CF. ○ represents dimension $d = 1$, △ represents $d = 2$, □ represents $d = 3$ and ∗ represents $d = 4$. Experiments were replicated with 10 random seeds and error bars denote standard error of the replicate mean. QMC points were generated from a shifted and scrambled Halton sequence. A Wendland regression kernel was used with $k = 1$.

parametrisation, which we write jointly as $\pi(\boldsymbol{\theta})$. Here $\sigma = 0.1$, $\alpha = \beta = \gamma = \delta = 2$.

To predict the value of the response Y_* corresponding to an unseen state vector $\boldsymbol{z}_*$, our estimator will be the Bayesian posterior mean

$$\hat{Y}_* := \mathbb{E}[Y_*|\boldsymbol{y}] = \int \mathbb{E}[Y_*|\boldsymbol{y}, \boldsymbol{\theta}]\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \qquad (11)$$

where we implicitly condition on the covariates $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n, \boldsymbol{z}_*$. Phrasing in terms of our earlier notation, the test function is

$$f(\boldsymbol{x}) = \mathbb{E}[Y_*|\boldsymbol{y}, \Pi^{-1}(\boldsymbol{x})] = \boldsymbol{C}_{*,n}(\boldsymbol{C}_n + \sigma^2 \boldsymbol{I}_{n\times n})^{-1}\boldsymbol{y}$$

where Π is the c.d.f for π, $(\boldsymbol{C}_n)_{i,j} = c(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\theta})$ and $(\boldsymbol{C}_{*,n})_{1,j} = c(\boldsymbol{z}_*, \boldsymbol{z}_j; \boldsymbol{\theta})$. Each evaluation of the integrand $f(\boldsymbol{x})$ requires $O(n^3)$ operations due to the matrix inversion and this entails a prohibitive level of computation. A partial solution is provided by a 'subset of regressors' approximation

$$f(\boldsymbol{x}) \approx \boldsymbol{C}_{*,n'}(\boldsymbol{C}_{n',n}\boldsymbol{C}_{n,n'} + \sigma^2 \boldsymbol{C}_{n'})^{-1}\boldsymbol{C}_{n',n}\boldsymbol{y} \qquad (12)$$

where $n' < n$ denotes a subset of the full data; see Sec. 8.3.1 of [25] for full details. However even Eqn. 12 still represents a substantial computational burden in general. To facilitate the illustration below, which investigates the sampling distribution of estimators, we took a random subset of $n = 1,000$ training points and a subset of regressors approximation with $n' = 100$. The total computational time needed to obtain these results was 268 core-hours.

For each test point $\boldsymbol{z}_*$ the sampling standard deviation of $\hat{Y}_*$ was estimated from 10 independent realisations of the QMC procedures. For CF we used a randomly-shifted, scrambled Halton sequence ($\alpha_L = 1$) and Wendland kernels with $k = 1$ ($\alpha_U = 2$), so that theory predicts an acceleration factor of $O(N^{-1/2})$. The estimator standard deviations were estimated for all 4,449 test points (with $N = 2^8$) and the full results are shown in Fig. 2. Note that each test point $\boldsymbol{z}_*$ corresponds to a different test function f and thus these results are quite objective, encompassing thousands of different integration problems. For the vast majority of integration problems, CF accelerated the standard QMC estimator. Here the computational time to construct a functional approximation (inverting a 16×16 matrix) was negligible (3%) in comparison to the cost of evaluating the function f once. The total additional computational time associated with the QMC+CF methodology was 2% greater than for QMC, which is easily justified by the substantial variance reductions ($\sim 10^3$%) that are realised in this application. Supplementary results (Fig. S9) compare QMC+CF to MC+CF (standard MC sampling).

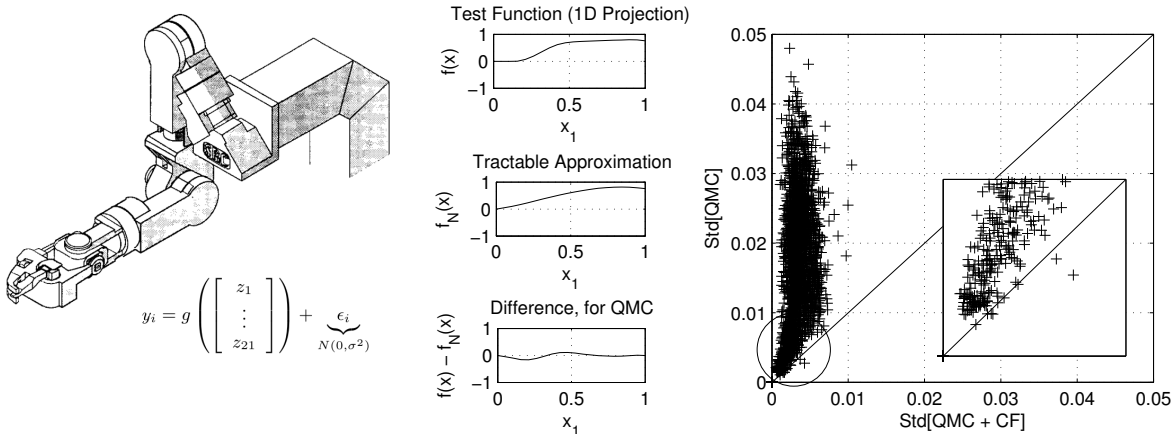

Figure 2: Application to robot arm data. *Left*: Posterior predictive means were computed for the mechanical torque experienced by one of the seven joints of the arm, for each of 4,449 joint configurations. Schematic reproduced from [32]. *Centre*: Model hyper-parameters were integrated out; for this task we compared standard QMC with the proposed QMC+CF approach (both implementations provided unbiased estimators). *Right*: Examining the estimator sampling standard deviations, we see that, for all but a handful of the configurations, QMC+CF was more accurate than QMC.

5 Discussion

QMC methods are becoming increasingly relevant in modern statistics applications [12, 35] and it is surely a priority to target the rate constants governing the practical performance of these algorithms. CFs provide one route to achieve this goal, providing substantial variance reductions in many of the examples we considered. Indeed, CFs allow us to use a sub-optimal QMC rule (e.g. as built into existing software packages) and yet, with minimal additional coding, obtain a QMC+CF algorithm that attains optimal convergence rates. The focus on unknown smoothness α distinguishes our work from previous literature on the connection between integration and functional approximation, e.g. [3, 13].

Functional approximation, and hence our QMC+CF methodology, has a computational cost associated with solution of a linear system. Whilst negligible in our experiments, this cost could be reduced if necessary using standard approximations and/or compactly supported kernels. On the other hand, we note that QMC is often used when f is expensive to evaluate and in such situations it is likely that evaluation of the integrand, rather than solution of a linear system, will be the main computational bottleneck.

Our focus was on Sobolev spaces, but it is known that a faster rate $O(N^{-\alpha+\epsilon})$ is possible in the subspace $SH^{\alpha}([0,1]^d)$, for any $\epsilon > 0$, and explicit point sets are available (for integer α) [6]. An immediate extension is to establish optimal rates for CFs in this class of functions. In a related direction, one can in principle obtain *dimension-independent* rates by imposing a (strong) assumption of polynomial tractability on the RKHS. This is achieved by generalising to weighted Sobolev spaces, such that the integrand f 'depends only weakly on most of the components of $\boldsymbol{x}$'. Further details are provided in [7, 20] and form part of our ongoing research.

The methods that we describe are immediately applicable in a range of applications including marginalisation of hyper-parameters in classification [10], probabilistic inference for differential equations [28, 5], computation of model evidence [21] and approximation of the partition function in social network models [26]. Finally we note that CFs generalise to other integration methods including Bayesian Quadrature [23, 4] and related kernel-based quadrature rules [2], in which the worst case error is also controlled by an RKHS norm $\|f\|_H$; this will be the focus of our ongoing research.

Acknowledgments

The authors are grateful to Dan Simpson, Mathieu Gerber and Ben Collyer for helpful discussions. CJO was supported by EPSRC [EP/D002060/1] and the ARC Centre of Excellence for Mathematics and Statistical Frontiers. MG was supported by EPSRC [EP/J016934/1, EP/K034154/1], an EPSRC Established Career Fellowship, the EU grant [EU/259348] and a Royal Society Wolfson Research Merit Award.

References

[1] C. Aistleitner and J. Dick. Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. *Acta Arithmetica*, 167(2):143–171, 2015.

[2] F. Bach. On the equivalence between quadrature rules and random features. *arXiv:1502.06800*, 2015.

[3] N.S. Bakhvalov. On the approximate calculation of multiple integrals. *Journal of Complexity*, 31(4):502-516, 2015.

[4] F.X. Briol, C.J. Oates, M. Girolami, M. Osborne and D. Sejdinovic. Probabilistic Integration: A Role for Statisticians in Numerical Analysis? *arXiv:1512.00933*, 2016.

[5] J. Cockayne, C.J. Oates, T. Sullivan and M. Girolami. Probabilistic Meshless Methods for Bayesian Inverse Problems. In preparation.

[6] J. Dick. Higher order scrambled digital nets achieve the optimal rate of the root mean square error for smooth integrands. *The Annals of Statistics*, 39(3):1372–1398, 2011.

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

[8] J. Dick and F. Pillichshammer. *Discrepancy theory and quasi-Monte Carlo integration*. Springer, Berlin, 2010.

[9] G.F. Fasshauer. *Meshfree approximation methods with MATLAB*. World Scientific Publishing Co., Inc., 2007.

[10] M. Filippone and M. Girolami. Pseudo-Marginal Bayesian Inference for Gaussian Processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11):2214–2226, 2014.

[11] A. Genz. Testing multidimensional integration routines. In *International Conference on Tools, Methods and Languages for Scientific and Engineering Computation*, pages 81–94. Elsevier North-Holland, Inc., 1984.

[12] M. Gerber and N. Chopin. Sequential Quasi-Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579, 2015.

[13] S. Heinrich. Random approximation in numerical analysis. *Lecture Notes in Pure and Applied Mathematics*, 150:123-171, 1994

[14] F.J. Hickernell, C. Lemieux, and A.B. Owen. Control variates for quasi-Monte Carlo. *Statistical Science*, 20(1):1–31, 2005.

[15] F. Huszár and D. Duvenaud. Optimally-Weighted Herding is Bayesian Quadrature. In *Uncertainty in Artificial Intelligence*, pages 377–385, 2012.

[16] L. Kocis and W.J. Whiten. Computational investigations of low-discrepancy sequences. *ACM Transactions on Mathematical Software (TOMS)*, 23(2):266–294, 1997.

[17] S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential Kernel Herding: Frank-Wolfe Optimization for Particle Filtering. In *International Conference on Artificial Intelligence and Statistics*, volume 18, 2015.

[18] J. Matoušek. On the L2-discrepancy for Anchored Boxes. *J. Complex.*, 14(4):527–556, 1998.

[19] E. Novak. *Deterministic and stochastic error bounds in numerical analysis*. Springer-Verlag Berlin, 1988.

[20] E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems: Standard information for functionals*, volume 2. European Mathematical Society, 2010.

[21] C.J. Oates, T. Papamarkou and M. Girolami. The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation. *J. Am. Stat. Assoc.*, 2016. To appear.

[22] C.J. Oates, M. Girolami, and N. Chopin. Control Functionals for Monte Carlo Integration. *J. R. Statist. Soc. B*, 2016. To appear.

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

[24] A.B. Owen. A constraint on extensible quadrature rules. *Numerische Mathematik*, 2015. To appear.

[25] C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. *MIT Press*, 2(3):4, 2006.

[26] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, 2007.

[27] R. Schaback and H. Wendland. Kernel techniques: from machine learning to meshless methods. *Acta Numerica*, 15:543–639, 2006.

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

[29] W. Sickel and T. Ullrich. Tensor products of Sobolev-Besov spaces and applications to approximation from the hyperbolic cross. *Journal of Approximation Theory*, 161(2):748–786, 2009.

[30] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[31] B. Tracey, D. Wolpert, and J.J. Alonso. Using Supervised Learning to Improve Monte Carlo Integral Estimation. *AIAA Journal*, 51(8):2015–2023, 2013.

[32] S. Vijayakumar and S. Schaal. Locally weighted projection regression: An O(n) algorithm for incremental real time learning in high dimensional space. In *International Conference on Machine Learning*, volume 16, pages 1079–1086, 2000.

[33] X. Wang. Enhancing Quasi-Monte Carlo Methods by Exploiting Additive Approximation for Problems in Finance. *SIAM Journal of Scientific Computing*, 34(1):A283–A308, 2012.

[34] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4(1):389–396, 1995.

[35] J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, volume 31, pages 485–493, 2014.