Bernoulli **25**(2), 2019, 1141–1159
https://doi.org/10.3150/17-BEJ1016

Convergence rates for a class of estimators based on Stein's method

CHRIS J. OATES[1,4], JON COCKAYNE[2], FRANÇOIS-XAVIER BRIOL[2,3] and MARK GIROLAMI[3,4]

[1]*School of Mathematics and Statistics, Newcastle University, UK. E-mail: chris.oates@ncl.ac.uk*
[2]*Department of Statistics, University of Warwick, UK*
[3]*Department of Mathematics, Imperial College London, UK*
[4]*Alan Turing Institute, UK*

Gradient information on the sampling distribution can be used to reduce the variance of Monte Carlo estimators via Stein's method. An important application is that of estimating an expectation of a test function along the sample path of a Markov chain, where gradient information enables convergence rate improvement at the cost of a linear system which must be solved. The contribution of this paper is to establish theoretical bounds on convergence rates for a class of estimators based on Stein's method. Our analysis accounts for (i) the degree of smoothness of the sampling distribution and test function, (ii) the dimension of the state space, and (iii) the case of non-independent samples arising from a Markov chain. These results provide insight into the rapid convergence of gradient-based estimators observed for low-dimensional problems, as well as clarifying a curse-of-dimension that appears inherent to such methods.

Keywords: asymptotics; control functionals; reproducing kernel; scattered data; variance reduction

1. Introduction

This paper considers methods to estimate the integral

$$\int f \, \mathrm{d}\Pi$$

of a test function f against a distribution Π based on evaluation of f at a finite number n of inputs. Our work is motivated by challenging settings in which (i) the variance $\sigma^2(f) = \int (f - \int f \, \mathrm{d}\Pi)^2 \, \mathrm{d}\Pi$ is large relative to n, and (ii) the distribution Π is only available up to an unknown normalisation constant. Such problems arise in Bayesian statistics when the cost of sampling from the posterior is prohibitive, requiring that posterior expectations are approximated based on a small number n of evaluations of the integrand. Indeed, the intrinsic accuracy of ergodic averages, such as obtained via Markov chain Monte Carlo (MCMC) methods [37], can lead to unacceptably high integration error when n is small. This paper considers a class of estimators inspired by Stein's method [41], based on integration-by-parts in this context:

$$\int f \, \mathrm{d}\Pi = -\int \left(\int f \, \mathrm{d}x \right) \cdot \frac{\mathrm{d}}{\mathrm{d}x} \log \pi \, \mathrm{d}\Pi, \tag{1}$$

subject to boundary conditions, where π is a density for Π. These estimators ensure an integration error $o_{\mathrm{P}}(n^{-\frac{1}{2}})$, provided that gradient information on the sampling distribution can be obtained. This is often the case; indeed, sophisticated software for automatic differentiation of statistical models has been developed (e.g., [9,25]).

Main contribution

The primary contribution of this paper is to establish convergence rates for a class of estimators based on Stein's method. These estimators, first described in [35], require as input both function evaluations $\{f(\boldsymbol{x}_i)\}_{i=1}^n$ and gradient evaluations $\{\nabla \log \pi(\boldsymbol{x}_i)\}_{i=1}^n$, where the states $\{\boldsymbol{x}_i\}_{i=1}^n$ themselves can be either independent or correlated draws from Π. Our central results are asymptotic rates for integration error; these enable us to compare and quantify the improvement in estimator precision relative to standard Monte Carlo methods and in doing so we fill a theoretical void.

The estimators that we consider can be viewed as a control variate (or 'control functional') method, and this concept is discussed next.

Control functionals

The classical control variate method proceeds by seeking a collection of non-trivial statistics $\{\psi_i\}_{i=1}^k$, such that each satisfies $\int \psi_i \,\mathrm{d}\Pi = 0$. Then a surrogate function

$$f' = f - a_1\psi_1 - \cdots - a_k\psi_k$$

is constructed such that automatically $\int f' \,\mathrm{d}\Pi = \int f \,\mathrm{d}\Pi$ and, for suitably chosen $\{a_i\}_{i=1}^k$, a variance reduction $\sigma^2(f') < \sigma^2(f)$ might be obtained; for further details see, for example, [39]. For specific problems it is sometimes possible to identify control variates, for example based on physical considerations (e.g., [3]). For Monte Carlo integration based on Markov chains, it is sometimes possible to construct control variates based on statistics relating to the sample path. In this direction, the problem of constructing control variates for discrete state spaces was essentially solved by [1] and for continuous state spaces, recent contributions include [14,20,22,29,30]. Control variates can alternatively be constructed based on gradient information on the sampling distribution [2,31,35,36].

The estimators considered here stem from a recent development that extends control variates to control *functionals*. This idea is motivated by the observation that the methods listed above are (in effect) solving a misspecified regression problem, since in general f does not belong to the linear span of the statistics $\{\psi_i\}_{i=1}^k$. The recent work by [29,35] alleviates model misspecification by increasing the number k of statistics alongside the number n of samples so that the limiting space spanned by the statistics $\{\psi_i\}_{i=1}^\infty$ is dense in a class of functions that contains the test function f of interest. Both methods provide a non-parametric alternative to classical control variates whose error is $o_{\mathrm{P}}(n^{-\frac{1}{2}})$. Of these two proposed solutions, [29] is not considered here since it is unclear how to proceed when Π is known only up to a normalisation constant. On the other hand, the control functional method of [35] is straight-forward to implement when gradients $\{\nabla \log \pi(\boldsymbol{x}_i)\}_{i=1}^n$ are provided. Understanding the theoretical properties of this method is the focus of the present research.

Technical contribution

This paper establishes that the estimators of [35] incur an integration error $O_{\mathrm{P}}(n^{-\frac{1}{2}-\frac{a\wedge b}{d}+\varepsilon})$, where a is related to the smoothness of the density π, b is related to the smoothness of the test function f, d is the dimension of the domain of integration and $\varepsilon > 0$ can be arbitrarily small (a notational convention used to hide logarithmic factors). This analysis provides important insight into the strong performance that has been observed for these estimators in certain low-dimensional applications [23,35]. Indeed, recall that the (naïve) computational cost associated with these methods, that is, the cost of solving a linear system, is $c = O(n^3)$. This cost can also involve a large constant factor when hyper-parameters are to be jointly estimated. Thus, whilst for standard Monte Carlo methods an estimator error of $O_{\mathrm{P}}(c^{-\frac{1}{2}})$ can be achieved at computational cost c, for gradient-based control functionals

$$\text{error for cost } c = O_{\mathrm{P}}\big((c^{\frac{1}{3}})^{-\frac{1}{2}-\frac{a\wedge b}{d}+\varepsilon}\big) = O_{\mathrm{P}}\big(c^{-\frac{1}{6}+\frac{d-a\wedge b}{3d}+\varepsilon}\big).$$

This demonstrates that gradient-based control functionals have asymptotically lower error for the same fixed computational cost c whenever $a \wedge b > d$, which occurs when both the density π and the test function f are sufficiently smooth. In the situation where the computational bottleneck is evaluation of f, not solution of the linear system, then the computational gain can be even more substantial. At the same time, the critical dependence on d highlights the curse-of-dimension that appears inherent to such methods. Going forward, these results provide a benchmark for future high-dimensional development.

Relation to other acceleration methods

Accelerated rates of convergence can be achieved by other means, including quasi-Monte Carlo (QMC; [33]). Consider the ratio estimator:

$$\int f \,\mathrm{d}\Pi \approx \frac{\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)}{\frac{1}{n}\sum_{i=1}^{n} \pi(\boldsymbol{x}_i)}. \tag{2}$$

For appropriate randomised point sets $\{\boldsymbol{x}_i\}_{i=1}^{n}$, the ratio estimator converges at a rate limited by the least smooth of $f \cdot \pi$ and f, that is, limited by $\frac{a\wedge b}{d}$ (at least, in the absence of additional conditions on the mixed partial derivatives, which we have not assumed).[5] See [16] for a recent study of this approach in the context of Bayesian inference for an unknown parameter in a partial differential equation model.

The method studied herein can be contrasted with QMC methods in at least two respects: (1) The states $\{\boldsymbol{x}_i\}_{i=1}^{n}$ can be independent (or correlated) draws from Π, which avoids the need to specifically construct a point set. This is an important benefit in cases where the domain of

[5] In this section the notation a and b is used as a shorthand for the "smoothness" of, respectively, π and f. The precise mathematical definition of a and b differs between manuscripts and the results discussed here should not be directly compared.

integration is complicated – indeed, our results hold for any domain of integration for which an interior cone condition can be established. (2) The estimator studied herein is unbiased, whereas ratio estimators of the form in Eq. (2) will be biased in general. The unbiased nature of the estimator, in common with standard Monte Carlo methods, facilitates convenient diagnostics to estimate the extent of Monte Carlo error and is therefore useful.

Recent work from [15] and [4] considered estimators of the form

$$\int f \, \mathrm{d}\Lambda \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)}{\hat{\pi}(\boldsymbol{x}_i)}, \tag{3}$$

where $\hat{\pi}$ is a kernel density estimate for $\pi = \mathrm{d}\Pi/\mathrm{d}\Lambda$ based on a collection of (possibly correlated) draws $\{\boldsymbol{x}_i\}_{i=1}^n$ from Π. Again, theoretical results established an error of $o_{\mathrm{P}}(n^{-\frac{1}{2}})$ with an explicit rate gated by a term of the form $\frac{a \wedge b}{d}$. However, this approach applies to integrals with respect to a known, normalised reference measure Λ rather than with respect to Π.

Outline

Below in Section 2 we describe the class of estimators that were considered and present our main theoretical results, including the case of non-independent samples arising from a Markov chain sample path. Our theoretical analysis combines error bounds from the scattered data approximation literature with stability results for Markov chains; proofs are contained in the electronic supplement [34]. Numerical results in Section 3 confirm these error rates are realised. Finally, the importance of our findings is discussed in Section 4.

2. Methods

First, we fix notation before describing the estimation method.

2.1. Set-up and notation

Consider an open and bounded set $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, with boundary $\partial\mathcal{X}$. Let $\mathcal{B} = \mathcal{B}(\mathcal{X} \cup \partial\mathcal{X})$ denote the Borel σ-algebra on $\mathcal{X} \cup \partial\mathcal{X}$ and equip $(\mathcal{X} \cup \partial\mathcal{X}, \mathcal{B})$ with the reference measure Λ induced from the restriction of Lebesgue measure on $\mathbb{R}^d$. Further, consider a random variable $\boldsymbol{X}$ on $\mathcal{X} \cup \partial\mathcal{X}$ with distribution Π and suppose Π admits a density $\pi = \mathrm{d}\Pi/\mathrm{d}\Lambda$.

The following notation will be used: $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $a \wedge b := \min(a, b)$, $a_+ := \max(a, 0)$, $\mathbf{1} = [1, \ldots, 1]^\top$, $\|\boldsymbol{x}\|_2^2 := \sum_{i=1}^d x_i^2$, $\nabla_{\boldsymbol{x}} := [\partial/\partial x_1, \ldots, \partial/\partial x_d]^\top$, $1_A(\boldsymbol{x}) = 1$ is the indicator of the event $\boldsymbol{x} \in A$. Write $L^2(\mathcal{X}, \Pi)$ for the vector space of measurable functions $f : \mathcal{X} \to \mathbb{R}$ for which $\sigma^2(f) := \int (f - \int f \, \mathrm{d}\Pi)^2 \, \mathrm{d}\Pi$ exists and is finite. Write $C^k(\mathcal{X})$ for the set of measurable functions for which continuous partial derivatives exist on $\mathcal{X}$ up to order $k \in \mathbb{N}_0$. A function $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be in $C_2^k(\mathcal{X})$ if $\partial^{2k} g / \partial x_{i_1} \cdots \partial x_{i_k} \partial x'_{j_1} \cdots \partial x'_{j_k}$ is $C^0(\mathcal{X} \times \mathcal{X})$ for all $i_1, \ldots, i_k, j_1, \ldots, j_k \in \{1, \ldots, d\}$. The notation $\|f\|_\infty := \sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x})|$ will be used.

2.2. Control functionals

This section introduces the control functional method for integration, a non-parametric extension of the classical control variate method. Recall that the trade-off between random sampling and deterministic approximation in the context of integration is well-understood [5]. Our starting point is, in a similar vein, to establish a trade-off between random sampling and *stochastic* approximation.

We assume throughout that the test function f belongs to $L^2(\mathcal{X}, \Pi)$ and that the boundary $\partial\mathcal{X}$ is piecewise smooth. Consider an independent sample from Π, denoted $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$. This is partitioned into disjoint subsets $\mathcal{D}_0 = \{\boldsymbol{x}_i\}_{i=1}^m$ and $\mathcal{D}_1 = \{\boldsymbol{x}_i\}_{i=m+1}^n$, where $1 \leq m < n$. Although m, n are fixed, we will be interested in the asymptotic regime where $m = O(n^\gamma)$ for some $\gamma \in [0, 1]$. Consider constructing an approximation $f_m \in L^2(\mathcal{X}, \Pi)$ to f, based on $\mathcal{D}_0$. Stochasticity in f_m is induced via the sampling distribution of elements in $\mathcal{D}_0$. The integral $\int f_m \,\mathrm{d}\Pi$ is required to be analytically tractable; we will return to this point.

The estimators that we study take the form

$$I_{m,n} := \frac{1}{n-m} \sum_{i=m+1}^{n} f(\boldsymbol{x}_i) - \left(f_m(\boldsymbol{x}_i) - \int f_m \,\mathrm{d}\Pi \right). \tag{4}$$

Such sample-splitting estimators are unbiased, i.e. $\mathbb{E}_{\mathcal{D}_1}[I_{m,n}] = \int f \,\mathrm{d}\Pi$, where the expectation here is with respect to the sampling distribution Π of the $n-m$ random variables that constitute $\mathcal{D}_1$, and is conditional on fixed $\mathcal{D}_0$. The corresponding estimator variance, again conditional on $\mathcal{D}_0$, is $\mathbb{V}_{\mathcal{D}_1}[I_{m,n}] = (n-m)^{-1}\sigma^2(f - f_m)$. This formulation encompasses control variates as a special case where $f_m = a_1\psi_1 + \cdots + a_k\psi_k$, $k \in \mathbb{N}$, and $\mathcal{D}_0$ are used to select suitable values for the coefficients $\{a_i\}_{i=1}^k$ (see e.g. [39]).

To go beyond control variates and achieve an error of $o_{\mathrm{P}}(n^{-1/2})$, we must construct increasingly accurate approximations f_m to f. Indeed, under the scaling $m = O(n^\gamma)$, if the expected functional approximation error satisfies $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - f_m)] = O(m^{-\delta})$ for some $\delta \geq 0$, then

$$\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}\left[\left(I_{m,n} - \int f \,\mathrm{d}\Pi\right)^2\right] = O\big(n^{-1-\gamma\delta}\big). \tag{5}$$

Here we have written $\mathbb{E}_{\mathcal{D}_0}$ for the expectation with respect to the sampling distribution Π of the m random variables that constitute $\mathcal{D}_0$. The rate above is optimised by taking $\gamma = 1$, so that an optimal sample-split satisfies $m/n \to \rho$ for some $\rho \in (0, 1]$ as $n \to \infty$; this will be assumed in the sequel.

When Π is given via an un-normalised density, this framework can only be exploited if it is possible to construct approximations f_m whose integrals $\int f_m \,\mathrm{d}\Pi$ are available in closed-form. If and when this is possible, the term in parentheses in Eq. (4) is known as a *control functional*. [35] showed how to build a flexible class of control functionals based on Stein's method; the key points are presented next.

2.3. Stein operator

To begin, we make the following assumptions on the density π:

(A1) $\pi \in C^{a+1}(\mathcal{X} \cup \partial\mathcal{X})$ for some $a \in \mathbb{N}_0$.
(A2) $\pi > 0$ in $\mathcal{X}$.

The gradient function $\nabla_{\boldsymbol{x}} \log \pi(\cdot)$ is well-defined and $C^a(\mathcal{X} \cup \partial\mathcal{X})$ by (A1,2). Crucially, gradients can be evaluated even when π is only available un-normalised. Consider the following Stein operator:

$$\begin{aligned} \mathbb{S}_\pi : C^1(\mathcal{X}) \times \cdots \times C^1(\mathcal{X}) &\to C^0(\mathcal{X}) \\ \boldsymbol{\phi}(\cdot) &\mapsto \mathbb{S}_\pi[\boldsymbol{\phi}](\cdot) := \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\phi}(\cdot) + \boldsymbol{\phi}(\cdot) \cdot \nabla_{\boldsymbol{x}} \log \pi(\cdot). \end{aligned} \tag{6}$$

This definition can be motivated in several ways, including via Schrödinger Hamiltonians [2] and via the generator method of Barbour applied to an overdamped Langevin diffusion [18]. The choice of Stein operator is not unique and some alternatives are listed in [17].

For functional approximation we follow [35] and study approximations of the form

$$f_m(\cdot) := \beta + \mathbb{S}_\pi[\boldsymbol{\phi}](\cdot), \tag{7}$$

where $\beta \in \mathbb{R}$ is a constant and $\mathbb{S}_\pi[\boldsymbol{\phi}](\cdot)$ acts as a flexible function, parametrised by the choice of $\boldsymbol{\phi} \in C^1(\mathcal{X}) \times \cdots \times C^1(\mathcal{X})$. Under regularity assumptions introduced below, integration-by-parts (Eq. (1)) can be applied to obtain $\int \mathbb{S}_\pi[\boldsymbol{\phi}] \, \mathrm{d}\Pi = 0$ (Lemma 1). Thus, for this class of functions, $\int f_m \, \mathrm{d}\Pi$ permits a trivial closed-form and $\mathbb{S}_\pi[\boldsymbol{\phi}]$ is a control functional (i.e., integrates to 0).

The choice of β and $\boldsymbol{\phi}$ can be cast as an optimisation problem over a Hilbert space and this will be the focus next.

2.4. Stein operators on Hilbert spaces

This section formulates the construction of f_m as approximation in a Hilbert space $\mathcal{H}_+ \subset L^2(\mathcal{X}, \Pi)$. This construction first appeared in [35] and was subsequently explored in several papers (e.g., [10,19,24]).

First, we restrict each component function $\phi_i : \mathcal{X} \to \mathbb{R}$ to belong to a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Moreover, we insist that $\mathcal{H}$ is a (non-trivial) reproducing kernel Hilbert space (RKHS), that is, there exists a (non-zero) symmetric positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that (i) for all $\boldsymbol{x} \in \mathcal{X}$ we have $k(\cdot, \boldsymbol{x}) \in \mathcal{H}$ and (ii) for all $\boldsymbol{x} \in \mathcal{X}$ and $h \in \mathcal{H}$ we have $h(\boldsymbol{x}) = \langle h, k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}}$ (see [6], for background). The vector-valued function $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$ is defined in the Cartesian product space $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$, itself a Hilbert space with the inner product $\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle \phi_i, \phi_i' \rangle_{\mathcal{H}}$.

To ensure $\mathcal{H} \subseteq C^1(\mathcal{X})$, we make an assumption on k that will be enforced by construction through selection of the kernel:

(A3) $k \in C_2^{b+1}(\mathcal{X} \cup \partial\mathcal{X})$ for some $b \in \mathbb{N}_0$.

2.4.1. *Boundary conditions*

Two further assumptions are made on π. To this end, denote by $\mathcal{Q}(k)$ the set of densities $q = \mathrm{d}Q/\mathrm{d}\Lambda$ on $(\mathcal{X} \cup \partial\mathcal{X}, \mathcal{B})$ such that (a) $q \in C^1(\mathcal{X} \cup \partial\mathcal{X})$, (b) $q > 0$ in $\mathcal{X}$, and (c) for all $i = 1, \dots, d$ we have $\nabla_{x_i} \log q \in L^2(\mathcal{X} \cup \partial\mathcal{X}, Q')$ for all distributions Q' on $(\mathcal{X} \cup \partial\mathcal{X}, \mathcal{B})$. Let $\mathcal{R}(k)$ denote the set of densities q for which $q(\boldsymbol{x})k(\boldsymbol{x}, \cdot) = 0$ for all $\boldsymbol{x} \in \partial\mathcal{X}$.

(A$\bar{2}$) $\pi \in \mathcal{Q}(k)$.
(A4) $\pi \in \mathcal{R}(k)$.

The assumption (A$\bar{2}$) was first discussed in [10]; note in particular that (A$\bar{2}$) implies (A2). A constructive approach to ensure (A4) holds is to start with an arbitrary RKHS $\tilde{\mathcal{H}}$ with reproducing kernel $\tilde{k}$ and let $B : \tilde{\mathcal{H}} \to \operatorname{im}(B)$ be a linear operator such that $B\varphi(\boldsymbol{x}) := \delta(\boldsymbol{x})\varphi(\boldsymbol{x})$, where $\delta(\cdot)$ is a smooth function such that $\pi(\cdot)\delta(\cdot)$ vanishes on $\partial\mathcal{X}$. Then $\mathcal{H} = \operatorname{im}(B)$ is a RKHS whose kernel k is defined by $k(\boldsymbol{x}, \boldsymbol{x}') = \delta(\boldsymbol{x})\delta(\boldsymbol{x}')\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')$. This construction will be used in Section 3. The following lemma shows that $\mathbb{S}_\pi[\boldsymbol{\phi}]$ is a control functional.

Lemma 1. *Under* (A1–4), *if* $\boldsymbol{\phi} \in \mathcal{H}^d$ *then* $\int \mathbb{S}_\pi[\boldsymbol{\phi}]\,\mathrm{d}\Pi = 0$.

Now, consider the set $\mathcal{H}_0 := \mathbb{S}_\pi[\mathcal{H}^d]$, whose elements $\mathbb{S}_\pi[\boldsymbol{\phi}]$ result from application of the Stein operator $\mathbb{S}_\pi$ to elements $\boldsymbol{\phi}$ of the Hilbert space $\mathcal{H}^d$. Oates *et al.* [35], Theorem 1, showed that $\mathcal{H}_0$ can be endowed with the gradient-based reproducing kernel

$$\begin{aligned} k_0(\boldsymbol{x}, \boldsymbol{x}') := {}& (\nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'})k(\boldsymbol{x}, \boldsymbol{x}') + (\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x})) \cdot (\nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}')) \\ & + (\nabla_{\boldsymbol{x}'} \log \pi(\boldsymbol{x}')) \cdot (\nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}')) + (\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x})) \cdot (\nabla_{\boldsymbol{x}'} \log \pi(\boldsymbol{x}'))k(\boldsymbol{x}, \boldsymbol{x}'). \end{aligned} \tag{8}$$

From (A1, $\bar{2}$, 3), it follows that $\mathcal{H}_0 \subseteq C^{a \wedge b}(\mathcal{X} \cup \partial\mathcal{X})$. Moreover, under (A1, $\bar{2}$, 3, 4), the kernel k_0 satisfies $\int k_0(\boldsymbol{x}, \boldsymbol{x}')\Pi(\mathrm{d}\boldsymbol{x}) = 0$ for all $\boldsymbol{x}' \in \mathcal{X}$. Indeed, the function $k_0(\cdot, \boldsymbol{x}')$ belongs to $\mathcal{H}_0$ by definition and Lemma 1 shows that all elements of $\mathcal{H}_0$ have zero integral.

2.4.2. *Approximation in $\mathcal{H}_+$*

Now we can be specific about how β and $\boldsymbol{\phi}$ are selected. Write $\mathcal{H}_\mathbb{R}$ for the RKHS of constant functions, characterised by the kernel $k_\mathbb{R}(\boldsymbol{x}, \boldsymbol{x}') = c$, $c > 0$, for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. Denote the norms associated to $\mathcal{H}_\mathbb{R}$ and $\mathcal{H}_0$ respectively by $\|\cdot\|_{\mathcal{H}_\mathbb{R}}$ and $\|\cdot\|_{\mathcal{H}_0}$. Write

$$\mathcal{H}_+ := \mathcal{H}_\mathbb{R} + \mathcal{H}_0 = \{\beta + \psi : \beta \in \mathcal{H}_\mathbb{R}, \psi \in \mathcal{H}_0\}.$$

Equip $\mathcal{H}_+$ with the norm $\|f\|^2_{\mathcal{H}_+} := \|\beta\|^2_{\mathcal{H}_\mathbb{R}} + \|\psi\|^2_{\mathcal{H}_0}$. It can be shown that $\mathcal{H}_+$ is a RKHS with kernel $k_+(\boldsymbol{x}, \boldsymbol{x}') := k_\mathbb{R}(\boldsymbol{x}, \boldsymbol{x}') + k_0(\boldsymbol{x}, \boldsymbol{x}')$ ([6], Theorem 5, page 24). From (A1–3), it follows that $\mathcal{H}_+ \subseteq C^{a \wedge b}(\mathcal{X})$.

The choice of β and $\boldsymbol{\phi}$ is cast as a least-squares optimisation problem:

$$f_m := \arg\min \|h\|^2_{\mathcal{H}_+} \quad \text{s.t.} \quad \forall i = 1, \dots, m, \qquad h(\boldsymbol{x}_i) = f(\boldsymbol{x}_i), \qquad h \in \mathcal{H}_+.$$

By the representer theorem [40], we have $f_m(\boldsymbol{x}) = \sum_{i=1}^m a_i k_+(\boldsymbol{x}, \boldsymbol{x}_i)$ where the coefficients $\mathbf{a} = [a_1, \dots, a_m]^\top$ are the solution of the linear system $\mathbf{K}_+\mathbf{a} = \mathbf{f}_0$ where $\mathbf{K}_+ \in \mathbb{R}^{m \times m}$, $[\mathbf{K}_+]_{i,j} =$

$k_+(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\mathbf{f}_0 \in \mathbb{R}^{m\times 1}$, $[\mathbf{f}_+]_i = f(\boldsymbol{x}_i)$. In situations where $\mathbf{K}_+$ is not full-rank, we define $f_m \equiv 0$. Numerical inversion of this system is associated with a $O(m^3)$ cost and may in practice require additional numerical regularisation; this is relatively standard.

2.5. Theoretical results

Our novel analysis, next, builds on results from the scattered data approximation literature [45] and the study of the stability properties of Markov chains [26].

2.5.1. *The case of independent samples*

First, we focus on scattered data approximation and state two assumptions that are central to our analysis:

(A5) $\pi > 0$ on $\mathcal{X} \cup \partial\mathcal{X}$.
(A6) $f \in \mathcal{H}_+$.

Here (A5) extends (A2) in requiring also that $\pi > 0$ on $\partial\mathcal{X}$. (A6) ensures that the problem is well-posed. Define the fill distance

$$h_{\mathcal{D}_0} := \sup_{\boldsymbol{x}\in\mathcal{X}} \min_{i=1,\dots,m} \|\boldsymbol{x} - \boldsymbol{x}_i\|_2.$$

The proof strategy that we present here decomposes into two parts; (i) first, error bounds are obtained on the functional approximation error $\sigma^2(f - f_m)$ in terms of the fill distance $h_{\mathcal{D}_0}$, (ii) second, the fill distance $h_{\mathcal{D}_0}$ is shown to vanish under sampling (with high probability). For (ii) to occur, we require an additional constraint on the geometry of $\mathcal{X}$:

(A7) The domain $\mathcal{X} \cup \partial\mathcal{X}$ satisfies an *interior cone condition*, that is, there exists an angle $\theta \in (0, \pi/2)$ and a radius $r > 0$ such that for every $\boldsymbol{x} \in \mathcal{X} \cup \partial\mathcal{X}$ there exists a unit vector $\boldsymbol{\xi}$ such that the cone

$$\mathcal{C}(\boldsymbol{x}, \boldsymbol{\xi}, \theta, r) := \{\boldsymbol{x} + \lambda\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^d, \|\boldsymbol{y}\|_2 = 1, \boldsymbol{y}^\top\boldsymbol{\xi} \geq \cos\theta, \lambda \in [0, r]\}$$

is contained in $\mathcal{X} \cup \partial\mathcal{X}$.

The purpose of (A7) is to rule out the possibility of ‘pinch points’ on $\partial\mathcal{X}$ (i.e., $\prec$-shaped regions), since intuitively sampling-based approaches can fail to ‘get into the corners’ of the domain. The limiting behaviour of the fill distance under sampling enters through the following technical result:

Lemma 2. *Let* $g : [0, \infty) \to [0, \infty)$ *be continuous, monotone increasing, and satisfy* $g(0) = 0$ *and* $\lim_{x\downarrow 0} g(x)\exp(x^{-3d}) = \infty$. *Then under* (A5,7) *we have*

$$\mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})] = O(g(m^{-\frac{1}{d}+\varepsilon})),$$

where $\varepsilon > 0$ *can be arbitrarily small.*

Our first main result can now be stated as the following.

Theorem 1. *Assume* (A1, $\bar{2}$, 3–7). *Recall that we partition the set $\mathcal{D}$ as $\mathcal{D}_0 \cup \mathcal{D}_1$ where $|\mathcal{D}_0| = m$ and $|\mathcal{D}_1| = n - m$. There exists $h > 0$, independent of m, n, such that the estimator $I_{m,n}$ is an unbiased estimator of $\int f \,\mathrm{d}\Pi$ with*

$$\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}\left[1_{h_{\mathcal{D}_0}<h}\left(I_{m,n} - \int f \,\mathrm{d}\Pi\right)^2\right] = O\big((n-m)^{-1}m^{-2\frac{a\wedge b}{d}+\varepsilon}\big),$$

where $\varepsilon > 0$ can be arbitrarily small.

Thus for $m = O(n)$, this result establishes an overall error of $O(n^{-1-2\frac{a\wedge b}{d}+\varepsilon})$, as claimed. This establishes that these estimates are more efficient than standard Monte Carlo estimators when $a \wedge b > 0$. Or, when the cost of solving a linear system is taken into account, the method is more efficient on a per-cost basis when $a \wedge b > d$. This provides new insight into the first set of empirical results reported in [35] where, for assessment purposes, samples were generated independently from known, smooth densities. There, control functionals were constructed based on smooth kernels and integration errors were shown to be substantially reduced.

On the negative side, this result illustrates a curse of dimension that appears to be intrinsic to the method. We return to this point in Section 4.

The results above hold for independent samples, yet the main area of application for control functionals is estimation based on the MCMC output. In the next section, we prove that the assumption of independence can be relaxed.

2.5.2. *The case of non-independent samples*

In practice, samples from posterior distributions are often obtained via MCMC methods. Our analysis must therefore be extended to the non-independent setting: Consider the case where $\{\boldsymbol{x}_i\}_{i=1}^n$ are generated by a reversible Markov chain targeting Π. We make the following stochastic stability assumption:

(A8) The Markov chain is uniformly ergodic.

Then our first step is to extend Lemma 2 to the non-independent setting.

Lemma 3. *The conclusion of Lemma 2 holds when $\{\boldsymbol{x}_i\}_{i=1}^n$ are generated via MCMC, subject to* (A8).

Non-independence presents us with the possibility that two of the states $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}_0$ are identical (for instance, when a Metropolis–Hastings sample is used and a rejection occurs). Under our current definition, such an event would cause the kernel matrix $\mathbf{K}_+$ to become singular and the control functional to become trivial $f_m = 0$. It is thus necessary to modify the construction. Specifically, we assume that $\mathcal{D}_0$ has been pre-filtered such that any repeated states have been removed. Note that this does not 'introduce bias', since we are only pre-filtering $\mathcal{D}_0$, not $\mathcal{D}_1$. This reduces the effective number m of points in $\mathcal{D}_0$ by at most a constant factor and has no impact on the asymptotics.

With this technical point safely surmounted, we present our second main result.

Theorem 2. *The conclusion of Theorem 1 holds when $\{\boldsymbol{x}_i\}_{i=1}^n$ are generated via MCMC, subject to* (A8).

This result again demonstrates that control functionals are more cost-efficient than standard Monte Carlo when $a \wedge b > d$ and that efficiency is limited by the rougher of the density π and the test function f. This helps to explain the second set of empirical results obtained in [35], where excellent performance was reported on problems that involved smooth densities, smooth kernels and MCMC sampling methods. On the other hand, we again observe a curse of dimension that is inherent to control functionals and, indeed, control variates in general.

2.6. Commentary

Several points of discussion are covered below, on the appropriateness of the assumptions, the strength of the results and aspects of implementation.

On the assumptions

Assumptions (A1, $\bar{2}$, 3, 7) are not unduly restrictive. The boundary condition (A4) has previously been discussed in [35]. Below we discuss the remaining assumptions, (A5, 6, 8).

Our entire analysis was predicated on (A5), the assumption that π is bounded away from 0 on the compact set $\mathcal{X} \cup \partial\mathcal{X}$. This ensured that π was equivalent to Lebesgue measure on $\mathcal{X} \cup \partial\mathcal{X}$ and enabled this change of measure in the proofs. This is clearly a restrictive set-up as certain distributions of interest do vanish, however the assumption was intrinsic to our theoretical approach.

Our analysis also relied on (A6), that is, that f belongs to the function space $\mathcal{H}_+$. It it is thus natural to examine this assumption in more detail. To this end, we provide the following lemma. Recall that a RKHS $\mathcal{H}$ is *c-universal* if it is dense as a set in $(C^0(\mathcal{X} \cup \partial\mathcal{X}), \|\cdot\|_\infty)$.

Lemma 4. *Assume* (A$\bar{2}$, 3, 4). *If $\mathcal{H}$ is c-universal, then $\mathcal{H}_+$ is dense as a set in* $(L^2(\mathcal{X} \cup \partial\mathcal{X}, \Pi), \|\cdot\|_2)$.

The notion of c-universality was introduced by [42], who showed that many widely-used kernels are c-universal on compact sets. Indeed, Proposition 1 of [27] proves that a RKHS with kernel k is c-universal if and only if the map $\Pi' \mapsto \Pi'[k(\cdot,\cdot)]$, from the space of finite signed Borel measures Π' to the RKHS $\mathcal{H}$, is injective, which is a weak requirement. It is *not*, however, clear whether (A4), (A5) can both hold when k is also c-universal. Further work will therefore be required to better assess the consequences of $f \notin \mathcal{H}_+$. This might proceed in a similar vein to the related work of [21,32].

The last assumption to discuss is (A8); uniform ergodicity of the Markov chain. Since π is absolutely continuous with respect to Lebesgue measure on $\mathcal{X} \cup \partial\mathcal{X}$, in practice any Markov chain that targets Π will typically be uniformly ergodic. Indeed, [38] constructed an example where a pinch point in the domain caused a Gibbs sampler targeting a uniform distribution to fail to be geometrically ergodic; their construction violates our (A7).

On the results

The intuition for the results in Theorems 1 and 2 can be described as 'accurate estimation with high probability', since the condition $h_{\mathcal{D}_0} < h$ is satisfied when the samples $\mathcal{D}_0$ cover the state space $\mathcal{X}$, which occurs with unit probability in the $m \to \infty$ limit. There are two equivalent statements that can be made unconditionally on $h_{\mathcal{D}_0} < h$: (i) First, one can simply re-define $f_m = 0$ whenever $h \geq h_0$, that is, when the states $\mathcal{D}_0$ are poorly spaced we revert to the usual Monte Carlo estimator. (ii) Second, one could augment $\mathcal{D}_0$ with additional fixed states, such as a grid, $\{\boldsymbol{g}_i\}_{i=1}^{G}$, to ensure that $h_{\mathcal{D}_0} < h$ is automatically satisfied. However, we find both of these equivalent approaches to be less aesthetically pleasing, since in practice this requires that h be explicitly computed.

The condition $h_{\mathcal{D}_0} < h$ suggests that the asymptotics hold in the same regime where QMC methods could also be successful. However, as explained in Section 1, the method of [35] carries some advantages over the QMC approach that could be important. First, it provides unbiased estimation of $\int f \,\mathrm{d}\Pi$, which enables straight-forward empirical assessment. Second, the fact that it is based on MCMC output renders it more convenient to implement.

On the sharpness of our results, we refer to Section 11.7 of [45] where an overview of the strengths and weaknesses of results in the scattered data approximation literature is provided.

On the data-split

It is required to partition samples into sets $\mathcal{D}_0$ and $\mathcal{D}_1$, whose sizes must be specified. Substituting $\rho = m/n$ into the conclusion of Theorem 1 and minimising this expression over $\rho \in (0, 1]$ leads to an optimal value

$$\rho^* = \frac{\nu}{1+\nu} \qquad \text{where } \nu = 2\frac{a \wedge b}{d}. \tag{9}$$

Thus, when $a \wedge b \gg d$ we have $\rho^* \approx 1$ and the optimal method is essentially a numerical quadrature method (i.e., all samples assigned to $\mathcal{D}_0$). Conversely, when $a \wedge b \ll d$ we have $\rho^* \approx 0$ and the optimal method becomes a Monte Carlo method (i.e., all samples assigned to $\mathcal{D}_1$).

On the bandwidth

For the experiments reported next, we considered radial kernels of the form

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') = \varphi\left(\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{h}\right),$$

where $h > 0$ is a bandwidth parameter and φ is a radial basis function, to be specified. An appropriate value for the bandwidth h must therefore be selected. An important consideration is that if h is selected based on $\mathcal{D}_0$ but not on $\mathcal{D}_1$ then the estimator $I_{m,n}$ remains unbiased. To this end, we propose to select h via maximisation of the log-marginal likelihood

$$\log p(\mathbf{f}_0|\mathcal{D}_0, h) = -\frac{1}{2}\mathbf{f}_0^\top \mathbf{K}_+^{-1}\mathbf{f}_0 - \frac{1}{2}\log|\mathbf{K}_+| - \frac{m}{2}\log 2\pi$$

which arises from the duality with Gaussian processes and approximation in RKHS (see e.g., [6]).

On an extension

An extension of the estimation method was also considered. Namely, for each i one can build an approximation $f^{(-i)} \in \mathcal{H}_+$ to be used as a control functional for $f(\boldsymbol{x}_i)$, based on $\mathcal{D} \setminus \{\boldsymbol{x}_i\}$. This results in a leave-one-out (LOO) estimator

$$I_n := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - \left(f^{(-i)}(\boldsymbol{x}_i) - \int f^{(-i)}\,\mathrm{d}\Pi\right) \tag{10}$$

that again remains unbiased. The performance of I_n can be expected to compare favourably with that of $I_{m,n}$, but the computational cost of I_n is larger at $O(n^4)$.

On computation

It is important to emphasise the ease with which these estimators can be implemented. In the $c \to \infty$ limit, explicit evaluation of Eq. (4) is particularly straight-forward:

$$I_{m,n} = \frac{1}{n-m}\mathbf{1}^\top\left\{\mathbf{f}_1 - \mathbf{K}_{10}\mathbf{K}_0^{-1}\left[\mathbf{f}_0 - \left(\frac{\mathbf{1}^\top\mathbf{K}_0^{-1}\mathbf{f}_0}{\mathbf{1}^\top\mathbf{K}_0^{-1}\mathbf{1}}\right)\mathbf{1}\right]\right\}, \tag{11}$$

where $\mathbf{f}_1 \in \mathbb{R}^{n-m\times 1}$, $[\mathbf{f}_1]_i = f(\boldsymbol{x}_{m+i})$, $\mathbf{K}_0 \in \mathbb{R}^{m\times m}$, $[\mathbf{K}_0]_{i,j} = k_0(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\mathbf{K}_{10} \in \mathbb{R}^{n-m\times m}$ and $[\mathbf{K}_{10}]_{i,j} = k_0(\boldsymbol{x}_{m+i}, \boldsymbol{x}_j)$. In a sense, this expression generalises the usual kernel quadrature estimator to obtain an estimator that is unbiased [7,8]. An implementation called `control_func.m` is available on the Matlab File Exchange to download.

3. Numerical results

First, in Section 3.1, we assessed whether the theoretical results are borne out in simulation experiments. Then, in Section 3.2, we applied the method to a topical parameter estimation problem in uncertainty quantification for a groundwater flow model.

3.1. Simulation

To construct a test-bed for the theoretical results, we considered the simple case where Π is the uniform distribution on $\mathcal{X} = [0, 1]^d$. The test functions that we considered took the form $f(\boldsymbol{x}) = 1 + \sin(2\pi\omega x_1)$ where ω was varied to create a problem that was either ‘easy’ ($\omega = 1$) or ‘hard’ ($\omega = 3$). The importance of the first coordinate x_1 aimed to reflect the ‘low effective dimension’ phenomena that is often encountered. From symmetry of the integrand, the true integral is 1.

For estimation, we took the radial basis function φ to have variable smoothness and compact support, as studied in [44]. Explicit formulae for the φ and their derivatives are contained in the electronic supplement. To enforce (A4), we took $\delta(\boldsymbol{x}) = \prod_{i=1}^{d} x_i(1 - x_i)$ which vanishes on $\partial\mathcal{X}$. The data-split fraction ρ and the bandwidth h were each optimised as described in Section 2.6. Optimisation for h was performed through 10 iterations of the Matlab function `fminbnd` constrained to $h \in [0, 10]$.

Three estimators were considered; the standard Monte Carlo estimator, the control functional (CF) estimator $I_{m,n}$ in Eq. (11) and the LOO estimator I_n. In the case of the LOO estimator, the bandwidth h was re-optimised in building each of the n control functionals $f^{(-i)}$. (A1, $\bar{2}$, 3–5, 7) were satisfied in this experiment. Thus, for $f \in \mathcal{H}_+$, Theorem 1 entails a mean squared integration error for $I_{m,n}$ of $O(n^{-1-2\frac{b}{d}+\varepsilon})$, since $\pi(\boldsymbol{x}) = 1 \in C^{a+1}$ for all $a \in \mathbb{N}_0$. However, the theoretical analysis does not take into account automatic selection of the bandwidth h; this will be assessed through experiment.

Independent samples

To study estimator performance, we repeatedly generated collections of n independent uniform random variables $\{\boldsymbol{x}_i\}_{i=1}^n$ and evaluated all three estimators on this set. The procedure was repeated several times to obtain estimates (along with standard errors) for the average mean square errors (MSE) that were incurred. Results are displayed in Figure 1. In these experiments, the MSE appeared to decrease at least as rapidly as the rates that were predicted. Also, as predicted, the es-

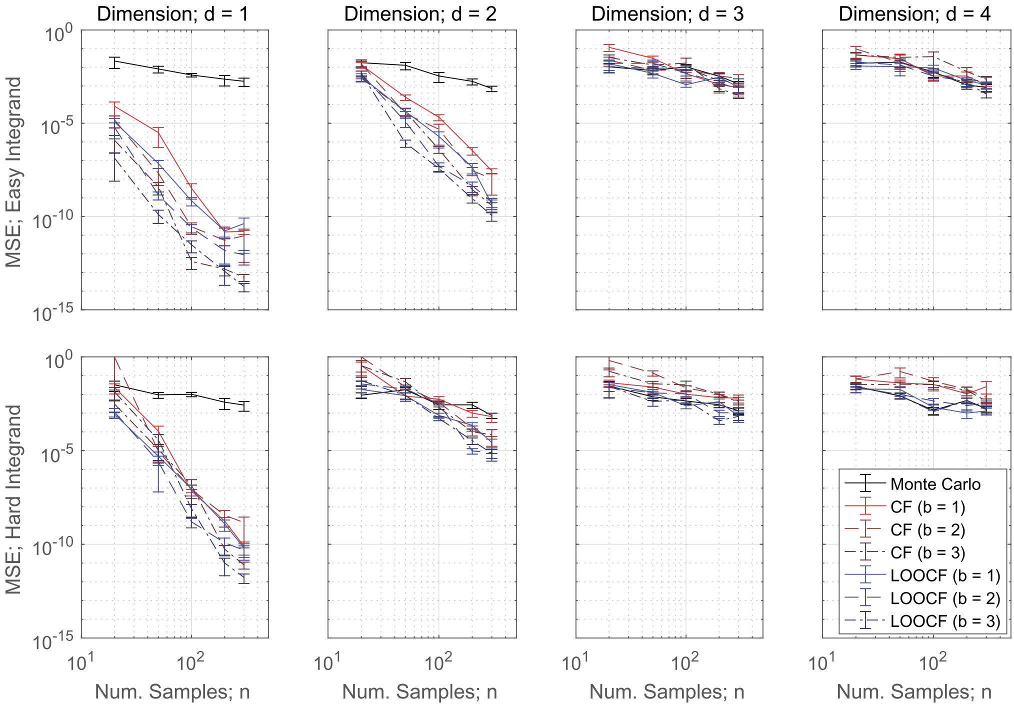

Figure 1. Simulation results; the case of independent samples. An 'easy' and a 'hard' integrand were considered. The mean square error (MSE) was estimated for the standard Monte Carlo estimator, the control functional (CF) estimator $I_{m,n}$ and the leave-one-out (LOOCF) estimator I_n, and plotted against the number n of samples used. The CF and LOOCF estimators were based on kernels of smoothness $b \in \{1, 2, 3\}$. Standard errors are also displayed.

timator performance quickly deteriorated as the dimension d was increased. Indeed, for $d = 3, 4$ an improvement over standard Monte Carlo was no longer observed. The LOO estimator I_n in general out-performed the CF estimator $I_{m,n}$, as expected, but at an increased computational cost. The integration error was in general larger for the hard integrand.

Dependent samples

The effect of correlation among the $\boldsymbol{x}_i$ was also explored. For this, we considered a random walk $\boldsymbol{x}_i = \boldsymbol{x}_{i-1} + \boldsymbol{e}_i$ on the d-torus with $\{\boldsymbol{e}_i\}_{i=1}^n$ drawn uniformly on $[-\varepsilon, \varepsilon]^d$ and $\boldsymbol{x}_0 = \mathbf{0}$. This is a Markov chain with invariant distribution Π. The objective was to assess estimator performance as a function of the step size parameter ε; results for $n = 100$ are shown in Figure 2. Compared to Figure 1, the MSE was larger in general when $\varepsilon < 0.5$. This reflects reduction in effective sample size of the set $\mathcal{D}_0$ used to build the control functional.

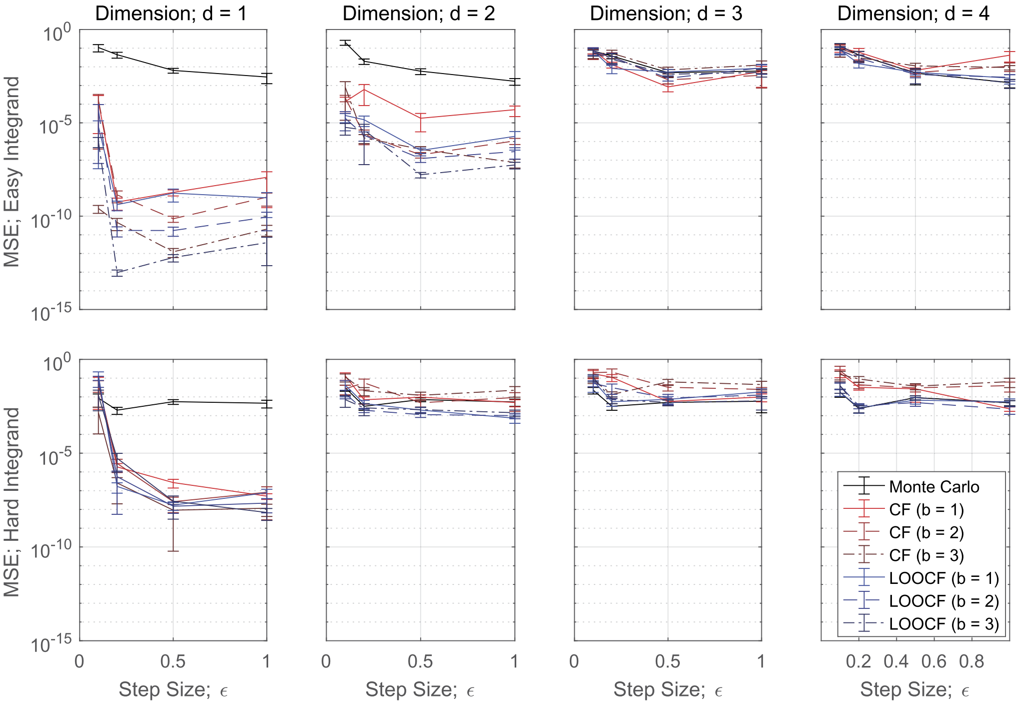

Figure 2. Simulation results; the case of dependent samples. An 'easy' and a 'hard' integrand were considered. The mean square error (MSE) was estimated for the standard Monte Carlo estimator, the control functional (CF) estimator $I_{m,n}$ and the leave-one-out (LOOCF) estimator I_n, where samples from a random walk of length $n = 100$ was used. The MSE was plotted against the step size ε of the random walk. The CF and LOOCF estimators were based on kernels of smoothness $b \in \{1, 2, 3\}$. Standard errors are also displayed.

3.2. Application to partial differential equations

Our theoretical results are illustrated with a novel application to an inverse problem arising in a partial differential equation (PDE) model. Specifically, we considered the following elliptic diffusion problem with mixed Dirichlet and Neumann boundary conditions:

$$\begin{aligned} \nabla_{\boldsymbol{x}} \cdot \big[\kappa(\boldsymbol{x};\boldsymbol{\theta})\nabla_{\boldsymbol{x}} w(\boldsymbol{x})\big] &= 0 \qquad \text{if } x_1, x_2 \in (0,1), \\ w(\boldsymbol{x}) &= \begin{cases} x_1 & \text{if } x_2 = 0, \\ 1 - x_1 & \text{if } x_2 = 1, \end{cases} \\ \nabla_{x_1} w(\boldsymbol{x}) &= 0 \qquad \text{if } x_1 \in \{0,1\}. \end{aligned}$$

This PDE serves as a simple model of steady-state flow in aquifers and other subsurface systems; κ can represent the permeability of a porous medium while w represents the hydraulic head. The aim is to make inferences on the field κ in a setting where the underlying solution w is observed with noise on a regular grid of M^2 points $\boldsymbol{x}_{i,j}$, $i, j = 1, \dots, M$. The observation model $p(\boldsymbol{y}|\boldsymbol{\theta})$ takes the form $\boldsymbol{y} = \{y_{i,j}\}$ where $y_{i,j} = w(\boldsymbol{x}_{i,j}) + \varepsilon_{i,j}$ and $\varepsilon_{i,j}$ are independent normal random variables with standard deviation $\sigma = 0.1$.

Following [43], the field κ was endowed with a prior distribution of the form $\log \kappa(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{d} \theta_i \kappa_i(\boldsymbol{x})$, where the κ_i are Fourier basis functions and θ_i are their associated coefficients. For the inference, we imposed a uniform prior $p(\boldsymbol{\theta}) \propto 1$ over the domain $[-10, 10]^d$. Our aim was to obtain accurate estimates for the posterior mean of the parameter $\boldsymbol{\theta}$. The posterior density $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$ is available up to an unknown normalising constant $p(\boldsymbol{y})$. Each evaluation of the likelihood necessitates the solution of the PDE; control functionals offer the possibility to reduce the number of likelihood evaluations, and hence the computational cost, required to achieve a given estimator precision.

As an aside, we note that the standard approach to inference employs a numerical integrator for the forward-solve, typically based on finite element methods. This would provide us with gradient information on the posterior, but would also introduce some bias due to discretisation error. To ensure that we obtain exact gradient information, we instead exploited a probabilistic meshless method due to [11] as our numerical integrator. See also [12,13]. Automatic differentiation was performed using the `Autograd` package [25].

The key assumptions of our theory were verified. Smoothness of the prior, together with ellipticity, imply (A1) holds for all $a \in \mathbb{N}$. (A$\bar{2}$, 5) hold since the prior and likelihood are well-behaved. (A7) holds since the domain of integration was a hyper-cuboid. Samples from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ were obtained using a Metropolis–adjusted Langevin sampler with fixed proposal covariance; this ensured that (A8) was satisfied. Remaining assumptions were satisfied by construction of the kernel k: Following the approach outlined in Section 2.4, we took $\tilde{k}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to be the standard Matérn kernel of order $\frac{7}{2}$, so that $b = 2$, and then formed $k(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as the product of $\tilde{k}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\delta(\boldsymbol{\theta})\delta(\boldsymbol{\theta}')$, where the boundary function δ satisfies $\delta(\boldsymbol{\theta}) = 1$ on $\boldsymbol{\theta} \in [-9, 9]^d$, $\delta(\boldsymbol{\theta}) = 0$ when $\theta_i \in \{-10, 10\}$ for some i, and δ was infinitely differentiable on $[-10, 10]^d$. With this construction, (A3) holds. (A4) holds since k has a root at $\theta_i \in \{-10, 10\}$ for each i. The constant $c = 1$ was fixed. However the conclusion of Lemma 4 cannot be directly applied here since $\mathcal{H}$ is not c-universal (k vanishes at $\theta_i = \pm 10$).

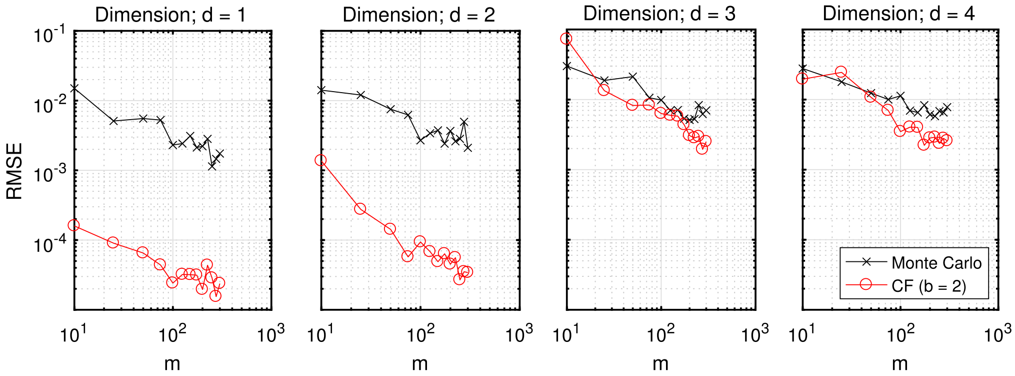

Figure 3. Experimental results; an experiment to approximate the posterior mean of the parameters $\boldsymbol{\theta} \in [-10, 10]^d$ that govern a permeability field. The figure shows root mean square error (RMSE) for (i) the standard Monte Carlo estimator based on $2m$ posterior samples, and (ii) the control functional (CF) estimator, where m samples are used to train the control functional and the remaining m samples are used to estimate the expectation. [Results are shown for the first parameter θ_1; results for other parameters were similar. The Matérn kernel of order $7/2$ was employed; $b = 2$ in our notation.]

Observations were generated from the model with data-generating parameter $\boldsymbol{\theta} = \mathbf{1}$ and collected over a coarse grid of $M^2 = 36$ locations. Samples of size n were obtained from the posterior and divided equally between the training set $\mathcal{D}_0$ and test set $\mathcal{D}_1$. The performance of gradient-based control functionals was benchmarked against that of standard Monte Carlo with all n samples used. We note that, in all experiments, all values of $\boldsymbol{\theta}$ encountered were contained in $[-9, 9]^d$. Thus it does not matter that we did not specify δ explicitly above, emphasising the weakness of assumption (A4) in practical application.

Results are shown in Figure 3. For dimensions $d = 1$ and 2, the estimator that uses control functionals achieved a dramatic reduction in asymptotic variance compared to the Monte Carlo benchmark. On the other hand, for $d = 3, 4$, the curse of dimension is clearly evident for the control functional method.

4. Conclusion

This paper has established novel asymptotic analysis for a class of estimators based on Stein's method. Our analysis makes explicit the contribution of the smoothness a of the distribution Π, the smoothness b of the test function f and the dimension d of the domain of integration. As such, these results provide a rigorous theoretical explanation for the excellent performance in low-dimensions observed in previous work.

Several extensions of this work are suggested: (i) Our results focused on compact domains, since this is the usual setting for results in the scattered data approximation literature. However, the estimation method does not itself require that the domain of integration be compact. Extending this analysis to the unbounded-domain setting appears challenging at present and remains a goal for future research. (ii) Alternative literatures to the scattered data literature could form the basis of an analysis of control functionals, such as e.g. recent work by [28]. These efforts have the advantage of providing L^2 error bounds, rather than L^∞ error bounds and might facilitate the extension to unbounded domains. (iii) Generally, our theoretical results clarify the need to develop estimation strategies that do not suffer from the curse of dimension. While this curse is intrinsic to functional approximation in general, due to the need to explore the state space, the observation that many test functions of interest are of low 'effective dimension' suggests that more regularity on the function space could reasonably be assumed. (iv) Recent work in [23] imposed an additional constraint on the coefficients a_i in Section 2.4.2. It would be interesting to extend our analysis to this context.

Acknowledgements

The authors wish to thank Aretha Teckentrup, Motonobu Kanagawa, Lester Mackey and the anonymous referees for their useful feedback. CJO and MG were supported by the Lloyds Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. CJO was also supported by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). FXB was supported by EPSRC [EP/L016710/1]. MG was supported by the EPSRC grants [EP/K034154/1, EP/R018413/1, EP/P020720/1, EP/L014165/1], and an EPSRC Established Career Fellowship, [EP/J016934/1].

Supplementary Material

Supplement to "Convergence rates for a class of estimators based on Stein's method" (DOI: 10.3150/17-BEJ1016SUPP; .pdf). Proofs of all theoretical results are provided.

References

[1] Andradóttir, S., Heyman, D.P. and Ott, T.J. (1993). Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Trans. Model. Comput. Simul.* **3** 167–189.

[2] Assaraf, R. and Caffarel, M. (1999). Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.* **83** 4682–4685.

[3] Assaraf, R. and Caffarel, M. (2003). Zero-variance zero-bias principle for observables in quantum Monte Carlo: Application to forces. *J. Chem. Phys.* **119** 10536.

[4] Azaïs, R., Delyon, B. and Portier, F. (2016). Integral estimation based on Markovian design. Available at arXiv:1609.01165.

[5] Bahvalov, N.S. (1959). Approximate computation of multiple integrals. *Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. Him.* **1959** 3–18. MR0115275

[6] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston: Kluwer Academic.

[7] Briol, F.-X., Oates, C.J., Cockayne, J., Chen, W.Y. and Girolami, M. (2017). On the sampling problem for kernel quadrature. In *Proceedings of the* 34*th International Conference on Machine Learning*.

[8] Briol, F.-X., Oates, C.J., Girolami, M., Osborne, M.A. and Sejdinovic, D. (2017). Probabilistic integration: A role in statistical computation? Available at arXiv:1512.00933.

[9] Carpenter, B., Hoffman, M.D., Brubaker, M., Lee, D., Li, P. and Betancourt, M. (2015). The Stan math library: Reverse-mode automatic differentiation in C++. Available at arXiv:1509.07164.

[10] Chwialkowski, K., Strathmann, H. and Gretton, A. (2016). A kernel test of goodness of fit. In *Proceedings of the* 33*rd International Conference on Machine Learning*.

[11] Cockayne, J., Oates, C.J., Sullivan, T. and Girolami, M. (2016). Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. In *Proceedings of the* 36*th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*.

[12] Cockayne, J., Oates, C.J., Sullivan, T. and Girolami, M. (2017). Probabilistic meshless methods for Bayesian inverse problems. Available at arXiv:1605.07811.

[13] Cockayne, J., Oates, C.J., Sullivan, T. and Girolami, M. (2017). Bayesian probabilistic numerical methods. Available at arXiv:1702.03673.

[14] Dellaportas, P. and Kontoyiannis, I. (2012). Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 133–161.

[15] Delyon, B. and Portier, F. (2016). Integral approximation by kernel smoothing. *Bernoulli* **22** 2177–2208. MR3498027

[16] Dick, J., Gantner, R.N., Gia, Q.T.L. and Schwab, C. (2016). Higher order quasi-Monte Carlo integration for Bayesian estimation. Available at arXiv:1602.07363.

[17] Gorham, J., Duncan, A.B., Vollmer, S.J. and Mackey, L. (2016). Measuring sample quality with diffusions. Available at arXiv:1611.06972.

[18] Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. In *Proceedings of the* 28*th Annual Conference on Neural Information Processing Systems*.

[19] Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the* 34*th International Conference on Machine Learning*.

[20] Hammer, H. and Tjelmeland, H. (2008). Control variates for the Metropolis–Hastings algorithm. *Scand. J. Stat.* **35** 400–414.

[21] Kanagawa, M., Sriperumbudur, B.K. and Fukumizu, K. (2016). Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Proceedings of the* 29*th Annual Conference on Neural Information Processing Systems*.

[22] Li, W., Chen, R. and Tan, Z. (2016). Efficient sequential Monte Carlo with multiple proposals and control variates. *J. Amer. Statist. Assoc.* **111** 298–313. MR3494661

[23] Liu, Q. and Lee, J.D. (2017). Black-box importance sampling. In *Proceedings of the* 21*st International Conference on Artificial Intelligence and Statistics*.

[24] Liu, Q., Lee, J.D. and Jordan, M.I. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *Proceedings of the* 33*rd International Conference on Machine Learning*.

[25] Maclaurin, D., Duvenaud, D., Johnson, M. and Adams, R.P. (2015). Autograd: Reverse-mode differentiation of native Python. Available at http://github.com/HIPS/autograd+.

[26] Meyn, S.P. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: Cambridge Univ. Press.

[27] Micchelli, C.A., Xu, Y. and Zhang, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667.

[28] Migliorati, G., Nobile, F. and Tempone, R. (2015). Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. *J. Multivariate Anal.* **142** 167–182.

[29] Mijatović, A. and Vogrinc, J. (2015). On the Poisson equation for Metropolis–Hastings chains. Available at arXiv:1511.07464.

[30] Mijatović, A. and Vogrinc, J. (2017). Asymptotic variance for random walk Metropolis chains in high dimensions: Logarithmic growth via the Poisson equation. Available at arXiv:1707.08510.

[31] Mira, A., Solgi, R. and Imparato, D. (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. *Stat. Comput.* **23** 653–662. MR3094805

[32] Narcowich, F.J., Ward, J.D. and Wendland, H. (2005). Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Math. Comp.* **74** 743–763. MR2114646

[33] Niederreiter, H. (2010). *Quasi-Monte Carlo Methods*. New York: Wiley.

[34] Oates, C.J., Cockayne, J., Briol, F.-X. and Girolami, M. (2019). Supplement to "Convergence rates for a class of estimators based on Stein's method." DOI:10.3150/17-BEJ1016SUPP.

[35] Oates, C.J., Girolami, M. and Chopin, N. (2017). Control functionals for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 695–718. MR3641403

[36] Oates, C.J., Papamarkou, T. and Girolami, M. (2016). The controlled thermodynamic integral for Bayesian model evidence evaluation. *J. Amer. Statist. Assoc.* **111** 634–645.

[37] Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. New York: Springer.

[38] Roberts, G.O. and Rosenthal, J.S. (1998). On convergence rates of Gibbs samplers for uniform distributions. *Ann. Appl. Probab.* **8** 1291–1302.

[39] Rubinstein, R.Y. and Marcus, R. (1985). Efficiency of multivariate control variates in Monte Carlo simulation. *Oper. Res.* **33** 661–677.

[40] Schölkopf, B., Herbrich, R. and Smola, A.J. (2001). A generalized representer theorem. *Lecture Notes in Comput. Sci.* **2111** 416–426.

[41] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 583–602. Berkeley, CA: Univ. California Press. MR0402873

[42] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2** 67–93.

[43] Stuart, A.M. and Teckentrup, A.L. (2018). Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Math. Comp.* To appear.

[44] Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4** 389–396. MR1366510

[45] Wendland, H. (2004). *Scattered Data Approximation*. Cambridge: Cambridge Univ. Press.

Received March 2017 and revised August 2017