Convergence Rates for a Class of Estimators Based on Stein's Method

Chris J. Oates[1], Jon Cockayne[2], François-Xavier Briol[2,3], Mark Girolami[3,4]
[1]University of Technology Sydney
[2]University of Warwick
[3]Imperial College London
[4]Alan Turing Institute

March 24, 2017

Abstract

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 one 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 much-needed 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 focuses on estimating the expectation $\int f \, \mathrm{d}\Pi$ of a test function f against a distribution Π on the basis of a finite number n of test function evaluations. 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 be approximated based on

a small number n of test function evaluations. Indeed, the intrinsic accuracy of ergodic averages, such as obtained via Markov chain Monte Carlo (MCMC) methods (Robert and Casella, 2013), can lead to unacceptably high integration error when n is small. Oates *et al.* (2016a) introduced a novel approach to this problem that can provide the "best of both worlds", in the sense of being compatible with un-normalised densities while delivering $o_P(n^{-\frac{1}{2}})$ estimation error. This was achieved by constructing estimators based on Stein's method, one component of which is integration by parts:

$$\int \nabla f \, \mathrm{d}\Pi \;\; = \;\; - \int f \cdot \nabla \log \pi \, \mathrm{d}\Pi, \tag{1}$$

subject to boundary conditions, where π is a density for Π. Stein (1972) used Eqn. 1 in the context of goodness-of-fit testing for the normal distribution. Since then, several statistical approaches have exploited Stein's method to assess how well one distribution approximates another. A comprehensive account on these approaches is provided in Ley *et al.* (2017).

A key motivation for interest in Stein's method, or more specifically Eqn. 1, is that the requirement of gradient information on the sampling distribution is often satisfied; it can be obtained in the presence of unknown normalising constants and is in fact a pre-requisite for established Langevin and Hamiltonian (Girolami and Calderhead, 2011) and emergent piecewise-deterministic (Bierkens *et al.*, 2017) MCMC methods. Sophisticated software for automatic differentiation of statistical models has been developed (e.g. Carpenter *et al.*, 2015; Maclaurin *et al.*, 2015) to circumvent hand-calculation on a per-model basis and offers considerable opportunity to exploit gradient information in the design of contemporary statistical methods.

Main Contribution: The primary contribution of this paper is to establish convergence rates for a class of estimators for $\int f \mathrm{d}\Pi$ based on Stein's method. These estimators are based on 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 explicit error rates; these enable us to compare and quantify the improvement in estimator precision relative to standard Monte Carlo methods and fill a theoretical void in the work of Oates *et al.* (2016a). The estimators considered in Oates *et al.* (2016a) can be viewed as a control functional method, and this concept is discussed next.

Control Functionals: Classical control variate methods proceed by seeking a collection of non-trivial statistics $\{\psi_i\}_{i=1}^k$, $k \in \mathbb{N}$, that satisfy $\int \psi_i \, \mathrm{d}\Pi = 0$. 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 $\{a_i\}_{i=1}^k \subset \mathbb{R}$, a variance reduction $\sigma^2(f') < \sigma^2(f)$ is obtained. An optimal choice of coefficients $\{a_i\}_{i=1}^k$ can be made using least-squares regression; for further details see e.g. Rubinstein and Marcus (1985). For specific problems it is sometimes possible to identify control variates, for example based on physical considerations (e.g. Assaraf and Caffarel, 2003). 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 Andradóttir *et al.* (1993) and for continuous state spaces, recent contributions include

Hammer and Tjelmeland (2008); Dellaportas and Kontoyiannis (2012); Li *et al.* (2016). Control variates can alternatively be constructed based on gradient information on the sampling distribution (Assaraf and Caffarel, 1999; Mira *et al.*, 2013; Oates *et al.*, 2016b).

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 Mijatović and Vogrinc (2015); Oates *et al.* (2016a) 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 $o_P(n^{-\frac{1}{2}})$. Of these two proposed solutions, Mijatović and Vogrinc (2015) 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 Oates *et al.* (2016a) 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 explicit convergence rates for the gradient-based control functional method of Oates *et al.* (2016a). It is demonstrated that integration error scales as $O_P(n^{-\frac{1}{2}-\frac{a\wedge b}{d}+\epsilon})$, 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 $\epsilon > 0$ 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 gradient-based control functionals in certain low-dimensional applications (Oates *et al.*, 2016a; Liu and Lee, 2017). Indeed, recall that the (naïve) computational cost associated with this method, i.e. the cost of solving a linear system, is $c = O(n^3)$. Thus, whilst for standard Monte Carlo methods an estimator error of $O_P(c^{-\frac{1}{2}})$ can be achieved at computational cost c, for gradient-based control functionals

$$\text{error for cost } c \quad = \quad O_P\left((c^{\frac{1}{3}})^{-\frac{1}{2}-\frac{a\wedge b}{d}+\epsilon}\right) \quad = \quad O_P\left(c^{-\frac{1}{2}+\frac{d-a\wedge b}{3d}+\epsilon}\right).$$

This demonstrates that gradient-based control functionals have 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, and are characteristic of quasi-Monte Carlo (QMC) methods (Niederreiter, 2010). Such methods can be used for estimation of posterior expectations

via ratio estimators:

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

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, i.e. 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)[1]. See Dick *et al.* (2016) for a recent study of this approach in the context of Bayesian inference for the unknown parameters of 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 integration is complicated - indeed, our results hold for any domain of integration for which an interior cone condition holds. (2) The estimator studied herein is unbiased, whereas ratio estimators of the form in Eqn. 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 Deylon and Portier (2016) and Azaïs *et al.* (2016) considered estimation problems 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 $o_P(n^{-\frac{1}{2}})$ with an explicit rate gated by a term of the form $\frac{a \wedge b}{d}$. The main distinctions with the present work are that (1) Deylon and Portier (2016) and Azaïs *et al.* (2016) treat f as known and π to be estimated; (2) the approach of Deylon and Portier (2016) and Azaïs *et al.* (2016) is not capable of estimating posterior expectations, rather the approach applies just to integrals with respect to a reference measure Λ. In contrast we do not require either of f and π to be known in closed-form, and our methods are applicable to posterior integrals.

Outline: Below in Sec. 2 we describe the gradient-based control functional method (2.1 - 2.5). Secs. 2.6 contain our main theoretical results, including the case of non-independent samples arising from a Markov chain sample path. Numerical results in Sec. 3 confirm these error rates are realised. 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. Finally the importance of our findings is discussed in Sec. 4.

[1] 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.

2 Methods

First we fix notation before introducing the gradient-based control functional estimators that are the focus of this research.

2.1 Set-up and Notation

Consider an open and bounded subset $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, 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: $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $a \wedge b := \min(a, b)$, $a_+ := \max(a, 0)$, $\|\boldsymbol{x}\|_2^2 := \sum_{i=1}^d x_i^2$, $\nabla_{\boldsymbol{x}} := [\partial/\partial x_1, \ldots, \partial/\partial x_d]^\top$, $1_A(\boldsymbol{x}) = 1$ is the indicator of the event $\boldsymbol{x} \in A$, $\mathrm{vol}(A) := \int 1_A \, \mathrm{d}\Lambda$, $\mathrm{d}\boldsymbol{x} := \mathrm{d}\Lambda(\boldsymbol{x})$ and $\mathbf{1} = [1, \ldots, 1]^\top$. 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 (*not* the π-a.s. equivalence classes of such functions). Similarly write $L^\infty(\mathcal{X})$ for the set of measurable functions for which $\sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x})| < \infty$ and 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 $\partial^{2k} g / \partial x_{i_1} \ldots \partial x_{i_k} \partial x'_{j_1} \ldots \partial x'_{j_k}$ 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 $\|f\|_2 := (\int f^2 \, \mathrm{d}\Pi)^{1/2}$ and $\|f\|_\infty := \sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x})|$ will be used. The notation $\oint_{\partial\mathcal{X}}$ will be used to denote a surface integral over $\partial\mathcal{X}$, $\boldsymbol{n}(\boldsymbol{x})$ to denote the unit normal vector to $\partial\mathcal{X}$ and $S(\mathrm{d}\boldsymbol{x})$ to denote the surface element.

2.2 Control Functionals

This section introduces the control functional method for integration, a non-parametric extension of classical control variate methods. Recall that a trade-off between random sampling and deterministic approximation has been studied on several separate occasions (Bakhvalov, 1959; Heinrich, 1995; Maire, 2003; Giles, 2013; Gobet and Surana, 2014; Oates and Girolami, 2016). Our starting point is, in a similar vein, to establish a trade-off between random sampling and *stochastic* approximation methods.

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 (i.e. infinitely differentiable). Consider an independent sample from Π, denoted $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$, which is partitioned into two 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 $s_{f,\mathcal{D}_0} \in L^2(\mathcal{X}, \Pi)$ to f, based on $\mathcal{D}_0$. Stochasticity in $s_{f,\mathcal{D}_0}$ is induced via the sampling distribution of elements in $\mathcal{D}_0$. The integral $\int s_{f,\mathcal{D}_0} \, \mathrm{d}\Pi$ is required to be analytically tractable; we will return to this point later.

The estimators that we study take the form

$$m_{f,\mathcal{D}_0,\mathcal{D}_1} := \frac{1}{n-m}\sum_{i=m+1}^{n} f(\boldsymbol{x}_i) - \underbrace{\left(s_{f,\mathcal{D}_0}(\boldsymbol{x}_i) - \int s_{f,\mathcal{D}_0}\,\mathrm{d}\Pi\right)}_{(*)}. \tag{4}$$

Such sample-splitting estimators are unbiased, i.e. $\mathbb{E}_{\mathcal{D}_1}[m_{f,\mathcal{D}_0,\mathcal{D}_1}] = \int f\,\mathrm{d}\Pi$, 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 $\mathbb{V}_{\mathcal{D}_1}[m_{f,\mathcal{D}_0,\mathcal{D}_1}] = (n-m)^{-1}\sigma^2(f - s_{f,\mathcal{D}_0})$. This formulation encompasses control variates as a special case where $s_{f,\mathcal{D}_0} = a_1\psi_1 + \cdots + a_k\psi_k$, $k \in \mathbb{N}$, and $\mathcal{D}_0$ are used to select suitable values for the coefficients $\{a_i\}_{i=1}^k$ (see e.g. Rubinstein and Marcus, 1985).

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

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

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 $m/n \to r$ for some $r \in (0, 1]$ as $n \to \infty$; this will be assumed in the sequel. The case $r = 1$ is essentially numerical quadrature; this will be a special case of the general results that we obtain in this paper.

When Π is given via an un-normalised density, this framework can only be exploited if it is possible to construct approximations $s_{f,\mathcal{D}_0}$ whose integrals $\int s_{f,\mathcal{D}_0}\,\mathrm{d}\Pi$ are available in closed-form. If and when this is possible, $(*)$ in Eqn. 4 is known as a *control functional*. Oates *et al.* (2016a) showed how to build a flexible class of control functionals based on Stein's method; the key points are presented next.

2.3 Stein Operators

To begin, we make the following assumptions:

(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 available only in un-normalised form, being equal to $\nabla_{\boldsymbol{x}}\pi(\boldsymbol{x})/\pi(\boldsymbol{x})$. This fact is the basis for several techniques in computational statistics,

including score matching (Hyvärinen, 2005) and differential-geometric MCMC (Girolami and Calderhead, 2011). Consider the following Stein operator (Ley *et al.*, 2017):

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

This definition can be motivated in several ways, including via Schrödinger Hamiltonians (Assaraf and Caffarel, 1999) and via the generator method of Barbour (1988) applied to an overdamped Langevin diffusion (Gorham and Mackey, 2015). Our definition generalises, to multiple dimensions, the original construction of Stein (1972).

For functional approximation we follow Assaraf and Caffarel (1999); Mira *et al.* (2013); Oates *et al.* (2016a) and study approximations of the form;

$$s_{f,\mathcal{D}_0}(\cdot) \;\; := \;\; \beta + \mathbb{S}_\pi[\boldsymbol{\phi}](\cdot) \tag{7}$$

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 (Eqn. 1) can be applied to obtain

$$\int \mathbb{S}_\pi[\boldsymbol{\phi}] \, \mathrm{d}\Pi = 0, \quad \text{and so} \quad \int s_{f,\mathcal{D}_0} \, \mathrm{d}\Pi = \beta.$$

Thus, for this class of functions, $\int s_{f,\mathcal{D}_0} \mathrm{d}\Pi$ permits a trivial closed-form and $\mathbb{S}_\pi[\boldsymbol{\phi}]$ is a control functional (i.e. integrates to 0). The choice of $\boldsymbol{\phi}$ presents us with an infinite-dimensional regression problem; this justifies the *functional* nomenclature. The choice of Stein operator is not unique and some alternatives are listed in Gorham *et al.* (2016). However, a preliminary investigation suggested that the Stein operator in Eqn. 6 yields strong performance for the methods discussed next. 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 control functionals as approximation in a Hilbert space $\mathcal{H}_+$ whose elements are contained in the set $L^2(\mathcal{X}, \Pi)$. This construction first appeared in Oates *et al.* (2016a) and has subsequently been explored by Liu *et al.* (2016); Chwialkowski *et al.* (2016) in the context of goodness-of-fit assessment, Liu and Wang (2016) in the context of gradient descent, Liu and Feng (2016) in the context of variational inference, Liu and Lee (2017) in the context of importance sampling, Wang and Liu (2016) in the context of adversarial learning and Gorham and Mackey (2017) in the context of Markov chain convergence assessment.

2.4.1 Boundary Conditions

Specification of a gradient-based control functional $\mathbb{S}_\pi[\boldsymbol{\phi}]$ is equivalent to specification of $\boldsymbol{\phi}$. Following Oates *et al.* (2016a) we restrict each component function $\phi_i : \mathcal{X} \rightarrow \mathbb{R}$ 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), i.e. 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 Berlinet and Thomas-Agnan, 2004, for background).

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

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

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 $i = 1, \ldots, d$ 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 $\boldsymbol{x} \in \partial\mathcal{X}$.

(A2') $\pi \in \mathcal{Q}(k)$

(A4) $\pi \in \mathcal{R}(k)$

The assumption (A2') was first discussed in Chwialkowski *et al.* (2016) and is stronger than (A2). A constructive approach to (A4) starts with an arbitrary RKHS $\tilde{\mathcal{H}}$ with reproducing kernel $\tilde{k}$. Let $B : \tilde{\mathcal{H}} \to \mathrm{im}(B)$ be a linear operator which enforces the boundary conditions. e.g. $B\varphi(\boldsymbol{x}) := \delta(\boldsymbol{x})\varphi(\boldsymbol{x})$, where $\delta(\cdot)$ is a deterministic, smooth function such that $\pi(\cdot)\delta(\cdot)$ vanishes on $\partial\mathcal{X}$. Then $\mathcal{H} = \mathrm{im}(B)$ is a RKHS, e.g. 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 Sec. 3.

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 $\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle \phi_i, \phi_i' \rangle_{\mathcal{H}}$.

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

This shows that $\mathbb{S}_\pi[\boldsymbol{\phi}]$ is a control functional. (Proofs are in the electronic supplement.)

2.4.2 $\mathcal{H}_0$, the Control Functional Space

Consider the set $\mathcal{H}_0 := \mathbb{S}_\pi[\mathcal{H}^d]$, whose elements $\mathbb{S}_\pi[\boldsymbol{\phi}]$ for $\boldsymbol{\phi} \in \mathcal{H}^d$ result from application of the Stein operator $\mathbb{S}_\pi$ to the Hilbert space $\mathcal{H}^d$. Oates *et al.* (2016a, Thm. 1) showed that $\mathcal{H}_0$ can be endowed with the gradient-based reproducing kernel

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

From (A1,2',3) it follows that $\mathcal{H}_0 \subseteq C^{a \wedge b}(\mathcal{X} \cup \partial\mathcal{X})$. Moreover, under (A1,2',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.

2.5 A Class of Control Functionals

The aim of this section is to be specific about how β and $\boldsymbol{\phi}$ are selected.

2.5.1 $\mathcal{H}_+$, the Approximation Space

Write $\mathcal{H}_{\mathbb{R}}$ for the RKHS characterised by the kernel $k_{\mathbb{R}}(\boldsymbol{x}, \boldsymbol{x}') = c$, $c > 0$, for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, consisting of constant functions. 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 structure of a vector space, with addition operator $(\beta+\psi)+(\beta'+\psi') = (\beta+\beta')+(\psi+\psi')$ and multiplication operator $\lambda(\beta+\psi) = (\lambda\beta)+(\lambda\psi)$, each well-defined due to uniqueness of the representation $f = \beta + \psi$, $f' = \beta' + \psi'$ with $\beta, \beta' \in \mathcal{H}_{\mathbb{R}}$ and $\psi, \psi' \in \mathcal{H}_0$. In addition, equip $\mathcal{H}_+$ with the norm $\|f\|^2_{\mathcal{H}_+} := \|\beta\|^2_{\mathcal{H}_{\mathbb{R}}} + \|\psi\|^2_{\mathcal{H}_0}$, again, well-defined. It can be shown that $\mathcal{H}_+$ is a RKHS with kernel $k_+(\boldsymbol{x}, \boldsymbol{x}') := k_{\mathbb{R}}(\boldsymbol{x}, \boldsymbol{x}') + k_0(\boldsymbol{x}, \boldsymbol{x}')$ (Berlinet and Thomas-Agnan, 2004, Thm. 5, p24). From (A1-3) it follows that $\mathcal{H}_+ \subseteq C^{a \wedge b}(\mathcal{X})$.

2.5.2 Constructive Approximation

To realise the gradient-based control functional method we formulate the choice of β and $\boldsymbol{\phi}$ as a least-squares optimisation problem:

$$s_{f,\mathcal{D}_0} \ := \ \underset{h \in \mathcal{H}_+}{\arg\min} \ \|h\|^2_{\mathcal{H}_+} \ \text{ s.t. } \ h(\boldsymbol{x}_i) = f(\boldsymbol{x}_i) \quad \forall\, i = 1, \ldots, m.$$

By the representer theorem (Schölkopf *et al.*, 2001) we have

$$s_{f,\mathcal{D}_0}(\boldsymbol{x}) = \sum_{i=1}^{m} a_i k_+(\boldsymbol{x}, \boldsymbol{x}_i)$$

where the coefficients $\boldsymbol{a} = [a_1, \ldots, a_m]^\top$ are the solution of the linear system $\boldsymbol{K}_0 \boldsymbol{a} = \boldsymbol{f}_0$ where $\boldsymbol{K}_0 \in \mathbb{R}^{m \times m}$, $[\boldsymbol{K}_0]_{i,j} = k_+(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{f}_0 \in \mathbb{R}^{m \times 1}$, $[\boldsymbol{f}_0]_i = f(\boldsymbol{x}_i)$. In situations where $\boldsymbol{K}_0$ is not full-rank, we define $s_{f,\mathcal{D}_0} \equiv 0$. Numerical inversion of this system is associated with a $O(m^3)$ cost and may require additional numerical regularisation; this is standard and briefly discussed in Section 2.6.3.

2.6 Theoretical Results

Our novel analysis, next, builds on results from the scattered data approximation literature (Wendland, 2004) and the study of the stability properties of Markov chains (Meyn and Tweedie, 2009).

2.6.1 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 $\sigma^2(f - s_{f,\mathcal{D}_0})$ 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*, i.e. 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

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

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

The purpose of (A7) is to rule our 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*

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

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

Our first main result can now be stated:

Theorem 1. *Assume (A1,2',3-7). Then there exists $h > 0$, independent of m, n, such that the estimator $m_{f,\mathcal{D}_0,\mathcal{D}_1}$ is an unbiased estimator of $\int f \, \mathrm{d}\Pi$ with*

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

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

This result demonstrates that control functionals 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 Oates *et al.* (2016a) 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.

A second consequence of this result is to show that efficiency is limited by the rougher of the density π and the test function f. An important application of these estimators is the computation of marginal likelihood (Oates *et al.*, 2016b) via thermodynamic integration; in that setting the regularity of the test function coincides with the regularity of the density, so $a = b$. Thus for marginal likelihood computation, estimator accuracy is intrinsic to the density that is being marginalised.

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

2.6.2 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 all n samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ 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 $\mathcal{D}_0$ 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 ($\boldsymbol{x}_i = \boldsymbol{x}_j$; 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 $\boldsymbol{K}_0$ to become singular and the control functional to become trivial $s_{f,\mathcal{D}_0} = 0$. 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 $\mathcal{D}_0$ 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 Oates *et al.* (2016a), 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.

2.6.3 Commentary

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

On the Assumptions: Assumptions (A1,2',3,7) are not unduly restrictive. The boundary condition (A4) has previously been discussed in Oates *et al.* (2016a). 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 and was also assumed in the related work of Deylon and Portier (2016) and Azaïs *et al.* (2016).

Our analysis also relied on (A6), i.e. 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 (A2',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 Steinwart (2001), who showed that many widely-used kernels are c-universal on compact $\mathcal{X} \cup \partial\mathcal{X} \subseteq \mathbb{R}^d$. Indeed, Prop. 1 of Micchelli *et al.* (2006) 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 Narcowich *et al.* (2005); Kanagawa *et al.* (2016).

The last assumption that deserves discussion 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, Roberts and Rosenthal (1998b) constructed an example where a 'pinch point' in the domain causes a Gibbs sampler targeting a uniform distribution to fail to be geometrically ergodic; their construction violates our (A7). As a concrete example, under our assumptions, a Metropolis-Hastings sampler is uniformly ergodic whenever the proposal density is bounded below on $\mathcal{X} \times \mathcal{X}$.

On the Results: The intuition for the results in Thms. 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) Firstly, one can simply re-define $s_{f,\mathcal{D}_0} \equiv 0$ whenever $h \geq h_0$, i.e. when the states $\mathcal{D}_0$ are poorly spaced we revert to the usual Monte Carlo estimator. (ii) Secondly, 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. This explains our presentational approach.

The condition $h_{\mathcal{D}_0} < h$ suggests that the asymptotics hold in the same regime where a ratio of QMC methods could also be successful. However, as explained in Sec. 1, the method of Oates *et al.* (2016a) 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 Sec. 11.7 of Wendland (2004) where an overview of the strengths and weaknesses of results in the scattered data approximation literature is provided. As discussed in Sec. 3.1 next, empirical results suggest that these results are either optimal or close to optimal.

On Computation: It is important to emphasise the ease with which gradient-based control functional estimators can be implemented. Indeed, there is an explicit closed-form expression:

$$m_{f,\mathcal{D}_0,\mathcal{D}_1} = \frac{1}{n-m}\mathbf{1}^\top(\boldsymbol{f}_1 - \boldsymbol{K}_{1,0}\boldsymbol{K}_0^{-1}\boldsymbol{f}_0) + c\mathbf{1}^\top\boldsymbol{K}_0^{-1}\boldsymbol{f}_0 \tag{9}$$

where $\boldsymbol{f}_1 \in \mathbb{R}^{n-m\times 1}$, $[\boldsymbol{f}_1]_i = f(\boldsymbol{x}_{m+i})$, $\boldsymbol{K}_{1,0} \in \mathbb{R}^{n-m\times m}$ and $[\boldsymbol{K}_{1,0}]_{i,j} = k_+(\boldsymbol{x}_{m+i}, \boldsymbol{x}_j)$. The last term in Eqn. 9 corresponds to $\int s_{f,\mathcal{D}_0}\mathrm{d}\Pi$. A full derivation can be found in Oates *et al.* (2016a) and an implementation that uses automatic differentiation, called `control_func.m`, is available on the Matlab File Exchange.

Numerical solution of the linear system $\boldsymbol{K}_0^{-1}\boldsymbol{f}_0$ can benefit from regularisation; such considerations are standard (e.g. Shawe-Taylor and Cristianini, 2004). In addition there is an active research effort to obtain computationally efficient approximations to kernel matrix inverses (e.g. Rudi *et al.*, 2015). All of these methods can be used in combination with control functionals and could be used to mitigate the computational cost associated with the solution of a linear system. Recall from Sec. 1 that when $a \wedge b > d$, the at most cubic computational cost of solving the linear system is outweighed by gains in estimator precision in the setting of a finite-but-large computational budget.

Smoothness	Function	Expression
$b = 1$	$\varphi(z)$	$(4z+1)(1-z)_+^4$
(i.e. C^2)	$\varphi^{(1)}(z)$	$-20z(1-z)_+^3$
	$\varphi^{(2)}(z)$	$20(4z-1)(1-z)_+^2$
$b = 3$	$\varphi(z)$	$(35z^2+18z+3)(1-z)_+^6$
(i.e. C^4)	$\varphi^{(1)}(z)$	$-56z(5z+1)(1-z)_+^5$
	$\varphi^{(2)}(z)$	$56(35z^2-4z-1)(1-z)_+^4$
$b = 5$	$\varphi(z)$	$(32z^3+25z^2+8z+1)(1-z)_+^8$
(i.e. C^6)	$\varphi^{(1)}(z)$	$-22z(16z^2+7z+1)(1-z)_+^7$
	$\varphi^{(2)}(z)$	$22(160z^3+15z^2-6z-1)(1-z)_+^6$

Table 1: Formulae for derivatives of the compact support radial functions of Wendland (1995). Note that z_+^k should be interpreted as $(z_+)^k$ rather than $(z^k)_+$.

3 Numerical Results

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

3.1 Convergence Assessment

To construct a test-bed for the theoretical results we considered the simple case where Π is the uniform distribution on $[0,1]^d$. For estimation we restricted to radial kernels of the form

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

where $h > 0$ is a bandwidth parameter and φ is positive definite function, to be specified. In particular we consider the φ to be the compact support functions of variable smoothness studied in Wendland (1995). Explicit formulae for φ and its derivatives are contained in Table 1.

To enforce (A4) we considered kernels of the form $k(\boldsymbol{x}, \boldsymbol{x}') = \delta(\boldsymbol{x})\delta(\boldsymbol{x}')\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')$ based on the smooth function

$$\delta(\boldsymbol{x}) = \prod_{i=1}^{d} x_i(1 - x_i)$$

which vanishes on $\partial\mathcal{X}$. This construction ensures k inherits the smoothness parameter b of the first kernel $\tilde{k}$, shown in Table 1. To derive the kernel k_+, let $r(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|$ so that

$\tilde{k} \equiv \tilde{k}(r)$. Then the identities

$$\begin{aligned} \nabla_{\boldsymbol{x}} \tilde{k} &= -\nabla_{\boldsymbol{x}'} \tilde{k} &= \frac{\boldsymbol{x} - \boldsymbol{x}'}{hr} \varphi^{(1)}\left(\frac{r}{h}\right) \\ \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} \tilde{k} &= & -\frac{1}{h^2} \varphi^{(2)}\left(\frac{r}{h}\right) \end{aligned}$$

and

$$\begin{aligned} \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}') &= [\nabla_{\boldsymbol{x}} \delta(\boldsymbol{x})] \delta(\boldsymbol{x}') \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') + \delta(\boldsymbol{x}) \delta(\boldsymbol{x}') \nabla_{\boldsymbol{x}} \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') \\ \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') &= \delta(\boldsymbol{x}) [\nabla_{\boldsymbol{x}'} \delta(\boldsymbol{x}')] \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') + \delta(\boldsymbol{x}) \delta(\boldsymbol{x}') \nabla_{\boldsymbol{x}'} \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') \\ \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') &= [\nabla_{\boldsymbol{x}} \delta(\boldsymbol{x})] \cdot [\nabla_{\boldsymbol{x}'} \delta(\boldsymbol{x}')] \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') + \delta(\boldsymbol{x}') [\nabla_{\boldsymbol{x}} \delta(\boldsymbol{x})] \cdot [\nabla_{\boldsymbol{x}'} \tilde{k}(\boldsymbol{x}, \boldsymbol{x}')] \\ &\quad + \delta(\boldsymbol{x}) [\nabla_{\boldsymbol{x}'} \delta(\boldsymbol{x}')] \cdot [\nabla_{\boldsymbol{x}} \tilde{k}(\boldsymbol{x}, \boldsymbol{x}')] + \delta(\boldsymbol{x}) \delta(\boldsymbol{x}') \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} \tilde{k}(\boldsymbol{x}, \boldsymbol{x}') \end{aligned}$$

can be used to obtain the kernel $k_+ = c + k_0$ in closed-form, in combination with Eqn. 8. For the uniform distribution Π, the score function $\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x}) = \mathbf{0}$ on $\mathcal{X}$, so that we have $k_+(\boldsymbol{x}, \boldsymbol{x}') = c + \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}')$. For the experiments $h = 0.1$ and $c = 1$ were fixed.

(A1,2',3-5,7) are satisfied in this experiment. Thus, for $f \in \mathcal{H}_+$, i.e. (A6), Thm. 1 entails a mean squared integration error of $O(n^{-1-2\frac{b}{d}+\epsilon})$, since $\pi(\boldsymbol{x}) = 1 \in C^{a+1}$ for all $a \in \mathbb{N}_0$. To assess the sharpness of this theoretical result we considered

$$\mathbb{E}_{\mathcal{D}_0} \left[\sup_{\|f\|_{\mathcal{H}_+} \leq 1} \left| m_{f,\mathcal{D}_0,\emptyset} - \int f \mathrm{d}\Pi \right|^2 \right]. \tag{10}$$

This quantifies performance of the simplified (biased) estimator $m_{f,\mathcal{D}_0,\emptyset} = c\mathbf{1}^\top \boldsymbol{K}_0^{-1} \boldsymbol{f}_0$ that allocates all function evaluations to $\mathcal{D}_0$ and none to $\mathcal{D}_1$. Indeed, from the standard control variate argument in Eqn. 5, the performance of the general estimator $m_{f,\mathcal{D}_0,\mathcal{D}_1}$ is always "one Monte Carlo" better than the performance of $m_{f,\mathcal{D}_0,\emptyset}$, so that we can focus on the simplified estimator for assessment. In particular, the theoretical results in this paper entail a rate of $O(n^{-2\frac{b}{d}+\epsilon})$ for Eqn. 10 and it is this rate which is assessed. The quantity inside the expectation in Eqn. 10 is identical to the (squared) maximum mean discrepancy (MMD; Smola *et al.*, 2007) and can be computed in closed-form as $c - c^2 \mathbf{1}^\top \boldsymbol{K}_0^{-1} \mathbf{1}$. This enables us to present results which are not dependent on the specific choice of test function $f \in \mathcal{H}_+$. The MMD in this case first appeared in 2015 in arXiv:1410.2392v2 and was later termed the *kernelised Stein discrepancy* (KSD) in Chwialkowski *et al.* (2016); Liu *et al.* (2016). The recent work of Gorham and Mackey (2017) establishes that, for certain choices of k, the KSD controls weak convergence to Π.

To estimate Eqn. 10 we repeatedly generated collections of n independent uniform random variables $\{\boldsymbol{x}_i\}_{i=1}^n$ and evaluated the MMD on this set. The procedure was repeated several times to obtain a Monte Carlo estimate of Eqn. 10 and standard errors were recorded. Results are displayed in Fig. 1. These results clearly demonstrate the predicted deterioration of the estimator as the dimension d is increased. However, the fastest theoretical rates of convergence were not observed at the values of n considered, which serves as a reminder on the limited relevance of asymptotic arguments to finite sample behaviour.

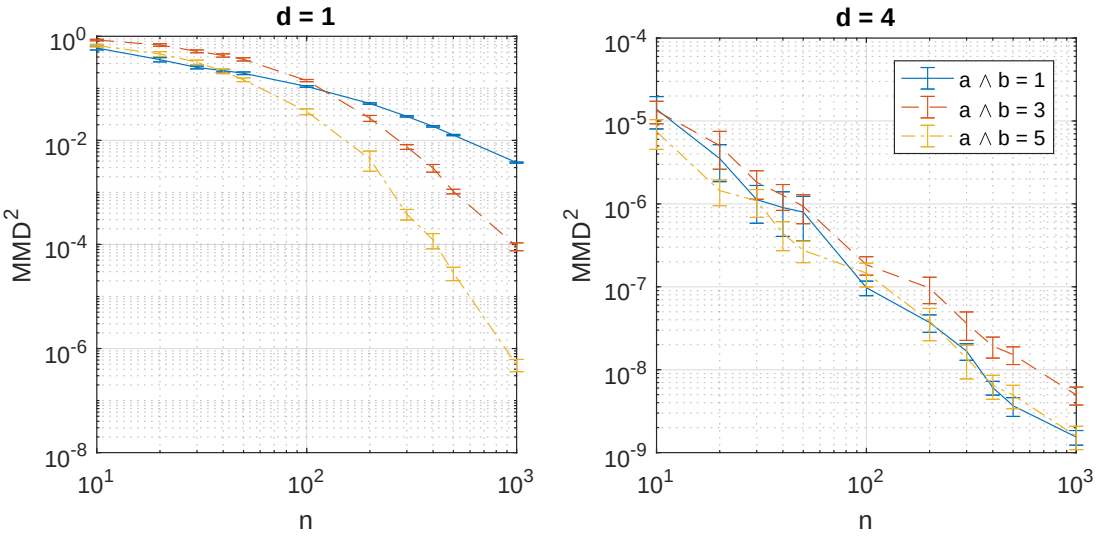


Figure 1: Simulation Results. The (squared) maximum mean discrepancy (MMD) was assessed based on a range of smoothness parameters ($b = 1, 3, 5$) and dimensions ($d = 1, 4$). [Interest here is in the slope of each curve; since the MMD is defined in terms of the kernel, which depends in turn on $a \wedge b$, the absolute values cannot be compared.]

3.2 Application to Partial Differential Equations

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

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

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 (Lan *et al.*, 2015). 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 $\boldsymbol{x}_{i,j}$, $i, j = 1, \ldots, M$. The observation model $p(\boldsymbol{y}|\boldsymbol{\theta})$ takes the form $\boldsymbol{y} = \{y_{i,j}\}$ where $y_{i,j} = w(\boldsymbol{x}_{i,j}) + \epsilon_{i,j}$ and $\epsilon_{i,j}$ are independent normal random variables with standard deviation $\sigma = 0.1$.

Consider now the Bayesian approach to this inverse problem (Stuart, 2010), in which the field κ is endowed with a prior distribution of the form

$$\log \kappa(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{d} \theta_i \kappa_i(\boldsymbol{x}),$$

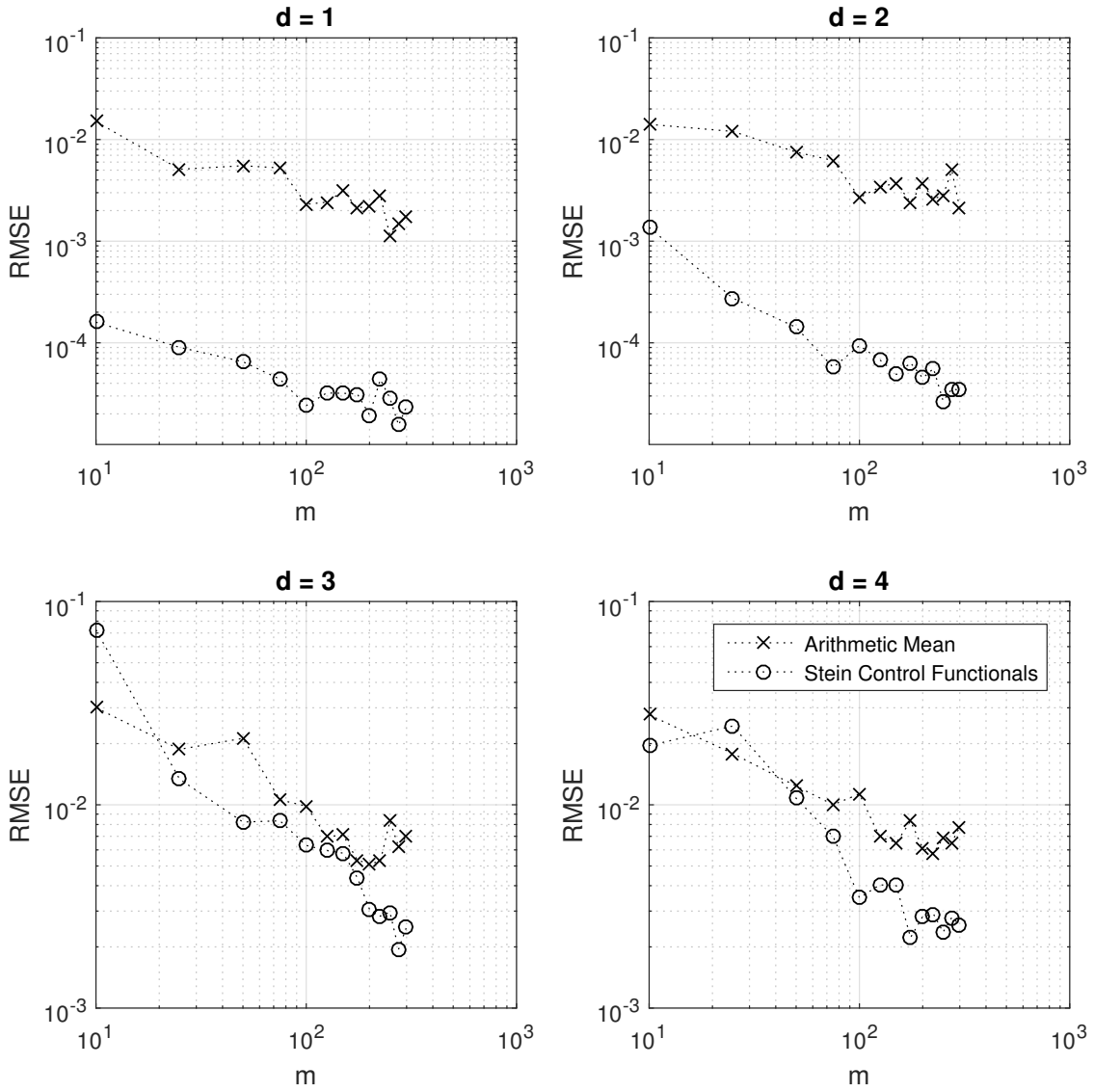


Figure 2: Experimental Results. An experiment to approximate the posterior mean of the parameters $\boldsymbol{\theta} \in [-10, 10]^d$ that govern a permeability field in a porous medium. The figure shows root mean square error (RMSE) for (i) the arithmetic mean computed based on $2m$ posterior samples, and (ii) the control functional 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; $b = 1$ in our notation.]

where $d \in \mathbb{N}$ and κ_i are orthonormal basis functions (Adler, 1981). For this illustration we follow Lan *et al.* (2015); Del Moral *et al.* (2016); Stuart and Teckentrup (2017) who each took a truncated Fourier basis. Our aim here is to obtain accurate estimates for the posterior mean of the parameter $\boldsymbol{\theta}$. For the inference we imposed a uniform prior $p(\boldsymbol{\theta}) \propto 1$ over the

domain $[-10, 10]^d$. 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 Cockayne *et al.* (2016) as our numerical integrator. Automatic differentiation was performed using the `Autograd` package (Maclaurin *et al.*, 2015).

The key assumptions of our theory are verified. Smoothness of the prior, together with ellipticity, imply (A1) holds for all $a \in \mathbb{N}$. (A2',5) hold since the prior and likelihood are well-behaved. (A7) holds since the domain of integration is hyper-cuboid. Samples from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ were obtained using a Metropolis-adjusted Langevin sampler with fixed proposal covariance; this ensures that (A8) is satisfied. Remaining assumptions are satisfied by construction of the kernel k: Following the approach outlined in Section 2.4.2, 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 is 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 $\theta_i = \pm 10$).

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 a standard arithmetic mean taken over all n samples. 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 2. 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 shown for the control functional method.

4 Conclusion

This paper has established novel asymptotic analysis for gradient-based control functionals, inspired from 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 of gradient-based control functionals 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 gradient-based control functional 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 Migliorati *et al.* (2015). 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 stronger assumptions could reasonably be placed on the function space. Active subspace methods (Constantine *et al.*, 2014), sparse grid methods (Nobile *et al.*, 2008), and weighted function spaces (Niederreiter, 2010) may have an important role to play in future research efforts. (iv) In the case of the simplified estimator, with $\mathcal{D}_1 = \emptyset$, it is known that the states $\{\boldsymbol{x}_i\}_{i=1}^n$ need not target Π. Recent work in Briol *et al.* (2017) might help to address the optimal choice of sampling distribution $\Pi' \neq \Pi$ in this context. (v) Recent work in Liu and Lee (2017) imposed an additional constraint on the coefficients a_i in Sec. 2.5.2. It would be interesting to extend our analysis to this context.

While we focused on control functionals, our analysis may have consequences for related ongoing work on Stein's method (Ley *et al.*, 2017; Gorham and Mackey, 2017).

Acknowledgements: The authors wish to thank Aretha Teckentrup, Motonobu Kanagawa and Lester Mackey for their useful feedback. CJO was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers. FXB was supported by EPSRC [EP/L016710/1]. MG was supported by the EPSRC grants [EP/J016934/3, EP/K034154/1, EP/P020720/1], an EPSRC Established Career Fellowship, the EU grant [EU/259348], a Royal Society Wolfson Research Merit Award and the Lloyds Register Foundation Programme on Data-Centric Engineering.

References

Adler, R. (1981) *The Geometry of Random Fields*. John Wiley and Sons, New York.

Andradóttir, S., Heyman, D. P. and Ott, T. J. (1993) Variance reduction through smoothing and control variates for Markov Chain simulations. *ACM T. M. Comput. S.*, **3**, 167-189.

Assaraf, R. and Caffarel, M. (1999) Zero-Variance Principle for Monte Carlo Algorithms. *Phys. Rev. Lett.*, **83**(23), 4682-4685.

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.

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

Bakhvalov, N.S. (1959) On the approximate calculation of multiple integrals (in Russian). *Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem.*, **4**, 3-18.

Barbour, A. D. (1988) Stein's method and Poisson process convergence. *J. Appl. Probab.*, **25A**, 175-184.

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

Bierkens, J., Bouchard-Côté, A., Doucet, A., Duncan, A.B., Fearnhead, P., Roberts, G. and Vollmer, S.J. (2017) Piecewise Deterministic Markov Processes for Scalable Monte Carlo on Restricted Domains. arXiv:1701.04244.

Briol, F.X., Oates, C.J., Cockayne, J. and Girolami, M. (2017) On the Sampling Problem for Kernel Quadrature. In submission.

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++. arXiv:1509.07164.

Chwialkowski, K., Strathmann, H. and Gretton, A. (2016) A Kernel Test of Goodness of Fit. Proceedings of the 33rd International Conference on Machine Learning.

Cockayne, J., Oates, C. J., Sullivan, T. and Girolami, M. (2016) Probabilistic Numerical Methods for PDE-constrained Bayesian Inverse Problems. Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering.

Constantine, P. G., Dow, E. and Wang, Q. (2014) Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM J. Sci. Comput.*, **36**(4), A1500-A1524.

Del Moral, P., Jasra, A., Law, K. and Zhou, Y. (2016) Multilevel Sequential Monte Carlo Samplers for Normalizing Constants. arXiv:1603.01136.

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

Delyon, B. and Portier, F. (2016). Integral approximation by kernel smoothing. *Bernoulli*, **22**(4), 2177-2208.

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

Giles, M. B. (2013) Multilevel Monte Carlo Methods. In *Monte Carlo and Quasi-Monte Carlo Methods* (pp. 83-103). Springer, Berlin Heidelberg.

Girolami, M. and Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B. Stat. Methodol.*, **73**, 1-37.

Gobet, E. and Surana, K. (2014) A new sequential algorithm for L2-approximation and application to Monte-Carlo integration. hal-00972016.

Gorham, J. and Mackey, L. (2015) Measuring Sample Quality with Stein's Method. Proceedings of the 28th Annual Conference on Neural Information Processing Systems.

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

Gorham, J. and Mackey, L. (2017) Measuring sample quality with kernels. arXiv:1703.01717.

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

Heinrich, S. (1995) Variance reduction for Monte Carlo methods by means of deterministic numerical computation. *Monte Carlo Methods Appl.*, **1**(4), 251-278.

Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, **6**, 695-709.

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

Lan, S., Bui-Thanh, T., Christie, M. and Girolami, M. (215) Emulation of Higher-Order Tensors in Manifold Monte Carlo Methods for Bayesian Inverse Problems. *J. Comput. Phys.*, **308**, 81-101.

Ley, C., Reinert, G. and Swan, Y. (2017) Stein's method for comparison of univariate distributions. *Probab. Surveys*, **14**, 1-52.

Li, W., Chen, R. and Tan, Z. (2016) Efficient Sequential Monte Carlo with Multiple Proposals and Control Variates. *J. Am. Stat. Assoc.*, **111**(513), 298-313.

Liu, Q., Lee, J.D. and Jordan, M.I. (2016) A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. Proceedings of the 33rd International Conference on Machine Learning.

Liu, Q. and Wang, D. (2016) Stein variational gradient descent: A general purpose Bayesian inference algorithm. Proceedings of the 29th Annual Conference on Neural Information Processing Systems.

Liu, Q. and Feng, Y. (2016) Two Methods For Wild Variational Inference. arXiv:1612.00081.

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

Maclaurin, D., Duvenaud, D., Johnson, M., Adams, R.P. (2015) Autograd: Reverse-mode differentiation of native Python. `http://github.com/HIPS/autograd`.

Maire, S. (2003) Reducing variance using iterated control variates, *J. Stat. Comput. Sim.*, **73**, 1-29.

Meyn, S.P. and Tweedie, R.L. (2009) *Markov Chains and Stochastic Stability, 2ed.* Cambridge University Press.

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

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.

Mira, A., Solgi, R. and Imparato, D. (2013) Zero Variance Markov Chain Monte Carlo for Bayesian Estimators. *Stat. Comput.*, **23**, 653-662.

Mijatović, A. and Vogrinc, J. (2015) On the Poisson equation for Metropolis-Hastings chains. arXiv:1511.07464.

Narcowich, F., Ward, J. and Wendland, H. (2005) Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Math. Comput.*, **74**(250), 743-763.

Niederreiter, H. (2010) *Quasi-Monte Carlo Methods*. John Wiley & Sons, Ltd.

Nobile, F., Tempone, R. and Webster, C.G. (2008) A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, **46**(5), 2309-2345.

Oates, C. J., Chopin, N. and Girolami, M. (2016a) Control Functionals for Monte Carlo Integration. *J. R. Statist. Soc. B. Stat. Methodol.*, to appear.

Oates, C. J., Papamarkou, T. and Girolami, M. (2016b) The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation. *J. Am. Stat. Assoc.*, **111**(514):634-645.

Oates, C. J. and Girolami, M. (2016) Control Functionals for Quasi Monte Carlo Integration. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics.

Robert, C. and Casella, G. (2013) *Monte Carlo statistical methods*. Springer.

Roberts, G. O. and Rosenthal, J. S. (1998b) On convergence rates of Gibbs samplers for uniform distributions. *Ann. Appl. Prob.*, **8**(4), 1291-1302.

Rubinstein, R. Y. and Marcus, R. (1985) Efficiency of Multivariate Control Variates in Monte Carlo Simulation. *Oper. Res.*, **33**, 661-677.

Rudi, A., Camoriano, R. and Rosasco, L. (2015) Less is More: Nyström Computational Regularization. Proceedings of the 28th Annual Conference on Neural Information Processing Systems.

Schölkopf, B., Herbrich, R. and Smola, A. J. (2001) A Generalized Representer Theorem. *Lect. Notes Comput. Sc.*, **2111**, 416-426.

Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Smola, A., Gretton, A., Song, L. and Schölkopf, B. (2007) A Hilbert space embedding for distributions. In *Proc. 18th International Conference on Algorithmic Learning Theory*, 13-31. Springer-Verlag, Berlin.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, **2**, 583-602.

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

Stuart, A.M. (2010) Inverse problems: a Bayesian perspective. *Acta Numer.*, **19**, 451-559.

Stuart, A. M. and Teckentrup, A. L. (2017) Posterior Consistency for Gaussian Process Approximations of Bayesian Posterior Distributions. *Math. Comput.*, to appear.

Wang, D. and Liu, Q. (2016) Learning to draw samples: With application to amortized MLE for generative adversarial learning. arXiv:1611.01722.

Wendland, H. (1995) Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, **4**(1), 389-396.

Wendland, H. (2004) *Scattered Data Approximation*. Cambridge University Press.

5 Supplement

In this electronic supplement we establish correctness of the theoretical results in the main text.

Proof of Lemma 1. From the Moore-Aronszajn theorem (Aronszajn, 1950), $\mathcal{H}^d$ can be expressed as

$$\mathcal{H}^d = \left\{ \boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d \text{ s.t. } \phi_i(\boldsymbol{x}) = \sum_{j=1}^{\infty} a_{i,j} k(\boldsymbol{x}, \boldsymbol{x}_{i,j}) \text{ where } \sum_{j=1}^{\infty} a_{i,j}^2 k(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{i,j}) < \infty \right\}.$$

Note that from (A3) the kernel k is continuous and thus bounded on the compact set $\mathcal{X} \cup \partial\mathcal{X}$. It follows that the series representation given in the Moore-Aronszajn theorem is uniformly convergent. Then, for $\boldsymbol{\phi} \in \mathcal{H}^d$ represented in Moore-Aronszajn form,

$$\begin{aligned}
\int \mathbb{S}_\pi[\boldsymbol{\phi}](\boldsymbol{x}) \Pi(\mathrm{d}\boldsymbol{x}) &= \int \sum_{i=1}^{d} (\nabla_{x_i} + \nabla_{x_i} \log \pi(\boldsymbol{x})) \sum_{j=1}^{\infty} a_{i,j} k(\boldsymbol{x}, \boldsymbol{x}_{i,j}) \, \Pi(\mathrm{d}\boldsymbol{x}) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{\infty} a_{i,j} \int (\nabla_{x_i} + \nabla_{x_i} \log \pi(\boldsymbol{x})) k(\boldsymbol{x}, \boldsymbol{x}_{i,j}) \, \Pi(\mathrm{d}\boldsymbol{x}) \\
&= \sum_{i=1}^{d} \sum_{j=1}^{\infty} a_{i,j} \int_{\mathcal{X} \cup \partial\mathcal{X}} \nabla_{x_i} \{ \pi(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{x}_{i,j}) \} \, \mathrm{d}\boldsymbol{x} \\
&= \sum_{i=1}^{d} \sum_{j=1}^{\infty} a_{i,j} \oint_{\partial\mathcal{X}} \underbrace{\pi(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{x}_{i,j})}_{(*)} n_i(\boldsymbol{x}) S(\mathrm{d}\boldsymbol{x}) \quad = \quad 0,
\end{aligned}$$

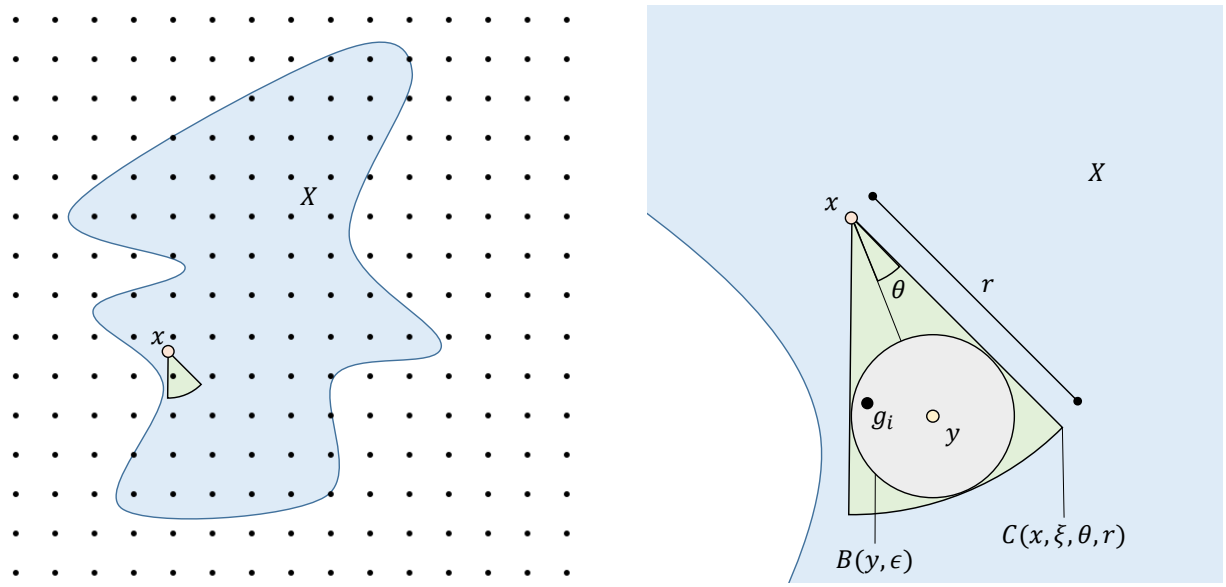
where the order of summation and calculus operations can be interchanged due to uniform convergence of the series representation on the compact set $\mathcal{X} \cup \partial\mathcal{X}$. The term $(*)$ equals zero for all $\boldsymbol{x}_{i,j} \in \mathcal{X} \cup \partial\mathcal{X}$ by (A4). □

Proof of Lemma 2. In rigorously establishing this result there are six main steps. Initially we fix $\boldsymbol{x} \in \mathcal{X} \cup \partial\mathcal{X}$ and aim to show that with high probability there exists a state $\boldsymbol{x}_j$ close to $\boldsymbol{x}$. Then we consider implications for the distribution of the fill distance.

Step #1: "Construct a reference grid." Since $\mathcal{X}$ is bounded, without loss of generality suppose $\mathcal{X} \cup \partial\mathcal{X} \subseteq [0,1]^d$. For $M \in \mathbb{N}$, define a uniform grid of reference points $\{\boldsymbol{g}_i\}_{i=1}^{G} \subset [0,1]^d$ consisting of all $G = M^d$ states of the form $\boldsymbol{g} = (g_1, \ldots, g_d)$ where $g_i \in \{0, \frac{1}{M-1}, \ldots, \frac{M-2}{M-1}, 1\}$ (Fig. 3a). We will require that this grid has a sufficiently fine resolution; specifically we suppose that

$$M \geq \frac{\sqrt{d}(1 + \sin\theta)}{2r \sin\theta} + 1,$$

where r, θ are defined by the interior cone condition that $\mathcal{X} \cup \partial\mathcal{X}$ is assumed to satisfy.



(a) Step #1 (b) Step #2

Figure 3: Schematic of the steps involved in proving Lemma 2.

Step #2: "Find a grid point $\boldsymbol{g}_i$ near to $\boldsymbol{x}$." Define

$$R = \frac{\sqrt{d}}{2}\left(\frac{1}{\sin\theta} + 1\right).$$

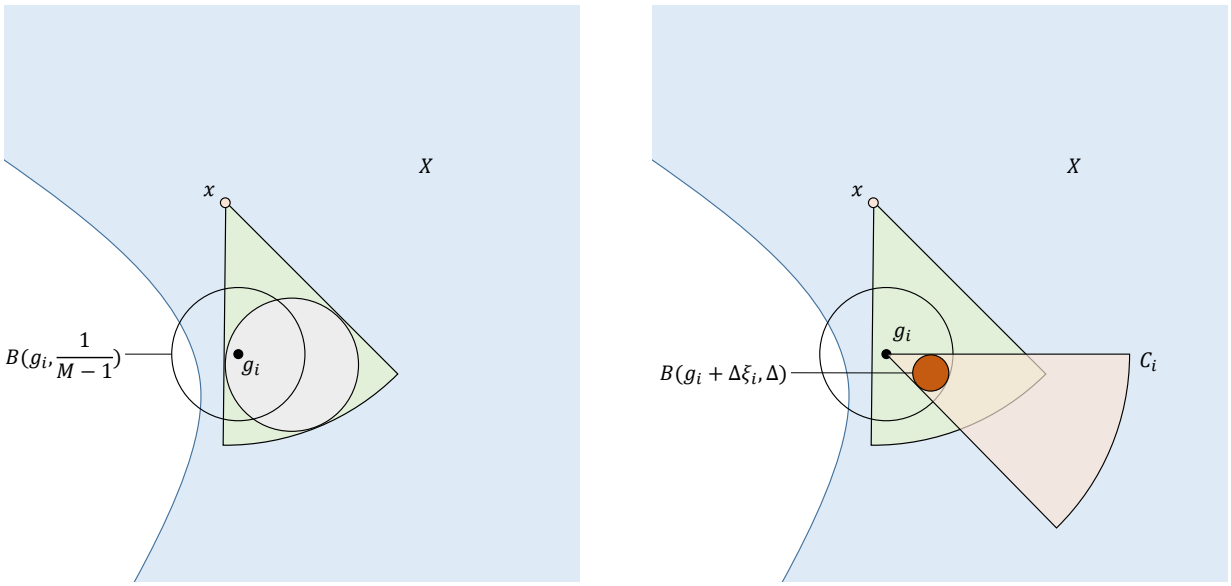
We claim that there exists a grid point $\boldsymbol{g}_i$ such that $\|\boldsymbol{x} - \boldsymbol{g}_i\|_2 < \frac{R}{M-1}$. Indeed, from the interior cone condition there exists a cone $C(\boldsymbol{x}, \boldsymbol{\xi}, \theta, r) \subseteq \mathcal{X}$ (Fig. 3a). Define

$$h = \frac{\sqrt{d}}{2(M-1)\sin\theta}.$$

The fine grid resolution implies $h \leq r/(1+\sin\theta)$. From Wendland (2004), Lemma 3.7, it follows that the cone $C(\boldsymbol{x}, \boldsymbol{\xi}, \theta, r)$ contains the ball $B(\boldsymbol{y}, \epsilon)$ with centre $\boldsymbol{y} = \boldsymbol{x} + h\boldsymbol{\xi}$ and radius $\epsilon = h\sin\theta$ (Fig. 3b). Now, the fill distance for the grid $\{\boldsymbol{g}_i\}_{i=1}^G$ relative to the space $[0,1]^d$ can be shown to equal $\sqrt{d}/(2(M-1))$, which is exactly equal to the radius ϵ. It follows that $B(\boldsymbol{y}, \epsilon)$ must contain a grid point $\boldsymbol{g}_i$ for some $i \in \{1, \ldots, G\}$. From the triangle inequality;

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{g}_i\|_2 &\leq \underbrace{\|\boldsymbol{x} - \boldsymbol{y}\|_2}_{=h} + \underbrace{\|\boldsymbol{y} - \boldsymbol{g}_i\|_2}_{\leq \epsilon} \\ &= \frac{\sqrt{d}}{2(M-1)\sin\theta} + \frac{\sqrt{d}}{2(M-1)} = \frac{R}{M-1}\end{aligned}$$

This establishes the claim.



(a) Step #3 (b) Step #4

Figure 4: Schematic of the steps involved in proving Lemma 2.

Step #3: "A set of points that 'cover the grid' must contain at least one point that is near to $\boldsymbol{x}$." Consider the event

$$E = \left[\forall i \exists j : \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 \leq \frac{1}{M-1}\right].$$

Conditional on E, there exists $\boldsymbol{x}_j$ such that $\|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 \leq \frac{1}{M-1}$. It follows that, conditional on E, we have

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x}_j\|_2 &\leq \|\boldsymbol{x} - \boldsymbol{g}_i\|_2 + \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 \\ &\leq \frac{R}{M-1} + \frac{1}{M-1} = \frac{R+1}{M-1} \end{aligned}$$

where we have used the result of Step #1. Thus the event E, or 'covering the grid', implies the event $[h_{\mathcal{D}_0} \leq \frac{R+1}{M-1}]$ (Fig. 4a).

Step #4: "The grid is covered with high probability." Next, we upper-bound the probability $\mathbb{P}_{\mathcal{D}_0}[E^c]$ of the event E^c. For this, note that for all $i \in \{1, \ldots, G\}$ there exists a unit vector $\boldsymbol{\xi}_i$ such that the cone $C_i = C(\boldsymbol{g}_i, \boldsymbol{\xi}_i, \theta, r)$ is contained in $\mathcal{X}$ (Fig. 4b). It follows that

$$\begin{aligned} \text{vol}\left(B\left(\boldsymbol{g}_i, \frac{1}{M-1}\right) \cap \mathcal{X}\right) &\geq \text{vol}\left(B\left(\boldsymbol{g}_i, \frac{1}{M-1}\right) \cap C_i\right) \\ &= \text{vol}\left(C\left(\boldsymbol{g}_i, \boldsymbol{\xi}_i, \theta, \frac{1}{M-1}\right)\right) \end{aligned}$$

where the final inequality follows since $\frac{1}{M-1} \leq r$ and the intersection of a cone with a ball is another cone. Now, from Wendland (2004), Lemma 3.7, the ball with centre $\boldsymbol{g}_i + \Delta\boldsymbol{\xi}_i$ and

radius

$$\Delta = \frac{\sin\theta}{(M-1)(1+\sin\theta)}$$

is contained in the cone $C(\boldsymbol{g}_i, \boldsymbol{\xi}_i, \theta, \frac{1}{M-1})$. Continuing,

$$\begin{aligned}\mathrm{vol}\left(C\left(\boldsymbol{g}_i, \boldsymbol{\xi}_i, \theta, \frac{1}{M-1}\right)\right) &\geq \mathrm{vol}(B(\boldsymbol{g}_i + \Delta\boldsymbol{\xi}_i, \Delta)) \\ &= V := \frac{(\frac{\sin\theta}{1+\sin\theta})^d \pi^{d/2}}{\Gamma(\frac{d}{2}+1)(M-1)^d}.\end{aligned}$$

Now, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_0}[E^c] &= \mathbb{P}_{\mathcal{D}_0}\left[\exists i : \forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 > \frac{1}{M-1}\right] \\ &\leq \sum_{i=1}^{G} \underbrace{\mathbb{P}_{\mathcal{D}_0}\left[\forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 > \frac{1}{M-1}\right]}_{(*)}.\end{aligned}$$

The probability $(*)$ can be bounded above using independence of the samples. Indeed, the probability that a random draw from Π lands in $B(\boldsymbol{g}_i, \frac{1}{M-1})$ can be lower-bounded by V times $\rho = \inf_{\boldsymbol{x}\in\mathcal{X}\cup\partial\mathcal{X}} \pi(\boldsymbol{x}) > 0$ since, under (A1,5), π can be decomposed as a mixture density with two components, the first uniform over $\mathcal{X} \cup \partial\mathcal{X}$ with mixture weight ρ and the second a component with density $\tilde{\pi}(\boldsymbol{x}) \propto \pi(\boldsymbol{x}) - \rho$ and mixture weight $1-\rho$. Thus we have

$$\mathbb{P}_{\mathcal{D}_0}\left[\forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 > \frac{1}{M-1}\right] \leq (1 - V\rho)^m. \tag{11}$$

Note that this is the only place in the proof that independence of the $\{\boldsymbol{x}_i\}$ is required.

Step #5: "The fill distance is small with high probability." The result of Step #4 can be used to derive a lower bound on the distribution function of the fill distance:

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_0}\left[h_{\mathcal{D}_0} > \frac{R+1}{M-1}\right] \leq \mathbb{P}_{\mathcal{D}_0}[E^c] &\leq G(1-V\rho)^m \\ &= G\left(1 - \frac{(\frac{\sin\theta}{1+\sin\theta})^d \pi^{d/2}\rho}{\Gamma(\frac{d}{2}+1)(M-1)^d}\right)^m\end{aligned}$$

Letting $\zeta = \frac{R+1}{M-1}$ implies that $M = 1 + \frac{R+1}{\zeta}$ and $G = (1 + \frac{R+1}{\zeta})^d$. In this parametrisation, when $M > R+2$ we have $0 < \zeta < 1$ and hence

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_0}[h_{\mathcal{D}_0} > \zeta] &\leq \left(1 + \frac{R+1}{\zeta}\right)^d \left[1 - \frac{(\frac{\sin\theta}{1+\sin\theta})^d \pi^{d/2}\rho}{\Gamma(\frac{d}{2}+1)(R+1)^d}\zeta^d\right]^m \\ &\leq \left(\frac{R+2}{\zeta}\right)^d (1 - C_d\zeta^d)^m,\end{aligned} \tag{12}$$

where we have written

$$C_d = \frac{(\frac{\sin\theta}{1+\sin\theta})^d \pi^{d/2} \rho}{\Gamma(\frac{d}{2}+1)(R+1)^d}.$$

While Eqn. 12 holds only for ζ of the form $\frac{R+1}{M-1}$, it can be made to hold for all $0 < \zeta < 1$ by replacing C_d with $\tilde{C}_d = 2^{-d} C_d$. This is because for any $0 < \zeta < 1$ there exists $M > 1$ such that $\tilde{\zeta} = \frac{R+1}{M-1}$ satisfies $\frac{\zeta}{2} \leq \tilde{\zeta} < \zeta$, along with the fact that $\mathbb{P}_{\mathcal{D}_0}[h_{\mathcal{D}_0} > \zeta] \leq \mathbb{P}_{\mathcal{D}_0}[h_{\mathcal{D}_0} > \tilde{\zeta}]$.

Step #6: "Putting it all together." From $\mathcal{X} \subseteq [0,1]^d$ we have that $h_{\mathcal{D}_0} \in [0,1]$ and hence, since g is continuous and $[0,1]$ is compact, $\sup_{h \in [0,1]} g(h) < \infty$. Without loss of generality, and with probability one, we have $g(h_{\mathcal{D}_0}) \leq 1$. (This is without loss of generality since we can simply redefine g as $g/g(1)$.) From the reverse Markov inequality, for all $\zeta < \mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})]$,

$$\mathbb{P}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0}) > \zeta] \;\; \geq \;\; \frac{\mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})] - \zeta}{1-\zeta}$$

and upon rearranging

$$\mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})] \;\; \leq \;\; \zeta + (1-\zeta)\mathbb{P}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0}) > \zeta]. \tag{13}$$

Since g is continuous and monotone, g^{-1} exists and we have $\mathbb{P}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0}) \leq \zeta] = \mathbb{P}_{\mathcal{D}_0}[h_{\mathcal{D}_0} \leq g^{-1}(\zeta)]$. This allows us to combine Eqns. 12 and 13, obtaining

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})] \quad &\leq \;\; \zeta + (1-\zeta)\left(\frac{R+2}{g^{-1}(\zeta)}\right)^d (1 - \tilde{C}_d (g^{-1}(\zeta))^d)^m \\
&\leq \;\; \zeta + \left(\frac{R+2}{g^{-1}(\zeta)}\right)^d (1 - \tilde{C}_d (g^{-1}(\zeta))^d)^m.
\end{aligned}$$

Now, letting $\zeta = g(m^{-\delta})$ for some fixed δ, subject to $\frac{1}{3d} \leq \delta < \frac{1}{d}$, and varying m, we have that

$$\mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})] \;\; \leq \;\; g(m^{-\delta}) + (R+2)^d \underbrace{m^{d\delta}(1 - \tilde{C}_d m^{-d\delta})^m}_{(**)} \tag{14}$$

where $(**) = O(m^{d\delta} \exp(-\tilde{C}_d m^{1-d\delta}))$. Indeed, since $\log(1-x) \leq -x$ for all $|x| < 1$,

$$\begin{aligned}
\log[m^{d\delta}(1 - \tilde{C}_d m^{-d\delta})^m] \;\; &= \;\; d\delta \log(m) + m \log(1 - \tilde{C}_d m^{-d\delta}) \\
&\leq \;\; d\delta \log(m) - \tilde{C}_d m^{1-d\delta} \;\; = \;\; \log[m^{d\delta} \exp(-\tilde{C}_d m^{1-d\delta})].
\end{aligned}$$

Writing $x = m^{-\delta}$,

$$\begin{aligned}
\frac{m^{d\delta} \exp(-\tilde{C}_d m^{1-d\delta})}{g(m^{-\delta})} \;\; &= \;\; \frac{x^{-d} \exp(-\tilde{C}_d x^{d-1/\delta})}{g(x)} \\
&\leq \;\; \frac{x^{-d} \exp(-\tilde{C}_d x^{-2d})}{g(x)} \quad (\text{take } \delta = \frac{1}{3d}).
\end{aligned}$$

Under the hypothesis on the limiting behavior of $g(x)$ as $x \downarrow 0$, we have that

$$\begin{aligned}\lim_{m\to\infty} \frac{m^{d\delta}\exp(-\tilde{C}_d m^{1-d\delta})}{g(m^{-\delta})} &= \lim_{x\downarrow 0} \frac{x^{-d}\exp(-\tilde{C}_d x^{-2d})}{g(x)} \\ &\leq \lim_{x\downarrow 0} \frac{x^{-d}\exp(-x^{-3d})}{g(x)} = 0\end{aligned}$$

Thus the right hand side of Eqn. 14 is asymptotically minimised by taking δ as large as possible, subject to $(**)$ converging exponentially fast. i.e. $\delta = \frac{1}{d} - \epsilon$ for $\epsilon > 0$ arbitrarily small. We therefore conclude that $\mathbb{E}_{\mathcal{D}_0}[g(h_{\mathcal{D}_0})] = O(g(m^{-1/d+\epsilon}))$, as required. □

Proof of Theorem 1. Unbiasedness follows directly from the structure of splitting estimators (Eqn. 4). From Eqn. 5 it suffices to consider the rate δ at which $\sigma^2(f - s_{f,\mathcal{D}_0})$ vanishes. First, note that

$$\begin{aligned}\sigma^2(f - s_{f,\mathcal{D}_0}) &= \int \left[f(\boldsymbol{x}) - s_{f,\mathcal{D}_0}(\boldsymbol{x}) - \int [f(\boldsymbol{x}') - s_{f,\mathcal{D}_0}(\boldsymbol{x}')]\Pi(\mathrm{d}\boldsymbol{x}')\right]^2 \Pi(\mathrm{d}\boldsymbol{x}) \\ &= \int [f(\boldsymbol{x}) - s_{f,\mathcal{D}_0}(\boldsymbol{x})]^2\Pi(\mathrm{d}\boldsymbol{x}) - \left[\int [f(\boldsymbol{x}') - s_{f,\mathcal{D}_0}(\boldsymbol{x}')]\Pi(\mathrm{d}\boldsymbol{x}')\right]^2 \\ &\leq \int [f(\boldsymbol{x}) - s_{f,\mathcal{D}_0}(\boldsymbol{x})]^2\Pi(\mathrm{d}\boldsymbol{x}).\end{aligned}$$

Second, observe that from (A1-3) the kernel $k_+ \in C_2^{a\wedge b}(\mathcal{X} \cup \partial\mathcal{X})$. Thus from Theorem 11.13 of Wendland (2004) there exists $h > 0$, $C > 0$ such that, whenever $h_{\mathcal{D}_0} < h$, we have

$$|f(\boldsymbol{x}) - s_{f,\mathcal{D}_0}(\boldsymbol{x})| \leq C h_{\mathcal{D}_0}^{a\wedge b} \|f\|_{\mathcal{H}_+}$$

for all $\boldsymbol{x} \in \mathcal{X}$. An unconditional bound, after squaring and integrating according to Π, is

$$1_{h_{\mathcal{D}_0}<h} \int [f(\boldsymbol{x}) - s_{f,\mathcal{D}_0}(\boldsymbol{x})]^2\Pi(\mathrm{d}\boldsymbol{x}) \leq C^2 h_{\mathcal{D}_0}^{2(a\wedge b)} \|f\|_{\mathcal{H}_+}^2.$$

Third, combining the first two parts and taking an expectation over the m samples in $\mathcal{D}_0$ gives

$$\mathbb{E}_{\mathcal{D}_0}[1_{h_{\mathcal{D}_0}<h}\sigma^2(f - s_{f,\mathcal{D}_0})] \leq C^2\|f\|_{\mathcal{H}_+}^2 \mathbb{E}_{\mathcal{D}_0}[h_{\mathcal{D}_0}^{2(a\wedge b)}].$$

Finally, from Lemma 2 with $g(x) = x^{2(a\wedge b)}$ we have that $\mathbb{E}_{\mathcal{D}_0}[h_{\mathcal{D}_0}^{2(a\wedge b)}] = O(n^{-2\frac{a\wedge b}{d}+\epsilon})$ for $\epsilon > 0$ arbitrarily small. The result follows from Eqn. 5. □

Proof of Lemma 3. In the proof of Lemma 2, independence of the samples $\{\boldsymbol{x}_i\}$ is only used to establish Eqn. 11. Below we derive an almost equivalent inequality that is valid for non-independent samples arising from the Markov chain sample path.

From Roberts and Rosenthal (1998a, Prop. 1) a uniformly ergodic, reversible Markov chain is strongly uniformly ergodic; i.e. there exists $N \in \mathbb{N}$ and $0 < \upsilon < 1$ such that for all

$\boldsymbol{x} \in \mathcal{X}$ the N-step transition density $P^N(\boldsymbol{x}, \cdot)$ satisfies the minorisation condition $P^N(\boldsymbol{x}, \cdot) \geq \upsilon\Pi[\cdot]$. Manipulating this minorisation condition gives $P^N(\boldsymbol{x}, A^c) = 1 - P^N(\boldsymbol{x}, A) \leq 1 - \upsilon\Pi(A)$ for $A \in \mathcal{A}$.

Fix $A = B(\boldsymbol{g}_i, \frac{1}{M-1})$. Denote the initial distribution of the Markov chain by Π_1 and, for notational simplicity only, assume Π_1 has a density $\pi_1 = \mathrm{d}\Pi_1/\,\mathrm{d}\Lambda$. Define the hitting time $\tau_A := \min\{n : \boldsymbol{x}_n \in A \text{ given } \boldsymbol{x}_1 \sim \Pi_1\}$. Then for $n \geq N$ we have the following bound:

$$
\begin{aligned}
\mathbb{P}[\tau_A > n] &= \int_{\boldsymbol{x}_1 \in A^c} \cdots \int_{\boldsymbol{x}_n \in A^c} \pi_1(\boldsymbol{x}_1) P(\boldsymbol{x}_1, \boldsymbol{x}_2) \ldots P(\boldsymbol{x}_{n-1}, \boldsymbol{x}_n)\, \mathrm{d}\boldsymbol{x}_n \ldots \mathrm{d}\boldsymbol{x}_1 \\
&= \int_{\boldsymbol{x}_1 \in A^c} \cdots \int_{\boldsymbol{x}_{n-N} \in A^c} \pi_1(\boldsymbol{x}_1) P(\boldsymbol{x}_1, \boldsymbol{x}_2) \ldots P(\boldsymbol{x}_{n-N-1}, \boldsymbol{x}_{n-N}) \\
&\times \underbrace{\int_{\boldsymbol{x}_{n-N+1} \in A^c} \cdots \int_{\boldsymbol{x}_n \in A^c} P(\boldsymbol{x}_{n-N}, \boldsymbol{x}_{n-N+1}) \ldots P(\boldsymbol{x}_{n-1}, \boldsymbol{x}_n)\, \mathrm{d}\boldsymbol{x}_{n-N+1} \ldots \mathrm{d}\boldsymbol{x}_n}_{\leq P^N(\boldsymbol{x}_{n-N}, A^c)} \mathrm{d}\boldsymbol{x}_{n-N} \ldots \mathrm{d}\boldsymbol{x}_1 \\
&\leq [1 - \upsilon\Pi(A)] \\
&\quad \times \int_{\boldsymbol{x}_1 \in A^c} \cdots \int_{\boldsymbol{x}_{n-N} \in A^c} \pi_1(\boldsymbol{x}_1) P(\boldsymbol{x}_1, \boldsymbol{x}_2) \ldots P(\boldsymbol{x}_{n-N-1}, \boldsymbol{x}_{n-N})\, \mathrm{d}\boldsymbol{x}_{n-N} \ldots \mathrm{d}\boldsymbol{x}_1 \\
\ldots &\leq [1 - \upsilon\Pi(A)]^{\lfloor n/N \rfloor} \\
&\times \int_{\boldsymbol{x}_1 \in A^c} \cdots \int_{\boldsymbol{x}_{n-\lfloor n/N \rfloor N} \in A^c} \pi_1(\boldsymbol{x}_1) P(\boldsymbol{x}_1, \boldsymbol{x}_2) \ldots P(\boldsymbol{x}_{n-\lfloor n/N \rfloor N-1}, \boldsymbol{x}_{n-\lfloor n/N \rfloor N})\, \mathrm{d}\boldsymbol{x}_{n-\lfloor n/N \rfloor N} \ldots \mathrm{d}\boldsymbol{x}_1 \\
&\leq [1 - \upsilon\Pi(A)]^{\lfloor n/N \rfloor}
\end{aligned}
$$

Employing this bound, we have

$$
\mathbb{P}_{\mathcal{D}_0}\left[\forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 > \frac{1}{M-1}\right] \;=\; \mathbb{P}_{\mathcal{D}_0}[\tau_A > m] \;\leq\; [1 - \upsilon\Pi(A)]^{\lfloor m/N \rfloor}.
$$

As before we have $\Pi(A) \geq V\rho$ and hence

$$
\mathbb{P}_{\mathcal{D}_0}\left[\forall j, \|\boldsymbol{g}_i - \boldsymbol{x}_j\|_2 > \frac{1}{M-1}\right] \;\leq\; (1 - \upsilon V\rho)^{\lfloor m/N \rfloor}.
$$

Eqn. 15 is essentially identical to Eqn. 11 up to the inclusion of a factor $0 < \upsilon < 1$ and a factor $1/N$. Arguing as in Lemma 2 with $\tilde{C}_d$ replaced by $\upsilon N^{-1} \times \tilde{C}_d$ completes the proof. □

Proof of Theorem 2. We appeal to Theorem 1 of Roberts and Rosenthal (2008). This states that reversible, geometrically ergodic Markov chains $(\boldsymbol{x}_i)_{i=1}^n$ are *variance bounding*, meaning that there exists $K < \infty$ such that

$$
\lim_{n \to \infty} \frac{1}{n} \mathbb{V}\left[\sum_{i=1}^{n} h(\boldsymbol{x}_i)\right] \leq K\sigma^2(h) \tag{15}
$$

holds for all $h \in L^2(\mathcal{X}, \Pi)$. Suppose that the chain starts at stationarity, so that Eqn. 15 is equivalent to the statement

$$\lim_{n\to\infty} n\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i) - \int h \,\mathrm{d}\Pi\right)^2\right] \leq K\sigma^2(h). \tag{16}$$

(An arbitrary initial condition $\boldsymbol{x}_1 \sim \Pi_1$ can be handled using standard renewal theory; we do not present the details here.) Applying Eqn. 16 to the control functional estimator produces

$$\lim_{n\to\infty} (n-m)\mathbb{E}_{\mathcal{D}_1}\left[\left(m_{f,\mathcal{D}_0,\mathcal{D}_1} - \int f \,\mathrm{d}\Pi\right)^2\right] \leq K\sigma^2(f - s_{f,\mathcal{D}_0}).$$

Arguing as in the proof of Theorem 1, we have

$$\mathbb{E}_{\mathcal{D}_0}[1_{h_{\mathcal{D}_0}<h}\sigma^2(f - s_{f,\mathcal{D}_0})] \leq C^2\|f\|^2_{\mathcal{H}_+}\mathbb{E}_{\mathcal{D}_0}[h_{\mathcal{D}_0}^{2(a\wedge b)}].$$

Finally, it remains only to show that the scaling relation $\mathbb{E}_{\mathcal{D}_0}[h_{\mathcal{D}_0}^{2(a\wedge b)}] = O(n^{-2(a\wedge b)/d+\epsilon})$, where $\epsilon > 0$ can be arbitrarily small, holds in the non-independent setting. This was precisely the content of Lemma 3. □

Denote $\boldsymbol{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ for the vector $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{1}k(\boldsymbol{x}, \boldsymbol{x}')$. The shorthand notation $\mathbb{S}_q[\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')]$ will be used to refer to $\mathbb{S}_q[\boldsymbol{k}(\cdot, \boldsymbol{x}')](\boldsymbol{x})$; i.e. the action of the Stein operator on the first argument of a bivariate function.

Lemma 5. *Assume (A3). For each $q \in \mathcal{Q}(k) \cap \mathcal{R}(k)$ it holds that*

$$\int \mathbb{S}_q[\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')]Q(\mathrm{d}\boldsymbol{x}) \;=\; 0, \tag{17}$$

where the left hand side exists for all $\boldsymbol{x}' \in \mathcal{X}$.

Proof. This is a generalisation of Lemma 1. From (A3) and $q \in \mathcal{Q}(k)$, the function $\mathbb{S}_q[\boldsymbol{k}(\cdot, \boldsymbol{x}')]$ exists on $\mathcal{X}$ for each $\boldsymbol{x}' \in \mathcal{X}$. In particular $\mathbb{S}_q[\boldsymbol{k}(\cdot, \boldsymbol{x}')] \in L^2(\mathcal{X} \cup \partial\mathcal{X}, Q)$, so that the left hand side of Eqn. 17 exists. Then

$$\begin{aligned}
\int \mathbb{S}_q[\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')]Q(\mathrm{d}\boldsymbol{x}) &= \int [\nabla_{\boldsymbol{x}} \cdot \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') \cdot \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})]\, Q(\mathrm{d}\boldsymbol{x}) \\
&= \int [q(\boldsymbol{x})\nabla_{\boldsymbol{x}} \cdot \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') \cdot \nabla_{\boldsymbol{x}} q(\boldsymbol{x})]\, \mathrm{d}\boldsymbol{x} \\
&= \int \nabla_{\boldsymbol{x}} \cdot \{q(\boldsymbol{x})\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')\}\, \mathrm{d}\boldsymbol{x} \\
&\overset{(*)}{=} \oint_{\partial\mathcal{X}} \underbrace{q(\boldsymbol{x})\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}')}_{(**)} \cdot \boldsymbol{n}(\boldsymbol{x})\, S(\mathrm{d}\boldsymbol{x}) \;=\; 0,
\end{aligned}$$

where $(*)$ is integration by parts and $(**)$ equals zero for all $\boldsymbol{x}' \in \partial\mathcal{X}$ since $q \in \mathcal{R}(k)$. □

Proof of Lemma 4. Recall that, for compact $\mathcal{X} \cup \partial\mathcal{X}$, the notion of c-universality is equivalent to *cc-universality*, where the (weaker) topology of compact convergence is used in place of the $\|\cdot\|_\infty$ norm topology (Carmeli *et al.*, 2010, Defn. 4.1).

For densities q, q' on $(\mathcal{X} \cup \partial\mathcal{X}, \mathcal{B})$ with $q = \mathrm{d}Q/\,\mathrm{d}\Lambda$, $q' = \mathrm{d}Q'/\,\mathrm{d}\Lambda$, define

$$T(q, q', k) \;:=\; \left\| \int \mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q(\mathrm{d}\boldsymbol{x}) - \int \mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q'(\mathrm{d}\boldsymbol{x}) \right\|_{\mathcal{H}},$$

whenever the right hand side exists. Observe, as in Lemma 5, that $T(q, q', k)$ is well-defined for all $q, q' \in \mathcal{Q}(k)$.

This proof then proceeds in two steps:

Step #1: First, we follow the proof of Chwialkowski *et al.* (2016, Thm. 2.1) in order to establish that

$$q, q' \in \mathcal{Q}(k) \cap \mathcal{R}(k), \quad T(q, q', k) = 0 \quad \Longrightarrow \quad q = q'.$$

Indeed:

$$\begin{aligned}
T(q, q', k) &= \left\| \int \mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q(\mathrm{d}\boldsymbol{x}) - \underbrace{\int \mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q'(\mathrm{d}\boldsymbol{x})}_{=0 \text{ from Lemma 5}} \right\|_{\mathcal{H}} \\
&= \left\| \int \left[\sum_{i=1}^{d} \nabla_{x_i} k(\boldsymbol{x}, \cdot) + k(\boldsymbol{x}, \cdot) \nabla_{x_i} \log q'(\boldsymbol{x}) \right] Q(\mathrm{d}\boldsymbol{x}) \right\|_{\mathcal{H}} \\
&= \Bigg\| \underbrace{\int \left[\sum_{i=1}^{d} \nabla_{x_i} k(\boldsymbol{x}, \cdot) + k(\boldsymbol{x}, \cdot) \nabla_{x_i} \log q(\boldsymbol{x}) \right] Q(\mathrm{d}\boldsymbol{x})}_{(*)} \\
&\qquad + \underbrace{\int \left[\sum_{i=1}^{d} k(\boldsymbol{x}, \cdot) \nabla_{x_i} \log q'(\boldsymbol{x}) - k(\boldsymbol{x}, \cdot) \nabla_{x_i} \log q(\boldsymbol{x}) \right] Q(\mathrm{d}\boldsymbol{x})}_{(**)} \Bigg\|_{\mathcal{H}}
\end{aligned}$$

where

$$(*) = \int \mathbb{S}_q[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q(\mathrm{d}\boldsymbol{x}),$$

which is the zero function from Lemma 5 and $(**)$ is the mean embedding (Smola *et al.*, 2007) of the function

$$g(\boldsymbol{x}) := \sum_{i=1}^{d} \nabla_{x_i} \log \frac{q'(\boldsymbol{x})}{q(\boldsymbol{x})}.$$

Note that $g \in L^2(\mathcal{X} \cup \partial\mathcal{X}, Q)$ follows from $q, q' \in \mathcal{Q}(k)$. By assumption this embedding satisfies

$$\left\| \int \left[\sum_{i=1}^{d} \nabla_{x_i} \log \frac{q'(\boldsymbol{x})}{q(\boldsymbol{x})} \right] k(\boldsymbol{x}, \cdot) Q(\mathrm{d}\boldsymbol{x}) \right\|_{\mathcal{H}} = 0.$$

Since $\mathcal{H}$ is *cc*-universal and $g \in L^2(\mathcal{X} \cup \partial\mathcal{X}, \Pi)$, the embedding is zero if and only if $g = 0$ (Carmeli *et al.*, 2010, Thm. 4.4c). Thus

$$\sum_{i=1}^{d} \nabla_{x_i} \log \frac{q'(\boldsymbol{x})}{q(\boldsymbol{x})} = 0.$$

This implies that q' and q are proportional on $\mathcal{X}$ and, since q, q' are both densities in $C^1(\mathcal{X} \cup \partial\mathcal{X})$, we must conclude $q = q'$.

Step #2: It is sufficient to prove that $\mathcal{H}_+$ is dense in $(C^1(\mathcal{X} \cup \partial\mathcal{X}), \|\cdot\|_2)$, since this set is dense in $(L^2(\mathcal{X} \cup \partial\mathcal{X}, \Pi), \|\cdot\|_2)$. Now, suppose $\mathcal{H}_+$ is not dense in $(C^1(\mathcal{X} \cup \partial\mathcal{X}), \|\cdot\|_2)$. Then there exists $0 \neq f \in C^1(\mathcal{X} \cup \partial\mathcal{X})$ such that

$$\int f \, \mathrm{d}\Pi = 0, \qquad \int f\psi \, \mathrm{d}\Pi = 0 \quad \forall\, \psi \in \mathcal{H}_0,$$

the second requirement representing orthogonality of f with respect to $\mathcal{H}_0$. Compactness of $\mathcal{X} \cup \partial\mathcal{X}$ implies $f \in L^\infty(\mathcal{X} \cup \partial\mathcal{X})$. Let

$$Q := \frac{c+f}{c}\,\Pi, \quad Q' := \Pi, \quad c := 1 + \|f\|_\infty,$$

so that Q, Q' are both distributions with $Q \neq Q'$. Moreover, under (A2',4), both Q, Q' admit densities q, q' such that $q, q' \in \mathcal{Q}(k) \cap \mathcal{R}(k)$. Indeed, $q \in \mathcal{Q}(k)$ since

(a) $q \in C^1(\mathcal{X} \cup \partial\mathcal{X})$

(b) $q > 0$ on $\mathcal{X}$;

(c) $\nabla_{x_i} \log q = \frac{\nabla_{x_i} f}{c+f} + \nabla_{x_i} \log \pi \in L^2(\mathcal{X} \cup \partial\mathcal{X}, Q')$ for all distributions Q' on $(\mathcal{X} \cup \partial\mathcal{X}, \mathcal{B})$, since $\nabla_{x_i} f \in C^0(\mathcal{X} \cup \partial\mathcal{X})$ and $c + f \geq 1$ on $\mathcal{X}$;

and $q \in \mathcal{R}(k)$ since $q(\boldsymbol{x})k(\boldsymbol{x}, \boldsymbol{x}') = \frac{c+f(\boldsymbol{x})}{c}\pi(\boldsymbol{x})k(\boldsymbol{x}, \boldsymbol{x}')$, where $\frac{c+f(\boldsymbol{x})}{c}$ is bounded on $\mathcal{X}$ and $\pi(\boldsymbol{x})k(\boldsymbol{x}, \boldsymbol{x}') = 0$ for $\boldsymbol{x} \in \partial\mathcal{X}$ and $\boldsymbol{x}' \in \mathcal{X} \cup \partial\mathcal{X}$.

The purpose of this construction becomes clear from plugging this choice of q, q' into the operator T:

$$\begin{aligned} T(q, q', k) &= \left\| \int \mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q(\mathrm{d}\boldsymbol{x}) - \int \mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)] Q'(\mathrm{d}\boldsymbol{x}) \right\|_{\mathcal{H}} \\ &= \left\| \frac{1}{c} \int \underbrace{\mathbb{S}_{q'}[\boldsymbol{k}(\boldsymbol{x}, \cdot)]}_{(***)} f(\boldsymbol{x}) \Pi(\mathrm{d}\boldsymbol{x}) \right\|_{\mathcal{H}}. \end{aligned}$$

The function $(***)$ belongs to $\mathcal{H}_0$; the definition of f then implies that this integral is zero (orthogonality with respect to $\mathcal{H}_0$) and so we conclude that $T(q, q', k) = 0$. From Step #1 we then conclude that $q = q'$. This is a contradiction and so $\mathcal{H}_+$ has been shown to be dense in $(C^1(\mathcal{X} \cup \partial\mathcal{X}), \|\cdot\|_2)$. □

References

Aronszajn, N. (1950) Theory of reproducing kernels. *T. Am. Math. Soc.*, **68**(3), 337-404.

Carmeli, C., De Vito, E., Toigo, A. and Umanita, V. (2010) Vector valued reproducing kernel Hilbert spaces and universality. *Anal. Appl.*, **8**, 19-61.

Chwialkowski, K., Strathmann, H. and Gretton, A. (2016) A Kernel Test of Goodness of Fit. Proceedings of the 33rd International Conference on Machine Learning.

Roberts, G. O. and Rosenthal, J. S. (1998a) Two convergence properties of hybrid samplers. *Ann. Appl. Prob.*, **8**(2), 397-407.

Roberts, G. O. and Rosenthal, J. S. (2008) Variance bounding Markov chains. *Ann. Appl. Prob.*, **18**(3), 1201-1214.

Smola, A., Gretton, A., Song, L. and Schölkopf, B. (2007) A Hilbert space embedding for distributions. In *Proc. 18th International Conference on Algorithmic Learning Theory*, 13-31. Springer-Verlag, Berlin.

Wendland, H. (2004) *Scattered Data Approximation*. Cambridge University Press.