# Control functionals for Monte Carlo integration

Chris Oates, Mark Girolami, Nicolas Chopin

# Control Functionals for Monte Carlo Integration

Chris J. Oates[1,2,*], Mark Girolami[3,4] and Nicolas Chopin[5]
[1]University of Technology Sydney, Australia
[2]Australian Research Council Centre for Excellence in Mathematical and Statistical Frontiers
[3]University of Warwick, Coventry, UK
[4]Alan Turing Institute
[5]CREST-LS and ENSAE, Paris, France

April 5, 2016

**Abstract**

A non-parametric extension of control variates is presented. These leverage gradient information on the sampling density to achieve substantial variance reduction. It is not required that the sampling density be normalised. The novel contribution of this work is based on two important insights; (i) a trade-off between random sampling and deterministic approximation and (ii) a new gradient-based function space derived from Stein's identity. Unlike classical control variates, our estimators achieve super-root-$n$ convergence, often requiring orders of magnitude fewer simulations to achieve a fixed level of precision. Theoretical and empirical results are presented, the latter focusing on integration problems arising in hierarchical models and models based on non-linear ordinary differential equations.

*Keywords:* control variates, non-parametric, reproducing kernel, Stein's identity, variance reduction

## 1 Introduction

Statistical methods are increasingly being employed to analyse complex models of physical phenomena (e.g. in climate forecasting or simulations of molecular dynamics; Slingo *et al.*, 2009; Angelikopoulos *et al.*, 2012). Analytic intractability of complex models has inspired the development of sophisticated Monte Carlo methodologies to facilitate computation (Robert

*Address for correspondence: School of Mathematical and Physical Sciences, University of Technology Sydney, NSW 2007, Australia. E-mail: christopher.oates@uts.edu.au

and Casella, 2004). In their most basic form, Monte Carlo estimators converge as the reciprocal of root-$n$ where $n$ is the number of random samples. For complex models it may only be feasible to obtain a limited number of samples (e.g. a recent Met Office model for future climate simulations required the order of $10^6$ core-hours per simulation; Mizielinski *et al.*, 2014). In these situations, root-$n$ convergence is too slow and leads in practice to high-variance estimation. Our contribution is motivated by resolving this issue and provides novel methodology that is both formal and general.

The focus of this paper is the estimation of an expectation $\mu(f) = \int f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, where $f$ is a test function of interest and $\pi$ is a probability density associated with a random variable $\boldsymbol{X}$. Provided that $f(\boldsymbol{X})$ has variance $\sigma^2(f) < \infty$, the arithmetic mean estimator

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i),$$

based on $n$ independent and identically distributed (IID) samples $\{\boldsymbol{x}_i\}_{i=1}^n$ of the random variable, satisfies the central limit theorem and converges to $\mu(f)$ at the rate $O_P(n^{-1/2})$, or simply at "root-$n$". When working with complex models, root-$n$ convergence can be problematic, as highlighted in e.g. Ba and Joseph (2012). A model is considered complex when either (i) $\boldsymbol{X}$ is expensive to simulate, or (ii) $f$ is expensive to evaluate, in each case relative to the required estimator precision. Both situations are prevalent in scientific and engineering applications (e.g. Kohlhoff *et al.*, 2014; Higdon *et al.*, 2015). This paper introduces a class of estimators that converge more quickly than root-$n$. The significance of our contribution is made clear in the comparative overview below.

Generic approaches to reduction of variance are well-known in both statistics and numerical analysis. These include (i) importance sampling and its extensions (Cornuet *et al.*, 2012; Li *et al.*, 2013), (ii) stratified sampling and related techniques (Rubinstein and Kroese, 2011), (iii) antithetic variables (Green and Han, 1992) and more generally (randomised) quasi-Monte Carlo (QMC/RQMC; Dick and Pillichshammer, 2010), (iv) Rao-Blackwellisation (Robert and Casella, 2004; Douc and Robert, 2011; Ghosh and Clyde, 2011; Olsson and Ryden, 2011), (v) Riemann sums (Philippe, 1997), (vi) control variates (Glasserman, 2004; Mira *et al.*, 2013; Li *et al.*, 2016), (vii) multi-level Monte Carlo and related techniques (e.g. Heinrich, 1995; Giles, 2013; Giles and Szpruch, 2014), (viii) Bayesian Monte Carlo (BMC; O'Hagan, 1991; Rasmussen and Ghahramani, 2003; Briol *et al.*, 2015), and (ix) a plethora of sophisticated Markov chain Monte Carlo sampling schemes (MCMC; Łatuszyński *et al.*, 2015). Classical introductions to many of the above techniques include Robert and Casella (2004, Chap. 4) and Rubinstein and Kroese (2011, Chap. 5).

Motivated by contemporary statistical applications, we state four *desiderata* for a variance reduction technique: (I) *Unbiased estimation:* Monte Carlo (MC) methods based on IID samples produce unbiased estimators, whilst techniques such as MCMC generally produce biased estimators. (II) *Compatibility with an un-normalised density $\pi$:* An "un-normalised" density is known only up to proportionality so that, for example, MCMC techniques are required for sampling. (III) *Super-root-n convergence* (for sufficiently regular $f$): The convergence rates of (R)QMC are well studied and can be super-root-$n$. Riemann sums can

| Estimation Method | Unbiased | Un-normalised $\pi$ | Super-root-$n$ | Post-hoc |
|---|---|---|---|---|
| MC(/MCMC) + Arithmetic Mean | ✓(/×) | ×(/✓) | × | × |
| MC + Importance Sampling | ✓(/×) | ×(/✓) | × | × |
| MC + Antithetic Variables | ✓ | × | × | × |
| MC(/MCMC) + Stratified Sampling | ✓(/×) | ×(/✓) | × | × |
| Quasi-MC (QMC) | × | × | ✓ | × |
| Randomised QMC (RQMC) | ✓ | × | ✓ | × |
| MC(/MCMC) + Rao-Blackwellisation | ✓(/×) | ×(/✓) | × | ✓ |
| MC(/MCMC) + Control Variates | ✓(/×) | ×(/✓) | × | ✓ |
| MC(/MCMC) + Riemann Sums | × | ×(/✓) | ✓ | ✓ |
| Bayesian MC (BMC) | × | × | ✓ | ✓ |
| MC(/MCMC) + Control Functionals | ✓(/×) | ×(/✓) | ✓ | ✓ |

Table 1: A comparison of estimation methods for integrals. ["Unbiased" = the estimator is unbiased for $\mu(f)$. "Un-normalised $\pi$" = the estimator can handle sampling densities that are only available up to proportionality. "Super-root-$n$" = the estimator converges faster than root-$n$. "Post-hoc" = the estimator places no restriction on how the samples $\boldsymbol{x}_i$ are generated, i.e. requires no modification to computer code for sampling. Estimator properties may change in order to handle un-normalised densities $\pi$; these are shown in parentheses.]

also achieve super-root-$n$ rates and Briol *et al.* (2016) showed the same holds for BMC. (IV) *Post-hoc schemes:* Rao-Blackwellisation, Riemann sums, BMC and control variates can all be conceived as *post-hoc* schemes; i.e. schemes that can be applied retrospectively after samples have been obtained. In contrast, the remaining methods require modification to computer code for the sampling process itself. The former are appealing from both a theoretical and a practical perspective since they separate the challenge of sampling from the challenge of variance reduction.

Table 1 summarises existing techniques in relation to these *desiderata*; note that no technique fulfils all four criteria. In contrast, the method proposed here, called "control functionals", is able to satisfy all four *desiderata*. Control functionals appear to be similar, in this sense, to Riemann sums i.e. they are a super-root-$n$, *post-hoc* approach that applies to un-normalised sampling densities. However, Riemann sums are rarely used in practice due to (i) the fact that estimators are biased at finite sample sizes, and (ii) there is a prohibitive increase in methodological complexity for multi-dimensional state spaces. Control functionals do not posses either of these drawbacks.

The control functional method that we develop below can be intuitively considered as a non-parametric development of control variates. In control variate schemes one seeks a basis $\{s_i\}_{i=1}^m$, $m \in \mathbb{N}$, that have expectation $\mu(s_i) = 0$. Then a surrogate function $\tilde{f} = f - a_1 s_1 - \cdots - a_m s_m$ is constructed such that $\mu(\tilde{f}) = \mu(f)$ and, for suitably chosen $a_1, \ldots, a_m \in \mathbb{R}$, a variance reduction $\sigma^2(\tilde{f}) < \sigma^2(f)$ is obtained (see e.g. Rubinstein and Marcus, 1985). The statistics $s_i$ are known as control variates and the variance $\sigma^2(\tilde{f})$ can be reduced to zero if and only if there is perfect canonical correlation between $f$ and the basis $\{s_i\}_{i=1}^m$. For estimation based on Markov chains, control variates for the discrete state space case were

provided by Andradóttir *et al.* (1993). For continuous state spaces, statistics relating to the chain can be used as control variates (Hammer and Tjelmeland, 2008; Dellaportas and Kontoyiannis, 2012; Li *et al.*, 2016). Alternatively control variates can be constructed based on gradient information (Assaraf and Caffarel, 1999; Mira *et al.*, 2013).

The control variates described above are solving a misspecified regression problem, since in general $f$ will not be a linear combination of the $s_i$ basis functions. As such they achieve at most a constant factor reduction in estimator variance. Intuitively, one would like to increase the number $m$ of basis functions to increase in line with the number $n$. Mijatović and Vogrinc (2015) explored this approach within the Metropolis-Hastings method. However, their solution requires the user to partition of the state space, which limits its wider appeal. This paper introduces a powerful new perspective on variance reduction that fully resolves these issues, satisfying all the *desiderata* described above. To realise our method we developed a gradient-based function space that leads to closed-form estimators whose convergence can be guaranteed. The functional analysis perspective works "out of the box", without requiring the user to partition the state space. Extensive empirical support is provided in favour of the proposed method, including applications to hierarchical models and models based on non-linear differential equations. In each case state-of-the-art estimation is achieved.

All results can be reproduced using `MATLAB R2015a` code that is available to download from `http://warwick.ac.uk/control_functionals`.

# 2 Methodology

## 2.1 Set-up and notation

Consider a random vector $\boldsymbol{X}$ taking values in an open set $\Omega \subseteq \mathbb{R}^d$. Assume $\boldsymbol{X}$ admits a positive density on $\Omega$ with respect to $d$-dimensional Lebesgue measure, written $\pi(\boldsymbol{x}) > 0$. For bounded $\Omega$ with boundary $\partial\Omega$, we assume $\partial\Omega$ is piecewise smooth (i.e. infinitely differentiable). Write $\mathcal{L}^2(\pi)$ for the space of measurable functions $g : \Omega \to \mathbb{R}$ for which $\int_\Omega g(\boldsymbol{x})^2\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is finite. Write $C^k(\Omega, \mathbb{R}^j)$ for the space of (measurable) functions from $\Omega$ to $\mathbb{R}^j$ with continuous partial derivatives up to order $k$. Consider a test function $f : \Omega \to \mathbb{R}$ of interest, assume $f \in \mathcal{L}^2(\pi)$ and write $\mu(f) := \int_\Omega f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, $\sigma^2(f) := \int_\Omega (f(\boldsymbol{x}) - \mu(f))^2\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Denote by $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ a collection of states $\boldsymbol{x}_i \in \Omega$. At each state $\boldsymbol{x}_i$ the corresponding function values $f(\boldsymbol{x}_i)$ and gradients $\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x}_i)$ are assumed to have been pre-computed and cached. The method that we develop does not then require any further recourse to the statistical model $\pi$, nor any further evaluations of the function $f$, and is in this sense a widely-applicable *post-hoc* scheme.

## 2.2 From control variates to control functionals

### 2.2.1 Deterministic approximation

Our starting point is establish a trade-off between random sampling and deterministic approximation, as suggested on several separate occasions by authors including Bakhvalov (1959); Heinrich (1995); Speight (2009); Giles (2013).

Consider a dichotomy of available states $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ 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 surrogate functions of the form

$$f_{\mathcal{D}_0}(\boldsymbol{x}) := f(\boldsymbol{x}) - s_{f,\mathcal{D}_0}(\boldsymbol{x}) + \mu(s_{f,\mathcal{D}_0}),$$

where $s_{f,\mathcal{D}_0} \in \mathcal{L}^2(\pi)$ is an approximation to $f$, based on $\mathcal{D}_0$, whose expectation $\mu(s_{f,\mathcal{D}_0})$ is analytically tractable. By construction $f_{\mathcal{D}_0} \in \mathcal{L}^2(\pi)$, $\mu(f_{\mathcal{D}_0}) = \mu(f)$ and $\sigma^2(f_{\mathcal{D}_0}) = \sigma^2(f - s_{f,\mathcal{D}_0})$. We study estimators of the form

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) := \begin{cases} \frac{1}{n-m} \sum_{i=m+1}^n f_{\mathcal{D}_0}(\boldsymbol{x}_i) & \text{for } m < n \\ \mu(s_{f,\mathcal{D}_0}) & \text{for } m = n. \end{cases}$$

For theoretical purposes the second subset $\mathcal{D}_1$ is assumed to be an IID sample from $\pi$, statistically independent from $\mathcal{D}_0$. Then, for $m < n$, we have unbiasedness, i.e. $\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)] = \mu(f)$, where the expectation here is with respect to the sampling distribution $\pi$ of the $n - m$ random variables that constitute $\mathcal{D}_1$, and is conditional on $\mathcal{D}_0$. The corresponding estimator variance, conditional on $\mathcal{D}_0$, is $\mathbb{V}_{\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)] = (n - m)^{-1}\sigma^2(f - s_{f,\mathcal{D}_0})$. This formulation encompasses control variates as the special case where $s_{f,\mathcal{D}_0}$ is constrained to a finite-dimensional space.

The insight required to go beyond control variates and achieve super-root-$n$ convergence is that we can use an infinite-dimensional space to construct an increasingly accurate approximations $s_{f,\mathcal{D}_0}$ as $m \to \infty$. We allow for the possibility that the first subset $\mathcal{D}_0$ are also random and write $\mathbb{E}_{\mathcal{D}_0}$ to denote an expectation with respect to the (marginal) distribution of these $m$ random variables.

**Proposition 1.** *Assume $m = O(n^\gamma)$ for some $\gamma \in [0, 1]$ and that the expected functional approximation error (EFAE) vanishes as*

$$\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(m^{-\delta})$$

*for some $\delta \geq 0$. Then $\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f))^2] = O(n^{-1-\gamma\delta})$.*

All proofs are reserved for Appendix A.

**Remark 1.** *Taking $\gamma = 1$ optimises the rate in Prop. 1 and we therefore assume in the sequel that $m/n \to r \in (0, 1)$.*

### 2.2.2 Control variates based on Stein's identity

To construct approximations $s_{f,\mathcal{D}_0}$ whose integrals $\mu(s_{f,\mathcal{D}_0})$ are analytically tractable, we make the assumption

(A1) The density $\pi$ belongs to $C^1(\Omega, \mathbb{R})$.

Denote the gradient function by $\boldsymbol{u}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x})$ where $\nabla_{\boldsymbol{x}} := [\partial/\partial x_1, \ldots, \partial/\partial x_d]^T$, well-defined by (A1). We study approximations of the form

$$\begin{aligned} s_{f,\mathcal{D}_0}(\boldsymbol{x}) &:= c + \psi(\boldsymbol{x}) \\ \psi(\boldsymbol{x}) &:= \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}) \end{aligned} \tag{1}$$

where $c \in \mathbb{R}$ is a constant and $\boldsymbol{\phi} \in C^1(\Omega, \mathbb{R}^d)$. Eqn. 1 appears in Stein's classical test for approximate normality (Stein, 1970) and related to (but simpler than) control variates proposed by Assaraf and Caffarel (1999); Mira *et al.* (2013). We make the following assumption (c.f. e.g. Eqn. 9 of Mira *et al.*, 2013):

(A2) Let $\boldsymbol{n}(\boldsymbol{x})$ be the unit normal to the boundary $\partial\Omega$ of the state space $\Omega$. Then

$$\oint_{\partial\Omega} \pi(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) S(\mathrm{d}\boldsymbol{x}) = 0.$$

(The notation $\oint_{\partial\Omega}$ denotes a surface integral over $\partial\Omega$ and $S(\mathrm{d}\boldsymbol{x})$ denotes the surface element at $\boldsymbol{x} \in \partial\Omega$.) Stein's identity implies that this class of approximations has integrals that are analytically tractable:

**Proposition 2.** *Assume (A1,2). Then* $\mu(\psi) = 0$ *and so* $\mu(s_{f,\mathcal{D}_0}) = c$.

When $\Omega$ is unbounded, all surface integrals are interpreted as tail conditions. i.e. (A2) should be replaced with $\oint_{\Gamma_r \cap \Omega} \pi(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) S(\mathrm{d}\boldsymbol{x}) \to 0$, where $\Gamma_r \subset \mathbb{R}^d$ is the sphere of radius $r$ centred at the origin and $\boldsymbol{n}(\boldsymbol{x})$ is the unit normal to the surface of $\Gamma_r$.

The statistic $\psi$ is recognised as a control variate. These control variates were explored in the case where $\boldsymbol{\phi}$ is a (gradient of a low-degree) polynomial by Assaraf and Caffarel (1999), Assaraf and Caffarel (2003) and Mira *et al.* (2013). This paper takes the innovative step of setting $\boldsymbol{\phi}$ within a function space to enable fully non-parametric approximation. The functional approximation perspective differs fundamentally from the control variate approach, in which the estimation problem is formally mis-specified (i.e. $\boldsymbol{\phi}$ is restricted to a low dimensional parametric family that does not contain the "true" function). We emphasise this key conceptual distinction by referring to $\psi$ as a control *functional* (CF; reflecting the use of terminology from functional analysis).

## 2.3 Theory

This section establishes $\psi$ as belonging to a Hilbert space $\mathcal{H}_0 \subset \mathcal{L}^2(\pi)$. This allows us to formulate and solve a functional approximation problem that targets the EFAE.

### 2.3.1 A Hilbert space of control functionals

Specification of $\psi$ is equivalent to specification of $\boldsymbol{\phi}$. We decide to restrict each component function $\phi_i : \Omega \to \mathbb{R}$ to a Hilbert space $\mathcal{H} \subset \mathcal{L}^2(\pi) \cap C^1(\Omega, \mathbb{R})$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Moreover we insist that $\mathcal{H}$ is a reproducing kernel Hilbert space. This implies that there exists a symmetric positive definite function $k : \Omega \times \Omega \to \mathbb{R}$ such that (i) for all $\boldsymbol{x} \in \Omega$ we have $k(\cdot, \boldsymbol{x}) \in \mathcal{H}$ and (ii) for all $\boldsymbol{x} \in \Omega$ and $h \in \mathcal{H}$ we have $h(\boldsymbol{x}) = \langle h, k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}}$ (Berlinet and Thomas-Agnan, 2004, Def. 1, p7, and Def. 2, p10). The vector-valued function $\boldsymbol{\phi} : \Omega \to \mathbb{R}^d$ 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 $\langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle \phi_i, \phi_i' \rangle_{\mathcal{H}}$.

We make an assumption on $k$ that will be enforced by construction:

(A3) The kernel $k$ belongs to $C^2(\Omega \times \Omega, \mathbb{R})$.

Now we can analyse the class of CFs induced by $k$:

**Theorem 1.** *Assume $\boldsymbol{\phi} \in \mathcal{H}^d$ and (A1,3). Then $\psi$ belongs to $\mathcal{H}_0$, the reproducing kernel Hilbert space with kernel $k_0(\boldsymbol{x}, \boldsymbol{x}') := \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{u}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{u}(\boldsymbol{x}') \cdot \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}') k(\boldsymbol{x}, \boldsymbol{x}')$.*

To gain some intuition for $\mathcal{H}_0$ we strengthen (A2) as follows:

(A2') For $\pi$-almost all $\boldsymbol{x} \in \Omega$ the kernel $k$ satisfies

$$\oint_{\partial\Omega} k(\boldsymbol{x}, \boldsymbol{x}') \pi(\boldsymbol{x}') \boldsymbol{n}(\boldsymbol{x}') S(\mathrm{d}\boldsymbol{x}') = \boldsymbol{0}$$

and

$$\oint_{\partial\Omega} \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}') \pi(\boldsymbol{x}') \cdot \boldsymbol{n}(\boldsymbol{x}') S(\mathrm{d}\boldsymbol{x}') = 0.$$

While (A2') must be verified on a case-by-case basis, it can in principle always be enforced with a suitable choice of $k$.

**Lemma 1.** *Under (A1,2',3), the gradient-based kernel $k_0$ satisfies*

$$\int_{\Omega} k_0(\boldsymbol{x}, \boldsymbol{x}') \pi(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}' = 0$$

*for $\pi$-almost all $\boldsymbol{x} \in \Omega$.*

Lemma 1 generalises Eqn. 1 of Mira *et al.* (2013) and implies that $\mathcal{H}_0$ consists of only valid CFs, i.e. $\psi \in \mathcal{H}_0 \implies \mu(\psi) = 0$. These ideas are illustrated in Fig. 1.

(A4) The gradient-based kernel $k_0$ satisfies

$$\int_{\Omega} k_0(\boldsymbol{x}, \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} < \infty.$$

**Lemma 2.** *Under (A1,2',3,4) we have $\mathcal{H}_0 \subset \mathcal{L}^2(\pi)$.*

**Remark 2.** *In general (A4) must be verified on a case-by-case basis. (A4) is easily verified for all examples in this paper.*

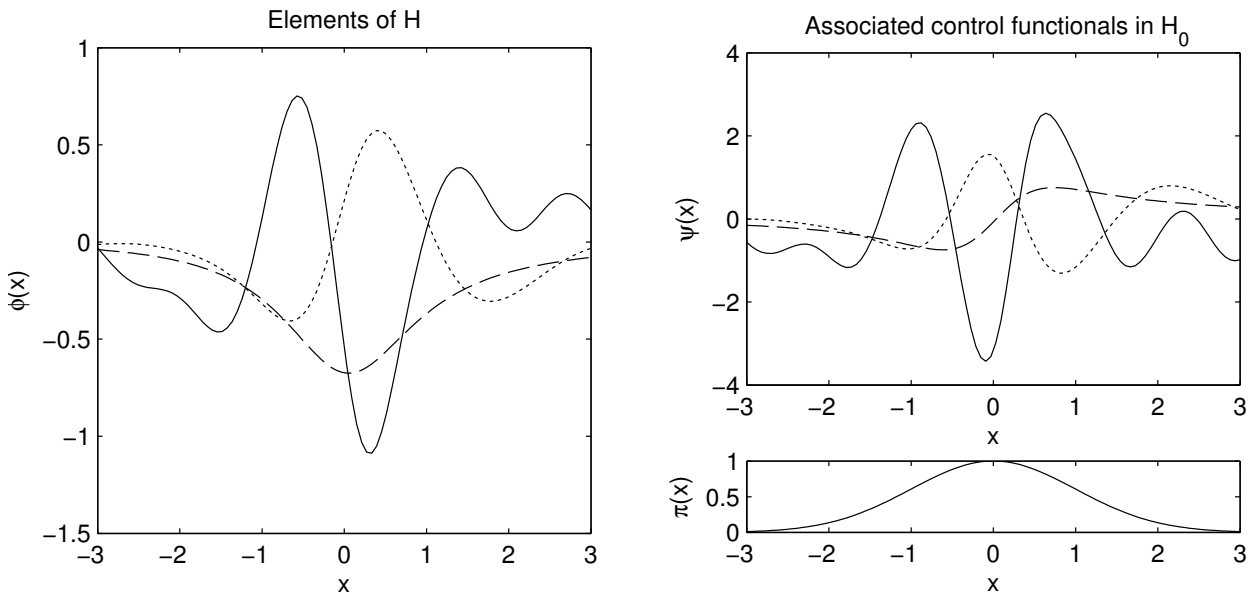


Figure 1: Constructing control functionals (in dimension $d = 1$): Representative elements $\phi$ from the reproducing kernel Hilbert space $\mathcal{H}$ (left panel) are plotted, along with their associated control functionals $\psi = \nabla\phi + \phi\nabla \log \pi$ in $\mathcal{H}_0$ (right, top panel). Each $\phi$ is unconstrained in expectation, but the corresponding control functional $\psi$ is automatically constrained to have expectation zero with respect to the (possibly un-normalised) probability density $\pi$ (right, bottom panel).

### 2.3.2 Consistent approximation and asymptotics

Now we establish theoretical results for consistent approximation of $f$ by $s_{f,\mathcal{D}_0}$. Write $\mathcal{C}$ for the reproducing kernel Hilbert space of constant functions with kernel $k_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{x}') = 1$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$. Denote the norms associated to $\mathcal{C}$ and $\mathcal{H}_0$ respectively by $\|\cdot\|_{\mathcal{C}}$ and $\|\cdot\|_{\mathcal{H}_0}$. Write $\mathcal{H}_+ = \mathcal{C} + \mathcal{H}_0$ for the set $\{c + \psi : c \in \mathcal{C},\ \psi \in \mathcal{H}_0\}$. Equip $\mathcal{H}_+$ with the structure of a vector space, with addition operator $(c + \psi) + (c' + \psi') = (c + c') + (\psi + \psi')$ and multiplication operator $\lambda(c + \psi) = (\lambda c) + (\lambda\psi)$, each well-defined due to uniqueness of the representation $f = c + \psi$, $f' = c' + \psi'$ with $c, c' \in \mathcal{C}$ and $\psi, \psi' \in \mathcal{H}_0$. In addition, equip $\mathcal{H}_+$ with the norm $\|f\|^2_{\mathcal{H}_+} := \|c\|^2_{\mathcal{C}} + \|\psi\|^2_{\mathcal{H}_0}$, again, well-defined by uniqueness of representation. It can be shown that $\mathcal{H}_+$ is a reproducing kernel Hilbert space with kernel $k_+(\boldsymbol{x}, \boldsymbol{x}') := k_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{x}') + k_0(\boldsymbol{x}, \boldsymbol{x}')$ (Berlinet and Thomas-Agnan, 2004, Thm. 5, p24).

For the analysis we assume a basic well-posedness condition:

(A5) $f \in \mathcal{H}_+$. i.e. $f = c + \psi$ for some $c \in \mathcal{C}$ and $\psi \in \mathcal{H}_0$.

**Remark 3.** *(A5) is equivalent to the existence of a solution $\boldsymbol{\phi} \in \mathcal{H}^d$ to the partial differential equation*

$$\nabla_{\boldsymbol{x}} \cdot [\pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})] = [f(\boldsymbol{x}) - \mu(f)]\pi(\boldsymbol{x}),$$

*called the "fundamental equation" in Assaraf and Caffarel (1999, Eqn. 5; see also Eqn. 4 in Mira et al. (2013)). With no initial or boundary conditions to satisfy, it is easy to show that there exist infinitely many solutions to the fundamental equation, so (A5) is automatically satisfied by choosing $k$ such that $\mathcal{H}$ is big enough to contain at least one solution.*

To realise the CF method we consider the regularised least-squares (RLS) functional approximation given by

$$s_{f,\mathcal{D}_0} := \arg\min_{g \in \mathcal{H}_+} \left\{ \frac{1}{m} \sum_{j=1}^{m} (f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j))^2 + \lambda \|g\|^2_{\mathcal{H}_+} \right\}$$

where $\lambda > 0$. For the special case where $m = n$, the CF estimator can be interpreted as kernel quadrature (Sommariva and Vianello, 2006) and also as empirical interpolation (Kristoffersen, 2013). The distinguishing feature of CFs from these methods is that the Stein construction is compatible with un-normalised $\pi$.

Below we will establish that the RLS estimate produces vanishing EFAE under a strengthening of (A4):

(A4') $\sup_{\boldsymbol{x} \in \Omega} k_0(\boldsymbol{x}, \boldsymbol{x}) < \infty$.

**Remark 4.** *(A4') would follow from (A3) and compactness of the state space $\Omega$, but we do not assume compactness here. All experiments in this paper have at worst $\boldsymbol{u}(\boldsymbol{x}) = O(\|\boldsymbol{x}\|_2)$, so that (A4') is automatically satisfied, for example, when we choose $k(\boldsymbol{x}, \boldsymbol{x}') = (1+\alpha_1\|\boldsymbol{x}\|_2^2 + \alpha_1\|\boldsymbol{x}'\|_2^2)^{-1} \exp(-(2\alpha_2^2)^{-1}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2)$ for some $\alpha_1, \alpha_2 > 0$.*

We can now state our main result:

**Theorem 2.** *Assume (A1,2',3,4',5) and take a RLS estimate with $\lambda = O(m^{-1/2})$. When $\mathcal{D}_0$ are IID samples from $\pi$, the estimator $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is an unbiased estimator of $\mu(f)$ with $\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f))^2] = O(n^{-7/6})$.*

CFs based on RLS therefore improve upon the Monte Carlo rate. The hypotheses on $\pi$ are weak, only requiring that $\pi$ be continuously differentiable. Empirical evidence (below) indicates stronger rates hold in more regular examples. Indeed we can prove sharper results under stronger conditions that include boundedness of $\Omega$. Details are reserved for a future publication (Oates *et al.*, 2016).

Importantly, the RLS estimate leads to a convenient closed-form expression for the CF estimator:

**Lemma 3.** *Assume (A1,3). The CF estimator based on RLS is*

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \underbrace{\frac{1}{n-m} \mathbf{1}^T (\boldsymbol{f}_1 - \hat{\boldsymbol{f}}_1)}_{(*)} + \underbrace{\frac{\mathbf{1}^T (\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1} \boldsymbol{f}_0}{1 + \mathbf{1}^T (\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1} \mathbf{1}}}_{(**)} \tag{2}$$

*where* $\boldsymbol{f}_0 = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]^T$, $\boldsymbol{f}_1 = [f(\boldsymbol{x}_{m+1}), \ldots, f(\boldsymbol{x}_n)]^T$, $\mathbf{1} = [1, \ldots, 1]^T$, $(\boldsymbol{K}_0)_{i,j} = k_0(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *and the vector*

$$\hat{\boldsymbol{f}}_1 := \boldsymbol{K}_{1,0}(\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1}\boldsymbol{f}_0 + (\mathbf{1} - \boldsymbol{K}_{1,0}(\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1}\mathbf{1}) \left( \frac{\mathbf{1}^T(\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1}\boldsymbol{f}_0}{1 + \mathbf{1}^T(\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1}\mathbf{1}} \right)$$

*contains predictions for* $\boldsymbol{f}_1$ *based only on* $\mathcal{D}_0$*, with* $(\boldsymbol{K}_{1,0})_{i,j} = k_0(\boldsymbol{x}_{m+i}, \boldsymbol{x}_j)$*.*

**Remark 5.** *The estimator is a weighted combination of function values* $\boldsymbol{f} = [\boldsymbol{f}_0^T, \boldsymbol{f}_1^T]^T$ *with weights summing to one. Estimates are readily obtained using standard matrix algebra. Moreover the weights are independent of the test function* $f$ *and can be re-used to estimate multiple expectations* $\mu(f_j)$ *for a collection* $\{f_j\}$*.*

**Remark 6.** *The samples* $\mathcal{D}_1$ *enter only through the term* $(*)$ *in Eqn. 2, which vanishes in probability as* $m \to \infty$*. Thus any randomness due to* $\mathcal{D}_1$ *vanishes and this gives another perspective on the source of super-root-n convergence of the estimator.*

**Remark 7.** *The term* $(**)$ *in Eqn. 2 is algebraically equivalent to BMC based on* $\mathcal{H}_+$*. i.e.* $(**)$ *is the posterior mean for* $\mu(f)$ *based on a Gaussian process (GP) prior* $f \sim \mathcal{GP}(0, k_+)$ *and data* $\mathcal{D}_0$ *(Rasmussen and Ghahramani, 2003, Eqn. 9). Our general construction in Lemma 3 therefore "heals" BMC in the sense of (i) de-biasing the BMC estimator, (ii) generalising BMC to un-normalised densities and (iii) remaining agnostic to statistical paradigm (e.g. frequentist vs. Bayesian).*

### 2.3.3 Non-asymptotic bounds

The naive computational complexity associated with the RLS estimate is $O(m^3)$ due to the solution of an $m \times m$ linear system. In situations where $\boldsymbol{X}$ is expensive to simulate or $f$ is expensive to evaluate, $m$ is necessarily small and this additional computational cost will be negligible relative to model-based computation. In such scenarios we are more interested in non-asymptotic behaviour:

**Theorem 3.** *Assume (A1,2',3,5). Let* $\lambda \searrow 0$ *to simplify presentation. Then*

$$|\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f)| \leq D(\mathcal{D}_0, \mathcal{D}_1)^{1/2} \|f\|_{\mathcal{H}_+}$$

*where*

$$D(\mathcal{D}_0, \mathcal{D}_1) = \frac{1}{(n-m)^2} \left[ \frac{(\mathbf{1}^T \boldsymbol{K}_{1,0} \boldsymbol{K}_0^{-1} \mathbf{1})^2}{1 + \mathbf{1}^T \boldsymbol{K}_0^{-1} \mathbf{1}} - \mathbf{1}^T \boldsymbol{K}_{1,0} \boldsymbol{K}_0^{-1} \boldsymbol{K}_{0,1} \mathbf{1} + \mathbf{1}^T \boldsymbol{K}_1 \mathbf{1} \right].$$

*Here* $(\boldsymbol{K}_1)_{i,j} = k_0(\boldsymbol{x}_{m+i}, \boldsymbol{x}_{m+j})$ *and* $\boldsymbol{K}_{0,1} = \boldsymbol{K}_{1,0}^T$*.*

Theorem 3 provides an explicit error bound for $f \in \mathcal{H}_+$, mimicking the approach of (R)QMC (Dick and Pillichshammer, 2010, Sec. 2.3.3). This offers a principled approach to

selection of the design points $\mathcal{D}_0$ since, writing $\mathbb{V}_{\mathcal{D}_0,\mathcal{D}_1}$ for the variance with respect to the joint distribution of $\mathcal{D}_0$ and $\mathcal{D}_1$, we have

$$\mathbb{V}_{\mathcal{D}_0,\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0,\mathcal{D}_1;f)] \leq \mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}[D(\mathcal{D}_0,\mathcal{D}_1)]\|f\|^2_{\mathcal{H}_+}.$$

In the extreme case where $\boldsymbol{x} \neq \boldsymbol{x}' \implies k_0(\boldsymbol{x},\boldsymbol{x}') = 0$, the discrepancy $D(\mathcal{D}_0,\mathcal{D}_1)$ reduces to $(n-m)^{-1}$ and we recover the usual root-$n$ rate. A similar bound forms the basis for recent work on two-sample testing by Chwialkowski *et al.* (2016); Liu *et al.* (2016).

### 2.3.4 Implementation

Several randomly chosen splits of the samples $\mathcal{D}$ into subsets $\mathcal{D}_0$ and $\mathcal{D}_1$ may be averaged over to reduce estimator variance. We note that a multi-splitting estimator remains unbiased. As an alternative to multi-splitting, for applications where consistency suffices and unbiased estimation is not essential, we also propose the simplified estimator $(**)$ in Eqn. 2 with $m = n$. Empirical results below show that bias is negligible for practical purposes and, to pre-empt our conclusions, we recommend this simplified estimator for use in applications due to its reduced variance compared to the multi-splitting estimator. In all cases the regularisation parameter $\lambda$ was taken to be the smallest power of 10 such that the kernel matrix $\boldsymbol{K}_0 + \lambda\boldsymbol{I}$ has condition number lower than $10^{10}$.

A kernel $k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\alpha})$ typically involves hyper-parameters $\boldsymbol{\alpha}$ that must be specified. Selection of $\boldsymbol{\alpha}$ can proceed via cross-validation, under the assumption that $\mathcal{D}_0$ are independent samples from $\pi$. Specifically, we randomly split the samples $\mathcal{D}_0$ into $m'$ training samples $\mathcal{D}_{0,0}$ and $m - m'$ test samples $\mathcal{D}_{0,1}$. Then we propose to select $\boldsymbol{\alpha}$ to minimise $\|\boldsymbol{f}_{(0,1)} - \hat{\boldsymbol{f}}_{(0,1)}\|_2$ where $\boldsymbol{f}_{(0,1)}$ is a vector of values $f_i$ for $\boldsymbol{x}_i \in \mathcal{D}_{0,1}$, and $\hat{\boldsymbol{f}}_{(0,1)}$ are the corresponding predicted values. In this way we are targeting the EFAE that reflects the variance of the CF estimator. We emphasise that the cross-validated estimator does not require additional sampling and will remain unbiased provided that preliminary cross-validation is performed only using $\mathcal{D}_0$. In this paper we employed the kernel defined in Remark 4, with hyper-parameters $\alpha_1, \alpha_2$. Full pseudocode is provided in the supplement.

## 2.4 Illustration

To illustrate the method we begin with simple, tractable examples. Consider the synthetic problem of estimating the expectation of $f(\boldsymbol{X}) = \sin(\frac{\pi}{d}\sum_{i=1}^{d} X_i)$ where $\boldsymbol{X}$ is a $d$-dimensional standard Gaussian random variable. By symmetry the true expectation is $\mu(f) = 0$. Initially we take $n = 50$ IID samples and consider the scalar case $d = 1$. Cross-validation was used to select tuning parameters. Specifically: (i) We selected the hyper-parameters $\alpha_1 = 0.1$, $\alpha_2 = 1$ on the basis that this approximately minimised the cross-validation error (Fig. S1a). (ii) We found that estimator variance due to sample-splitting was minimised when at least half of the samples were allocated to $\mathcal{D}_0$ (Fig. S1b). We therefore set a conservative default $m = \lceil n/2 \rceil$. (iii) Empirical results showed that little additional variance reduction occurs from employing multiple splits (Fig. S1c), so we chose to just use a single split. (iv) Finally,

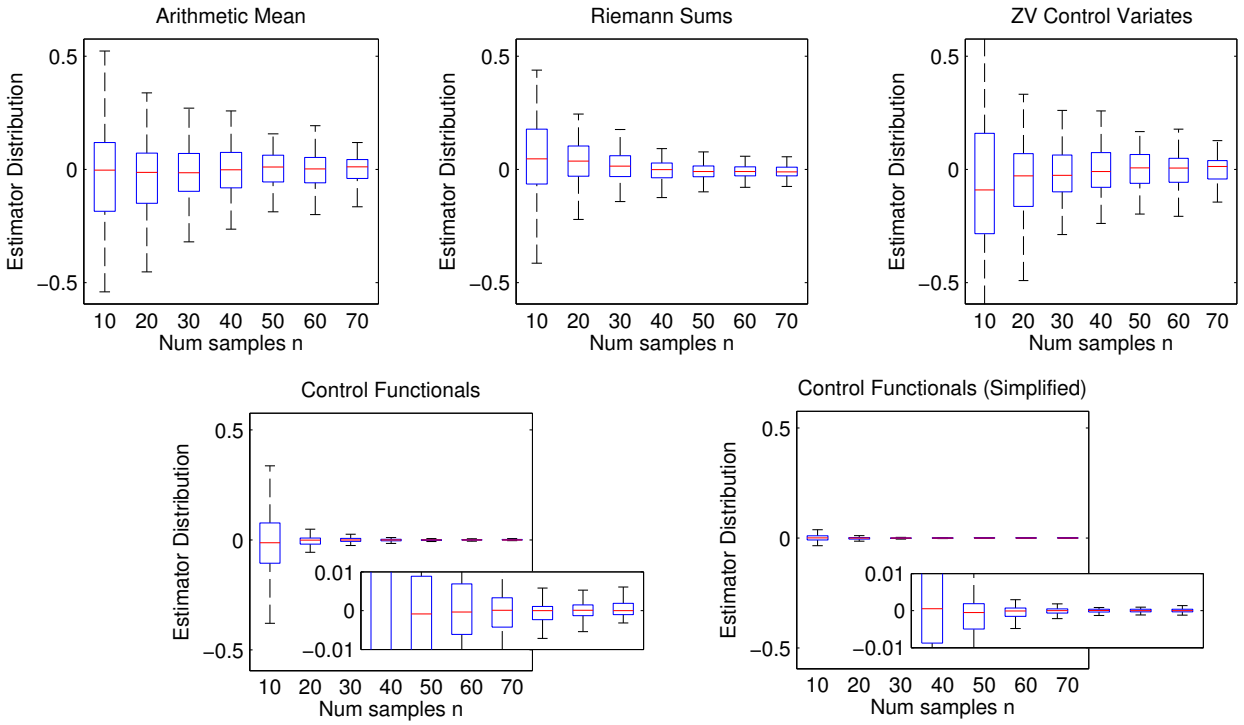


Figure 2: Illustration on a synthetic integration problem in $d = 1$ dimension. Here we display the empirical sampling distribution of Monte Carlo estimators, based on $n$ samples and 100 independent realisations. [The settings for all methods were as described in the main text.]

we found that the bias of the simplified estimator was negligible ($<\sim 10^{-3}$) compared to Monte Carlo error ($\sim 10^{-2}$) (Fig. S1d). This is in line with an analogous result for classical control variates, where estimator bias vanishes asymptotically with respect to Monte Carlo error (Glasserman, 2004, p.200).

In Fig. 2 we summarise the sampling distribution of both the sample-splitting and simplified CF estimators as the number of samples $n$ is varied. The alternative approaches of the arithmetic mean, Riemann sums and "zero variance" (ZV) control variates are also shown, the latter being based on quadratic polynomials (Mira *et al.*, 2013). It is visually apparent that CFs enjoy the lowest variance at all samples sizes considered. We note that in this synthetic example, where there are essentially no computational restrictions, the CF framework is unnecessary and gains in precision come with comparable increases in computational cost. However we emphasise that, in the serious applications that follow, the CF calculations requires negligible computational resources in comparison to simulation from the model. Super-root-$n$ convergence could perhaps be achieved by employing polynomials of increasing degree in the ZV method, but our implementation of this approach did not provide stable estimates in this example (full details in the supplement).

Since the performance of CF is so pronounced, in order to more clearly visualise the results for all sample sizes, in Fig. 3 we plot the estimator mean square error (MSE) scaled

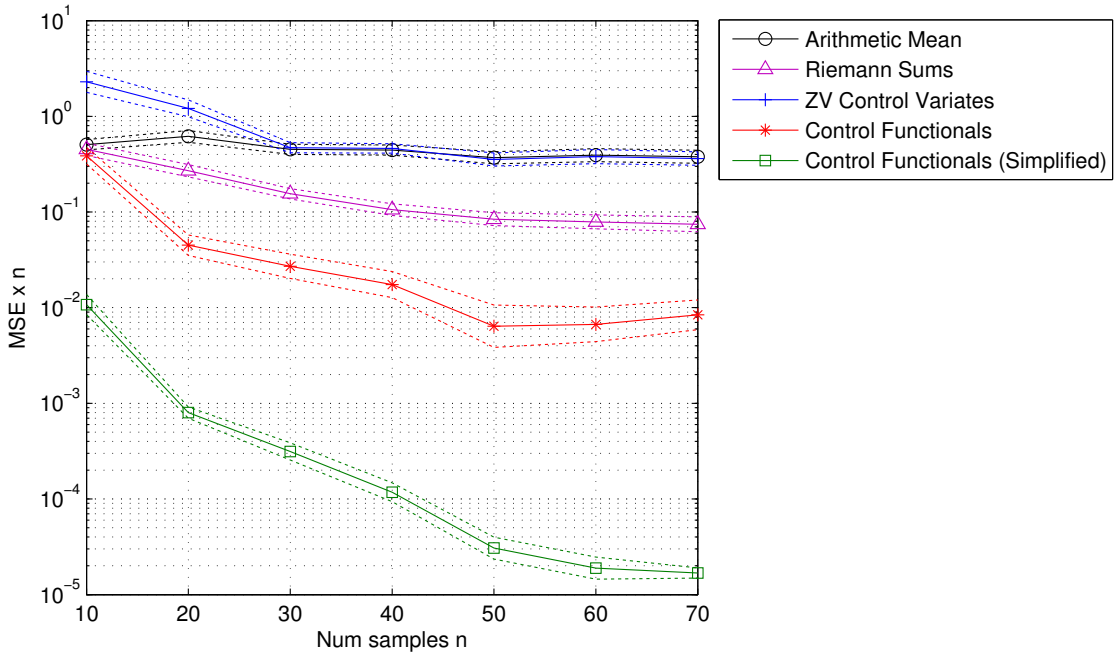


Figure 3: Illustration on a synthetic integration problem in $d = 1$ dimension (continued). Empirical assessment of asymptotic properties.

by $n$, so that root-$n$ convergence corresponds to a horizontal line. Empirical results here are consistent with theory, showing that the arithmetic mean and control variates all achieve a constant factor variance reduction, whereas Riemann sums and CFs achieve super-root-$n$ convergence. In this example CFs significantly outperformed Riemann sums, the latter being based on piecewise linear approximations. We plot results for both the sample-splitting CF estimator and the simplified CF estimator, observing that the latter has lower variance.

To assess the generality of our conclusions we considered going beyond the scalar case to examples with dimensions $d = 3$ and $d = 5$. The analogous results in Figs. S2a, S2b show that, whilst increasing dimensionality presents fundamental challenges for all the variance reduction methods, CF continues to out-perform alternatives. Going further we considered a variety of alternative problems, varying both the test function $f$ and the density $\pi$. These include several pathological cases, with results summarised in Table S1. The results marked (b) echo the conclusions of Mira *et al.* (2013), that ZV control variates are effective in many cases where $f$ is well-approximated by a low-degree polynomial and $\pi$ is a Gaussian or gamma density. However, when $f$ is not well-approximated by a low-degree polynomial, or when $\pi$ takes a more complex form, as in cases marked (c), ZV control variates can be outperformed by CFs, which have the potential to decrease variance dramatically. We then investigated how CFs can fail when theoretical assumptions are violated (see examples marked "CF $\times$"). As expected, violation of (A2) and (A5) in (e), (g) respectively led to poor performance of the CF estimator. Interestingly, violation of differentiability in example (f) did not lead to poor estimation, though this may be because $\pi$ was only non-differentiable at a single point.

We have not reported computational times for these experiments. Our work is motivated by settings in which either simulation from $\pi$ or evaluation of $f$ (or both) are computationally prohibitive, so that additional effort required to implement CFs is negligible by comparison; we illustrate this with two realistic applications the next section.

# 3 Applications

Two applications are considered that together present many of the challenges associated with complex models. Firstly we consider marginalisation of hyper-parameters in hierarchical models, focussing on a non-parametric prediction problem. Here evaluation of $f$ forms a computational bottleneck due to the required inversion of a large matrix. For this problem, CFs are shown to offer significant computational savings. Secondly we consider computation of normalising constants for models based on non-linear ordinary differential equations (ODEs). Evaluation of the likelihood function requires numerical integration of a system of ODEs and dominates computational expenditure in both sampling from $\pi$ and evaluation of $f$. Here CFs combine with gradient-based population MCMC and thermodynamic integration in order to deliver a state-of-the-art technique for low-variance estimation of normalising constants.

## 3.1 Marginalisation in hierarchical models

A fully Bayesian treatment of hierarchical models aims to marginalise over hyper-parameters, but this often entails a prohibitive level of computation. Here we explore the efficacy of CFs in such situations.

### 3.1.1 A hierarchical GP model

The marginalisation of hyper-parameters is a common problem in spatial statistics and Bayesian statistics in general (Besag and Green, 1993; Agapiou *et al.*, 2014; Filippone and Girolami, 2014). Here we consider one such model that is based on $p$-dimensional GP regression. Denote by $Y_i \in \mathbb{R}$ a measured response variable at state $\boldsymbol{z}_i \in \mathbb{R}^p$, assumed to satisfy $Y_i = g(\boldsymbol{z}_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ are independent for $i = 1, \ldots, N$ and $\sigma > 0$ will be assumed known. In order to use training data $(y_i, \boldsymbol{z}_i)_{i=1}^n$ to make predictions regarding an unseen test point $\boldsymbol{z}_*$, we place a GP prior $g \sim \mathcal{GP}(0, c(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\theta}))$ where $c(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\theta}) = \theta_1 \exp(-\frac{1}{2\theta_2^2}\|\boldsymbol{z} - \boldsymbol{z}'\|_2^2)$. Here $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are hyper-parameters that control how training samples are used to predict the response at a new test point. In the fully-Bayesian framework these are assigned hyper-priors, say $\theta_1 \sim \Gamma(\alpha, \beta)$, $\theta_2 \sim \Gamma(\gamma, \delta)$ in the shape/scale parametrisation, which we write jointly as $\pi(\boldsymbol{\theta})$.

### 3.1.2 Marginalising the GP hyper-parameters

We are interested in predicting the value of the response $Y_*$ corresponding to an unseen state vector $\boldsymbol{z}_*$. Our estimator will be the Bayesian posterior mean given by

$$\hat{Y}_* := \mathbb{E}[Y_*|\boldsymbol{y}] = \int \mathbb{E}[Y_*|\boldsymbol{y}, \boldsymbol{\theta}]\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{3}$$

where we implicitly condition on the covariates $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, \boldsymbol{z}_*$. Eqn. 3 is unavailable in closed form and we therefore naive a Monte Carlo estimate by sampling $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ independently from the prior $\pi(\boldsymbol{\theta})$ (more efficient QMC estimates are considered later). Phrasing in terms of our previous notation, the function of interest is

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

where $(\boldsymbol{C}_N)_{i,j} = c(\boldsymbol{z}_i, \boldsymbol{z}_j; \boldsymbol{\theta})$ and $(\boldsymbol{C}_{*,N})_{1,j} = c(\boldsymbol{z}_*, \boldsymbol{z}_j; \boldsymbol{\theta})$ and the underlying distribution is $\pi(\boldsymbol{\theta})$. Each evaluation of the integrand $f(\boldsymbol{\theta})$ requires $O(N^3)$ operations due to the matrix inversion; this can be reduced by employing a "subset of regressors" approximation

$$f(\boldsymbol{\theta}) \approx \boldsymbol{C}_{*,N'}(\boldsymbol{C}_{N',N}\boldsymbol{C}_{N,N'} + \sigma^2 \boldsymbol{C}_{N'})^{-1}\boldsymbol{C}_{N',N}\boldsymbol{y} \tag{4}$$

where $N' < N$ denotes a subset of the full data (see Sec. 8.3.1 of Rasmussen and Williams, 2006, for full details). To facilitate the illustration below, which investigates the sampling distribution of estimators, we take a random subset of $N = 1,000$ training points and a subset of regressors approximation with $N' = 100$. However we emphasise that evaluation of Eqn. 4 will typically be based on much larger $N$ and $N'$ and will be extremely expensive in general. In applications we would therefore have to proceed with Monte Carlo estimation based on only a small number $n$ of these function evaluations.

### 3.1.3 SARCOS robot arm

We used the hierarchical GP model in Sec. 3.1.2 to estimate the inverse dynamics of a seven degrees-of-freedom SARCOS anthropomorphic robot arm. The task, as described in Rasmussen and Williams (2006, Sec. 8.3.1), is to map from a 21-dimensional input space (7 positions, 7 velocities, 7 accelerations) to the corresponding 7 joint torques using the hierarchical GP model described in Sec. 3.1.1. Following Rasmussen and Williams (2006) we present results below on just one of the mappings, from the 21 input variables to the first of the seven torques. The dataset consists of 48,933 input-output pairs, of which 44,484 were used as a training set and the remaining 4,449 were used as a test set. The inputs were translated and scaled to have mean zero and unit variance on the training set. The outputs were centred so as to have mean zero on the training set. Here $\sigma = 0.1$, $\alpha = \gamma = 25$, $\beta = \delta = 0.04$, so that each hyper-parameter $\theta_i$ has a prior mean of 1 and a prior standard deviation of 0.2.

For each test point $\boldsymbol{z}_*$ we estimated the sampling standard deviation of $\hat{Y}_*$ over 10 independent realisations of the Monte Carlo sampling procedure. For CF we took default

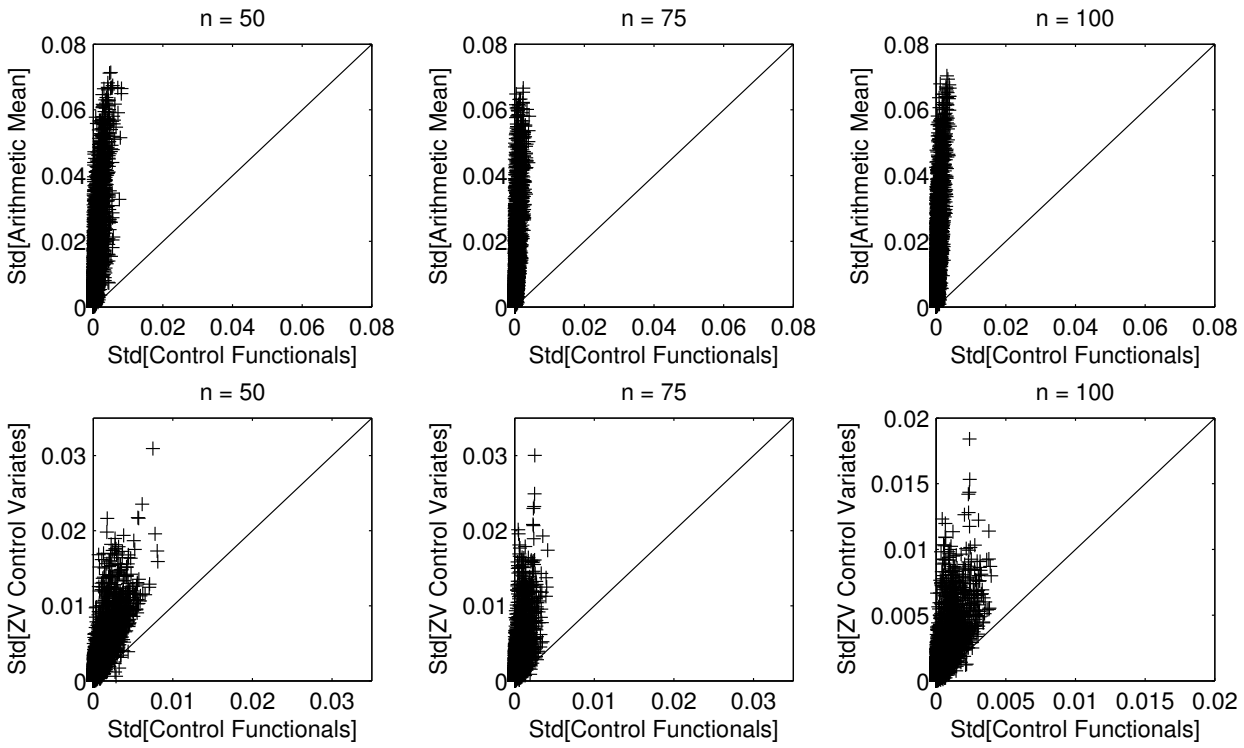


Figure 4: Marginalisation of hyper-parameters in hierarchical models. [Here we display the sampling standard deviation of Monte Carlo estimators for the posterior predictive mean $\mathbb{E}[Y_*|\boldsymbol{y}]$ in the SARCOS robot arm example, computed over 10 independent realisations. Each point, representing one Monte Carlo integration problem, is represented by a cross.]

hyper-parameters $\alpha_1 = 0.1$, $\alpha_2 = 1$, the latter reflecting the fact that the training data were standardised. The estimator standard deviations were estimated in this way for all 4,449 test samples and the full results are shown in Fig. 4. Note that each test sample corresponds to a different function $f$ and thus these results are quite objective, encompassing thousands of different Monte Carlo integration problems. Results show that, for the vast majority of integration problems, CF achieves a lower estimator variance compared with both the arithmetic mean estimator and ZV control variates. Here the cost of post-processing the Monte Carlo samples (using either ZV control variates or CF) is negligible in comparison to the cost of evaluating the function $f$, even once. Indeed, CF requires that we invert a $n \times n$ matrix once, where $n$ is no larger than $N'$ in this example.

In the supplement we investigate an extension that draws design points $\mathcal{D}_0$ using RQMC. Results show that CFs+RQMC outperforms RQMC alone.

## 3.2 Normalising constants for non-linear ODE models

Our second application concerns the estimation of normalising constants for non-linear ODE models (e.g. Calderhead and Girolami, 2009). Recent empirical investigations recommend

thermodynamic integration (TI) for this task (e.g. Friel and Wyse, 2012). The control variate method of Mira *et al.* (2003) was recently applied to TI by Oates *et al.* (2016), who found that this "controlled thermodynamic integral" (CTI) was extremely effective for standard regression models, but only moderately effective in complex models including non-linear ODEs due to poor approximation by low-degree polynomials. Below we study the application of CFs to TI in this setting where CTI is less effective.

### 3.2.1 Thermodynamic integration

Conditional on an inverse temperature parameter $t$, the "power posterior" for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{y}$ is defined as $p(\boldsymbol{\theta}|\boldsymbol{y}, t) \propto p(\boldsymbol{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta})$ (Friel and Pettitt, 2008). Varying $t \in [0, 1]$ produces a continuous path between the prior $p(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ and it is assumed here that all intermediate distributions exist and are well-defined. The standard thermodynamic identity is

$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]\mathrm{d}t, \tag{5}$$

where the expectation in the integrand is with respect to the power posterior. In TI, the one-dimensional integral in Eqn. 5 is evaluated numerically using a quadrature approximation over a discrete temperature ladder $0 = t_0 < t_1 < \cdots < t_m = 1$. Here we use the second-order quadrature recommended by Friel *et al.* (2014):

$$\log p(\boldsymbol{y}) \approx \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2}(\hat{\mu}_i + \hat{\mu}_{i+1}) - \frac{(t_{i+1} - t_i)^2}{12}(\hat{\nu}_{i+1} - \hat{\nu}_i),$$

where $\hat{\mu}_i$, $\hat{\nu}_i$ are Monte Carlo estimates of the posterior mean and variance respectively of $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ arises from $p(\boldsymbol{\theta}|\boldsymbol{y}, t_i)$. CTI uses ZV control variates to reduce the variance of these estimates. However, in complex models $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ will be poorly approximated by a low-degree polynomial and $p(\boldsymbol{\theta}|\boldsymbol{y}, t)$ will be non-Gaussian; this explains the mediocre performance of CTI in these cases. In contrast, CFs should still be able to deliver gains in estimation.

### 3.2.2 Non-linear ODE models

The approach is illustrated by computing the marginal likelihood for a non-linear ODE model (the van der Pol oscillator), described in full in the supplement. For TI, a temperature schedule $t_i = (i/30)^5$ was used, following the recommendation by Calderhead and Girolami (2009). The power posterior is not available in closed form, precluding the straight-forward generation of IID samples. Instead, samples from each of the power posteriors $p(\boldsymbol{\theta}|\boldsymbol{y}, t_i)$ were obtained using population MCMC, involving both (i) "within-temperature" proposals produced by the (simplified) m-MALA algorithm of Girolami and Calderhead (2011), and (ii) "between-temperature" proposals, as described previously by Calderhead and Girolami (2009). Gradient information is thus pre-computed in the sampling scheme and can be

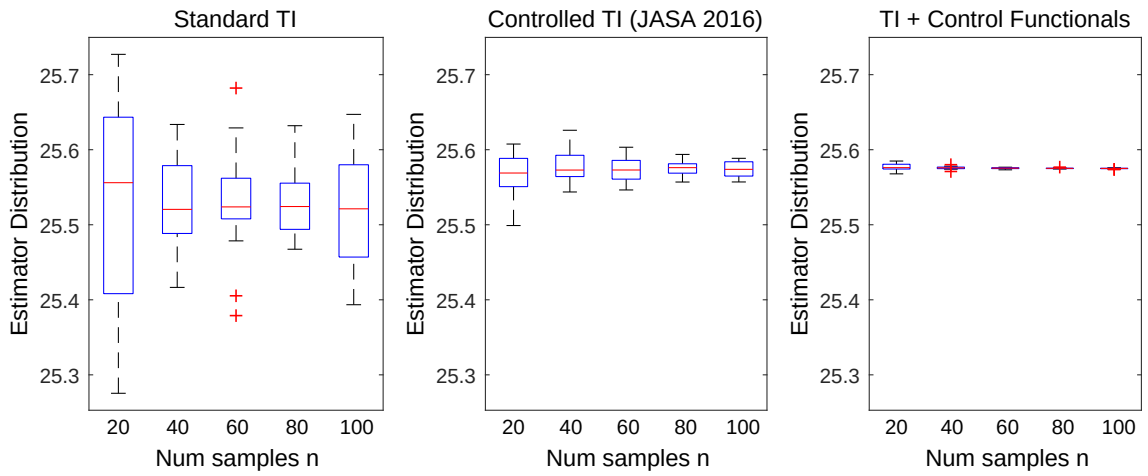


Figure 5: Estimation of normalising constants for non-linear ordinary differential equations using thermodynamic integration (TI); van der Pol oscillator example. [Here we show the distribution of 100 independent realisations of each estimator for $\log p(\boldsymbol{y})$. "Standard TI" is based on arithmetic means. "Controlled TI" is based on ZV control variates.]

leveraged "for free", as noted by Papamarkou *et al.* (2014). We denote the number of samples by $n$, such that for each of the 31 temperatures we obtained $n$ samples (a total of $31 \times n$ occasions where the system of ODEs was integrated numerically). Both sampling and evaluation of the integrand are computationally expensive, requiring the numerical solution of a system of ODEs.

Results in Fig. 5 show that the CTI estimator improves upon the standard TI estimator, but a more substantial reduction in estimator variance results from using the CF method. For the CF computation we have used the simplified but biased CF estimator, since TI in any case produces a biased estimate for the normalising constant due to numerical quadrature. The hyper-parameters $\alpha_1 = 0.1$, $\alpha_2 = 3$ were selected on the basis of cross-validation. The additional cost of using CF is essentially zero relative to running the population MCMC sampler, the latter requiring repeated solution of the ODE system.

# 4 Discussion

This paper developed a novel and general approach to integration that achieves super-root-$n$ convergence. An important feature of CFs is that variance reduction is formulated as a *post-hoc* step. This has several advantages: (i) No modification is required to existing computer code associated with either the sampling process or the model itself. (ii) Specific implementational choices, e.g. for the kernel, can be made *after* performing expensive simulations. Through exploitation of recent results in functional analysis we were able to realise our general framework and construct estimators with an analytic form. Empirical results evidenced the practical utility of CF estimators in settings where gradient information is available and

the dimensionality of the problem is not too large (e.g. $\leq 10$). The paper concludes below by suggesting directions for further research.

In terms of methodology: (i) The estimates we presented here are not parameterisation-invariant. Likewise the specification of $f$ and $\pi$ is not unique, as we can employ an importance sampling transformation $f \mapsto (f\pi)/\pi'$, $\pi \mapsto \pi'$. It would therefore be interesting to elicit effective parametrisations as an additional *post-hoc* step. (ii) The version of CFs presented here is limited in terms of the dimension of the problems for which it is effective. Techniques for high-dimensional functional approximation should be applicable in the context of CFs (e.g. Dick *et al.*, 2013) and this forms part of our ongoing research.

In terms of theory: (i) For bounded $\Omega$, sharper asymptotics are provided in a sequel, Oates *et al.* (2016). These account for various levels of smoothness of both $f$ and $\pi$ and help to explain the strong empirical results presented here. However the case of unbounded $\Omega$ seems considerably more challenging to characterise. (ii) For problems involving un-normalised densities $\pi$, sampling is naturally facilitated by MCMC. The analysis of CFs is carried out in Oates *et al.* (2016) under a uniform ergodicity assumption. For unbounded $\Omega$ this condition is too strong and future work will aim to relax this constraint.

In terms of application: (i) Our methods were motivated by the un-normalised densities arising in Bayesian computation. An extension should be possible to models with unknown, parameter-dependent normalising constants, which include e.g. Markov random fields (Everitt, 2012) and random network models (Friel *et al.*, 2015). (ii) An interesting direction would be to use the discrepancy $D(\mathcal{D}_0, \mathcal{D}_1)$ as a tool for assessment of MCMC convergence, providing a reproducing kernel Hilbert space alternative to Gorham and Mackey (2015).

Finally we note that Oates and Girolami (2016) provide a complementary study of CF strategies in the QMC setting.

**Acknowledgements:** The authors are grateful to the editor and referees, whose valuable feedback helped to improve the paper. The authors benefited from discussions with Sergios Agapiou, Michel Caffarel, Adam Johansen, Christian Robert, Daniel Simpson and Tim Sullivan. CJO was supported by EPSRC [EP/D002060/1] and the ARC Centre for Excellence in Mathematical and Statistical Frontiers. MG was supported by EPSRC [EP/J016934/1], EU [EU/259348], an EPSRC Established Career Fellowship and a Royal Society Wolfson Research Merit Award. NC was supported by the ANR (Agence Nationale de la Recherche) grant Labex ECODEC ANR [11-LABEX-0047].

# A Proofs

*Proof of Proposition 1.* We exploit the unbiasedness property $\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f)] = 0$ to show that

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f))^2] &= \mathbb{E}_{\mathcal{D}_0}\mathbb{V}_{\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)] \\ &= (n-m)^{-1}\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] \end{aligned}$$

where $(n-m)^{-1} = O(n^{-1})$ and by hypothesis $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(m^{-\delta})$. Thus using $m = O(n^\gamma)$ produces an overall rate $O(n^{-1-\gamma\delta})$, as required. □

*Proof of Proposition 2.* (A1) ensures $\psi$ is well-defined. Since $\partial\Omega$ is piecewise smooth we can apply the divergence theorem (e.g. Bourne and Kendall, 1977, p.159) to obtain

$$\mu(\psi) = \int_\Omega \psi(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \nabla_{\boldsymbol{x}} \cdot [\boldsymbol{\phi}(\boldsymbol{x})\pi(\boldsymbol{x})]\mathrm{d}\boldsymbol{x} = \oint_{\partial\Omega} \pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x})S(\mathrm{d}\boldsymbol{x}),$$

which is zero by (A2). The use of this identity in statistical applications is often attributed to Stein (1970). Thus $\mu(s_{f,\mathcal{D}_0}) = \mu(c) + \mu(\psi) = c$, as required. □

*Proof of Theorem 1. Stage 1:* We begin by defining the set of CFs $\mathcal{H}_0$. Given a reproducing kernel $k : \Omega \times \Omega \to \mathbb{R}$ for the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, define the canonical feature map $\Phi : \Omega \to \mathcal{H}$ by $\Phi(\boldsymbol{x}) = k(\cdot, \boldsymbol{x})$. Under (A1) the gradient function $\boldsymbol{u} : \Omega \to \mathbb{R}$ is well-defined. Under (A3) $k$ has mixed first order partial derivatives; it follows that all elements $\phi_i \in \mathcal{H}$ are differentiable and thus $(\partial/\partial x_i)\phi_i(\boldsymbol{x})$ is well-defined (Steinwart and Christmann, 2008, Cor. 4.36, p131). We then have that

$$\begin{aligned}
\psi(\boldsymbol{x}) &= \sum_{i=1}^{d} (\partial/\partial x_i)\phi_i(\boldsymbol{x}) + u_i(\boldsymbol{x})\phi_i(\boldsymbol{x}) \\
&= \sum_{i=1}^{d} (\partial/\partial x_i)\langle \phi_i, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}} + u_i(\boldsymbol{x})\langle \phi_i, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{d} \langle \phi_i, (\partial/\partial x_i)\Phi(\boldsymbol{x}) + u_i(\boldsymbol{x})\Phi(\boldsymbol{x})\rangle_{\mathcal{H}} = \sum_{i=1}^{d} \langle \phi_i, \Phi_i^*(\boldsymbol{x})\rangle_{\mathcal{H}},
\end{aligned}$$

where we have used the notation $\Phi_i^*(\boldsymbol{x}) = (\partial/\partial x_i)\Phi(\boldsymbol{x}) + u_i(\boldsymbol{x})\Phi(\boldsymbol{x})$. Write $\Phi^* : \Omega^d \to \mathbb{R}^d$ for the derived feature map with $i$th component $\Phi_i^*$. Define the set of all CFs $\psi$ of this form as

$$\mathcal{H}_0 = \{\psi : \Omega \to \mathbb{R} \text{ such that } \forall \boldsymbol{x} \in \Omega,\ \psi(\boldsymbol{x}) = \langle \boldsymbol{\phi}, \Phi^*(\boldsymbol{x})\rangle_{\mathcal{H}^d} \text{ for some } \boldsymbol{\phi} \in \mathcal{H}^d\}.$$

Clearly $\mathcal{H}_0$ is a vector space with addition and multiplication defined pointwise; $(\lambda\psi + \lambda'\psi')(\boldsymbol{x}) = \lambda\psi(\boldsymbol{x}) + \lambda'\psi'(\boldsymbol{x})$.

*Stage 2:* We now show that $\mathcal{H}_0$ can be endowed with the structure of a RKHS. To this end, define a norm on $\mathcal{H}_0$ by

$$\|\psi\|_{\mathcal{H}_0} := \inf_{\boldsymbol{\phi} \in \mathcal{H}^d} \{\|\boldsymbol{\phi}\|_{\mathcal{H}_d} \text{ such that } \forall \boldsymbol{x} \in \Omega,\ \psi(\boldsymbol{x}) = \langle \boldsymbol{\phi}, \Phi^*(\boldsymbol{x})\rangle_{\mathcal{H}^d}\}.$$

Theorem 4.21 (p121) of Steinwart and Christmann (2008) immediately gives that the normed space $(\mathcal{H}_0, \|\cdot\|_{\mathcal{H}_0})$ is a RKHS whose kernel $k_0$ satisfies $k_0(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi^*(\boldsymbol{x}), \Phi^*(\boldsymbol{x}')\rangle_{\mathcal{H}^d}$. Thus

we can directly calculate

$$\begin{aligned} k_0(\boldsymbol{x},\boldsymbol{x}') &= \langle \Phi^*(\boldsymbol{x}), \Phi^*(\boldsymbol{x}')\rangle_{\mathcal{H}^d} \\ &= \sum_{i=1}^{d} \langle (\partial/\partial x_i)\Phi(\boldsymbol{x}) + u_i(\boldsymbol{x})\Phi(\boldsymbol{x}), (\partial/\partial x_i')\Phi(\boldsymbol{x}') + u_i(\boldsymbol{x}')\Phi(\boldsymbol{x}')\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{d} \begin{array}{l} (\partial/\partial x_i)(\partial/\partial x_i')k(\boldsymbol{x},\boldsymbol{x}') + u_i(\boldsymbol{x})(\partial/\partial x_i')k(\boldsymbol{x},\boldsymbol{x}') \\ \qquad\qquad + u_i(\boldsymbol{x}')(\partial/\partial x_i)k(\boldsymbol{x},\boldsymbol{x}') + u_i(\boldsymbol{x})u_i(\boldsymbol{x}')k(\boldsymbol{x},\boldsymbol{x}'), \end{array} \end{aligned}$$

where the interchange of derivative and inner product is justified by (A3) and Lemma 4.34 (p130) in Steinwart and Christmann (2008). This completes the proof. □

*Proof of Lemma 1.* (A1,3) ensure the kernel $k_0$ is well-defined. Then

$$\begin{aligned} \int_\Omega k_0(\boldsymbol{x},\boldsymbol{x}')\pi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' &= \int_\Omega [\nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' + \int_\Omega [\boldsymbol{u}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' \\ &\quad + \int_\Omega [\boldsymbol{u}(\boldsymbol{x}') \cdot \nabla_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' + \int_\Omega [\boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}')k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' \\ &= \int_\Omega [\nabla_{\boldsymbol{x}'} \cdot \nabla_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}') + [\nabla_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')] \cdot [\nabla_{\boldsymbol{x}'}\pi(\boldsymbol{x}')]\mathrm{d}\boldsymbol{x}' \\ &\quad + \boldsymbol{u}(\boldsymbol{x}) \cdot \int_\Omega [\nabla_{\boldsymbol{x}'} k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}') + k(\boldsymbol{x},\boldsymbol{x}')[\nabla_{\boldsymbol{x}'}\pi(\boldsymbol{x}')]\mathrm{d}\boldsymbol{x}' \\ &= \int_\Omega \nabla_{\boldsymbol{x}'} \cdot \{[\nabla_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')]\pi(\boldsymbol{x}')\}\,\mathrm{d}\boldsymbol{x}' + \boldsymbol{u}(\boldsymbol{x}) \cdot \int_\Omega \nabla_{\boldsymbol{x}'} \{k(\boldsymbol{x},\boldsymbol{x}')\pi(\boldsymbol{x})\}\,\mathrm{d}\boldsymbol{x}'. \end{aligned}$$

Now using the divergence theorem (Bourne and Kendall, 1977, p.159) we obtain

$$= \underbrace{\oint_{\partial\Omega} \nabla_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}')\pi(\boldsymbol{x}') \cdot \boldsymbol{n}(\boldsymbol{x}')S(\mathrm{d}\boldsymbol{x}')}_{=0\ \pi\text{-a.e. from (A2')}} + \boldsymbol{u}(\boldsymbol{x}) \cdot \underbrace{\oint_{\partial\Omega} k(\boldsymbol{x},\boldsymbol{x}')\pi(\boldsymbol{x}')\boldsymbol{n}(\boldsymbol{x}')S(\mathrm{d}\boldsymbol{x}')}_{=\mathbf{0}\ \pi\text{-a.e. from (A2')}},$$

proving the claim. □

*Proof of Lemma 2.* From Theorem 1, (A1,3) ensure $\mathcal{H}_0$ is well-defined. Moreover, from Lemma 1, (A1,2',3) imply that $\mu(\psi) = 0$ and thus

$$\sigma^2(\psi) = \int_\Omega \psi(\boldsymbol{x})^2\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Now, given $\psi \in \mathcal{H}_0$, we need to show $\sigma^2(\psi) < \infty$. By the reproducing property followed by the Cauchy-Schwarz inequality, we have

$$|\psi(\boldsymbol{x})| = |\langle \psi, k_0(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}_0}| \leq \|\psi\|_{\mathcal{H}_0}\|k_0(\cdot,\boldsymbol{x})\|_{\mathcal{H}_0}.$$

Using the reproducing property again, we have $\|k_0(\cdot,\boldsymbol{x})\|^2_{\mathcal{H}_0} = k_0(\boldsymbol{x},\boldsymbol{x})$ and it follows from (A4) that

$$\sigma^2(\psi) = \int \psi(\boldsymbol{x})^2\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq \int_\Omega \|\psi\|^2_{\mathcal{H}_0} k_0(\boldsymbol{x},\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \|\psi\|^2_{\mathcal{H}_0}\int_\Omega k_0(\boldsymbol{x},\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < \infty,$$

as required. □

*Proof of Theorem 2.* From Theorem 1, (A1,3) ensure $\mathcal{H}_+$ is well-defined. Unbiasedness follows from (A1,2',3) and Lemma 2. Below we employ the standard notation

$$L^2(\pi) = \mathcal{L}^2(\pi) \setminus \{f \text{ such that } f = 0 \ \pi\text{-almost everywhere}\}$$

and denote the standard norm on this space by $\|\cdot\|_{L^2(\pi)}$.

For the remainder we appeal to the relatively recent work of Sun and Wu (2009), who considered convergence in a general setting where (i) $\Omega \cup \partial\Omega$ is not required to be compact in $\mathbb{R}^d$, and (ii) only weak assumptions are required on the kernel $k_+$, which can be easily satisfied in our setting. To this end, define the integral operator

$$(Tg)(\boldsymbol{x}) := \int_\Omega k_+(\boldsymbol{x},\boldsymbol{x}')g(\boldsymbol{x}')\pi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}', \qquad \boldsymbol{x}\in\Omega,\ g\in L^2(\pi).$$

In the well-posed setting of (A5), Theorem 1.1 of Sun and Wu (2009) establishes that if (i) $\sup_{\boldsymbol{x}\in\Omega} k_+(\boldsymbol{x},\boldsymbol{x}) < \infty$ and (ii) $T^{-1/2}f \in L^2(\pi)$, then with a RLS estimator based on $\lambda = O(m^{-1/2})$ we have $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(m^{-1/6})$. Inserting this rate into Proposition 1 with $\gamma = 1$, $\delta = 1/6$ would produce a MSE $\mathbb{E}_{\mathcal{D}_0}\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0,\mathcal{D}_1;f) - \mu(f))^2] = O(n^{-1-\gamma\delta}) = O(n^{-7/6})$.

It therefore remains to prove requirements (i) and (ii) above are satisfied. For (i) we have that $\sup_{\boldsymbol{x}\in\Omega} k_+(\boldsymbol{x},\boldsymbol{x}') = 1 + \sup_{\boldsymbol{x}\in\Omega} k_0(\boldsymbol{x},\boldsymbol{x})$, where the second term is finite by (A4'). For (ii), Prop. 3.3 of Sun and Wu (2009) (which does not depend on Theorem 1.1 of the same paper) shows that, when (i) holds, we have $\|T^{-1/2}h\|_{L^2(\pi)} = \|h\|_{\mathcal{H}_+}$ for all $h\in\mathcal{H}_+$. Since $f\in\mathcal{H}_+$ by (A5) we thus have $\|T^{-1/2}f\|_{L^2(\pi)} = \|f\|_{\mathcal{H}_+} < \infty$ and so $T^{-1/2}f \in L^2(\pi)$, as required. □

*Proof of Lemma 3.* From Theorem 1, (A1,3) ensure $\mathcal{H}_0$ is well-defined. The interpolation problem is equivalently expressed as $s_{f,\mathcal{D}_0} = \hat{c} + \hat{\psi}$ where

$$(\hat{c},\hat{\psi}) := \underset{c\in\mathcal{C},\ \psi\in\mathcal{H}_0}{\arg\min} \|c\|^2_{\mathcal{C}} + \|\psi\|^2_{\mathcal{H}_0} \text{ s.t. } f(\boldsymbol{x}_j) = c + \psi(\boldsymbol{x}_j) \text{ for } j = 1,...,m.$$

For fixed $c\in\mathcal{C}$, the representer theorem (Steinwart and Christmann, 2008, Thm. 5.5, p168) tells us that the solution

$$\hat{\psi} = \underset{\psi\in\mathcal{H}_0}{\arg\min}\|\psi\|_{\mathcal{H}_0} \text{ s.t. } f(\boldsymbol{x}_j) = c + \psi(\boldsymbol{x}_j) \text{ for } j = 1,...,m$$

takes the form $\psi(\boldsymbol{x}) = \sum_{i=1}^m \beta_i k_0(\boldsymbol{x}_i, \boldsymbol{x})$ where, due to the reproducing property, $\|\psi\|^2_{\mathcal{H}_0} = \boldsymbol{\beta}^T \boldsymbol{K}_0 \boldsymbol{\beta}$. Thus writing $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]^T$ reduces the problem to

$$(\hat{c}, \hat{\boldsymbol{\beta}}) = \underset{c \in \mathcal{C}, \boldsymbol{\beta} \in \mathbb{R}^m}{\arg\min}\, c^2 \text{ s.t. } f(\boldsymbol{x}_j) = c + \sum_{i=1}^m \beta_i k_0(\boldsymbol{x}_i, \boldsymbol{x}_j) \text{ for } j = 1, ..., m$$

Differentiating with respect to $c$ and $\boldsymbol{\beta}$ leads, via the Woodbury matrix inversion identity, to the solution

$$\hat{c} = \frac{\mathbf{1}^T \boldsymbol{K}_0^{-1} \boldsymbol{f}_0}{1 + \mathbf{1}^T \boldsymbol{K}_0^{-1} \mathbf{1}}, \qquad \hat{\boldsymbol{\beta}} = \boldsymbol{K}_0^{-1}(\boldsymbol{f}_0 - \hat{c}\mathbf{1})$$

and associated fitted values $\hat{\boldsymbol{f}}_1 = \hat{c}\mathbf{1} + \boldsymbol{K}_{1,0}\hat{\boldsymbol{\beta}}$ at the points $\mathcal{D}_1$. Putting this together, we have

$$\begin{aligned} \hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) \;=\; \frac{1}{n-m} \sum_{i=m+1}^n f_{\mathcal{D}_0}(\boldsymbol{x}_i) \;&=\; \frac{1}{n-m} \sum_{i=m+1}^n f(\boldsymbol{x}_i) - s_{f,\mathcal{D}_0}(\boldsymbol{x}_i) + \mu(s_{f,\mathcal{D}_0}) \\ &=\; \frac{1}{n-m} \mathbf{1}^T (\boldsymbol{f}_1 - \hat{\boldsymbol{f}}_1) + \hat{c}. \end{aligned}$$

This completes the proof. □

*Proof of Theorem 3.* From Theorem 1, (A1,3) ensure $\mathcal{H}_0$ is well-defined. The CF estimator takes the form

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \sum_{i=1}^n w_i f(\boldsymbol{x}_i) = \hat{c} + \sum_{i=1}^n w_i \psi(\boldsymbol{x}_i)$$

where, by Lemma 3, the vector of weights $\boldsymbol{w} = [w_1, \ldots, w_n]^T$ is given by

$$\boldsymbol{w} = \begin{bmatrix} -\frac{(\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1} \boldsymbol{K}_{0,1} \mathbf{1}}{n-m} + \frac{1}{n-m} \frac{(\mathbf{1}^T (\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1} \boldsymbol{K}_{0,1} \mathbf{1})(\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1} \mathbf{1}}{1 + \mathbf{1}^T (\boldsymbol{K}_0 + \lambda m \boldsymbol{I})^{-1} \mathbf{1}} \\ \frac{1}{n-m} \mathbf{1} \end{bmatrix} \tag{6}$$

and satisfies $\mathbf{1}^T \boldsymbol{w} = 1$. Using the reproducing property, the estimation error is

$$\begin{aligned} \hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f) \;&=\; \sum_{i=1}^n w_i f(\boldsymbol{x}_i) - \int_\Omega f(\boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &=\; \sum_{i=1}^n w_i \psi(\boldsymbol{x}_i) - \int_\Omega \psi(\boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &=\; \Big\langle \psi, \sum_{i=1}^n w_i k_0(\cdot, \boldsymbol{x}_i) - \underbrace{\int_\Omega k_0(\cdot, \boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}_{=0 \text{ from Lemma 1}} \Big\rangle_{\mathcal{H}_0}. \end{aligned}$$

It follows from the Cauchy-Schwarz inequality that

$$|\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f)| \leq \|\psi\|_{\mathcal{H}_0} \left\| \sum_{i=1}^n w_i k_0(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}_0}.$$

The first term satisfies $\|\psi\|^2_{\mathcal{H}_0} \leq \hat{c}^2 + \|\psi\|^2_{\mathcal{H}_0} = \|f\|^2_{\mathcal{H}_+}$ and, from the reproducing property, the second term satisfies

$$\left\| \sum_{i=1}^{n} w_i k_0(\cdot, \boldsymbol{x}_i) \right\|^2_{\mathcal{H}_0} = \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w}, \qquad \boldsymbol{K} = \begin{bmatrix} \boldsymbol{K}_0 & \boldsymbol{K}_{0,1} \\ \boldsymbol{K}_{1,0} & \boldsymbol{K}_1 \end{bmatrix}. \tag{7}$$

Finally, upon substituting Eqn. 6 into Eqn. 7 we obtain the required result with $D(\mathcal{D}_0, \mathcal{D}_1) = \boldsymbol{w}^T \boldsymbol{K} \boldsymbol{w}$. The special case $\lambda = 0$ is reported in the statement of the theorem. □

# References


Agapiou, S., Bardsley, J. M., Papaspiliopoulos, O. and Stuart, A. M. (2014) Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA J. Uncertainty Quantification*, **2**(1), 511-544.

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.

Angelikopoulos, P., Papadimitriou, C. and Koumoutsakos, P. (2012) Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework. *J. Chem. Phys.*, **137**, 144103.

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.

Ba, S. and Joseph, V. R. (2012) Composite Gaussian process models for emulating expensive functions. *Ann. Appl. Stat.*, **6**(4), 1838-1860.

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

Berlinet, A. and Thomas-Agnan, C. (2004) *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Boston.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B*, **55**, 25-37.

Bourne, D. E. and Kendall, P. C. (1977) *Vector analysis and Cartesian tensors* (2nd ed.). Nelson and Sons, UK.

Briol, F.X., Oates, C.J., Girolami, M., Osborne, M.A. (2016) Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees. *Adv. Neur. In.*, **28**.

Briol, F.X., Oates, C.J., Girolami, M., Osborne, M.A., Sejdinovic, D. (2016) Probabilistic Integration: A Role for Statisticians in Numerical Analysis? arXiv:1512.00933.

Calderhead, B. and Girolami, M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Stat. Data An.*, **53**, 4028-4045.

Chwialkowski, K., Strathmann, H. and Gretton, A. (2016) A Kernel Test of Goodness of Fit. arXiv:1602.02964.

Cornuet, J.-M., Marin, J.-M., Mira, A. and Robert, C. P. (2012) Adaptive Multiple Importance Sampling. *Scand. J. Stat.*, **39**, 798-812.

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

Dick, J. and Pillichshammer, F. (2010) *Discrepancy theory and quasi-Monte Carlo integration.* Springer, Berlin.

Dick, J., Kuo, F. Y. and Sloan, I. H. (2013) High-dimensional integration: the quasi-Monte Carlo way. *Acta Numer.*, **22**, 133-288.

Douc, R. and Robert, C. P. (2011) A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *Ann. Stat.*, **39**(1), 261-277.

Everitt, R. G. (2012) Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comp. Graph. Stat.*, **21**(4), 940-960.

Filippone, M. and Girolami, M. (2014) Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE T. Pattern Anal.*, **36**(11), 2214-2226.

Friel, N. and Pettitt, A. N. (2008) Marginal likelihood estimation via power posteriors. *J. R. Statist. Soc. B*, **70**, 589-607.

Friel, N. and Wyse, J. (2012) Estimating the statistical evidence - a review. *Stat. Neerl.*, **66**, 288-308.

Friel, N., Hurn, M. A. and Wyse, J. (2014) Improving power posterior estimation of statistical evidence. *Stat. Comp.*, **24**, 709-723.

Friel, N., Mira, A. and Oates, C. J. (2015) Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods. *Baysian Anal.*, **11**(1), 215-245.

Ghosh, J. and Clyde, M. A. (2011) Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *J. Am. Stat. Assoc.*, **106**(495), 1041-1052.

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

Giles, M. B. and Szpruch, L. (2014) Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *Ann. Appl. Prob.*, **24**(4), 1585-1620.

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

Glasserman, P. (2004) *Monte Carlo methods in financial engineering*. Springer, New York.

Gorham, J. and Mackey, L. (2015) Measuring Sample Quality with Stein's Method. *Adv. Neur. In.* **28**, 226-234.

Green, P. and Han, X. (1992) Metropolis methods, Gaussian proposals, and antithetic variables. *Lect. Notes Stat.*, **74**, 142-164.

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 and Applications*, **1**(4), 251-278.

Higdon, D., McDonnell, J. D., Schunck, N., Sarich, J. and Wild, S. M. (2015) A Bayesian approach for parameter estimation and prediction using a computationally intensive model. *J. Phys. G: Nucl. Part. Phys.*, **42**, 034009.

Kohlhoff, K. J., Shukla, D., Lawrenz, M., Bowman, G. R., Konerding, D. E., Belov, D., Altman, R. B. and Pande, V. S. (2014) Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.*, **6**(1), 15-21.

Kristoffersen, S. (2013) The empirical interpolation method. *Master's thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology.*

Łatuszyński, K., Green, P., Pereyra, M. and Robert, C. P. (2015) Bayesian computation: a perspective on the current state, and sampling backwards and forwards. *Stat. Comput.*, **25**(4), 835-862.

Li, W., Tan, Z. and Chen, R. (2013) Two-Stage Importance Sampling With Mixture Proposals. *J. Am. Stat. Assoc.*, **108**(504), 1350-1365.

Li, W., Chen, R. and Tan, Z. (2016) Efficient Sequential Monte Carlo with Multiple Proposals and Control Variates. *J. Am. Stat. Assoc.*, to appear.

Liu, Q., Lee, J. D. and Jordan, M. (2016) A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. arXiv:1602.03253.

Mira, A., Tenconi, P. and Bressanini, D. (2003) Variance reduction for MCMC. *Technical Report 2003/29, Universitá degli Studi dell' Insubria, Italy.*

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

Mizielinski, M. S., Roberts, M. J., Vidale, P. L., Schiemann, R., Demory, M. E., Strachan, J., Edwards, T., Stephens, A., Lawrence, B. N., Pritchard, M., Chiu, P., Iwi, A., Churchill, J., del Cano Novales, C., Kettleborough, J., Roseblade, W., Selwood, P., Foster, M., Glover, M. and Malcolm, A. (2014) High-resolution global climate modelling: the UPSCALE project, a large-simulation campaign. *Geosci. Model Dev.*, **7**(4), 1629-1640.

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

O'Hagan, A. (1991) Bayes-Hermite Quadrature. *J. Stat. Plan. Infer.*, **29**, 245-260.

Oates, C. J., Papamarkou, T. and Girolami, M. (2016) The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation. *J. Am. Stat. Assoc.*, to appear.

Oates, C. J. and Girolami, M. (2016) Control Functionals for Quasi Monte Carlo Integration. *In: Proc. 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, to appear.

Oates, C. J., Cockayne, J., Briol, F.-X. and Girolami, M. (2016) Convergence Rates for a Class of Estimators Based on Stein's Identity. arXiv:1603.03220.

Olsson, J. and Ryden, T. (2011) Rao-Blackwellization of particle Markov chain Monte Carlo methods using forward filtering backward sampling. *IEEE T. Signal Proces.*, **59**(10), 4606-4619.

Papamarkou, T., Mira, A. and Girolami, M. (2014) Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms. *Bayesian Anal.*, **9**:97-128,

Philippe, A. (1997) Processing simulation output by Riemann sums. *J. Statist. Comput. Simul.*, **59**, 295-314.

Rasmussen, C. E. and Ghahramani, Z. (2003) Bayesian Monte Carlo. *Adv. Neur. Inf.*, **17**, 505-512.

Rasmussen, C.E. and Williams, C.K. (2006) *Gaussian Processes for Machine Learning*. MIT Press.

Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods (2nd ed.)*. Springer-Verlag, New York.

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

Rubinstein, R. Y. and Kroese, D. P. (2011) *Simulation and the Monte Carlo method*. John Wiley and Sons, New Jersey.

Slingo, J., Bates, K., Nikiforakis, N., Piggott, M., Roberts, M., Shaffrey, L., Stevens, L., Vidale, P. L. and Weller, H. (2009) Developing the next-generation climate system models: challenges and achievements. *Philos. T. R. Soc. A*, **367**, 815-831.

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

Speight, A. (2009) A multilevel approach to control variates. *J. Comp. Finance*, **12**, 1-25.

Stein, C. (1970) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. 6th Berkeley Symp. Math. Statist. Prob.* **2**, 583-602.

Steinwart, I. and Christmann, A. (2008) *Support Vector Machines*. Springer, New York.

Sun, H. and Wu, Q. (2009) Application of integral operator for regularized least-square regression. *Math. Comput. Model.*, **49**(1), 276-285.