Biometrika (2018), **103**, 1, *pp.* 1–18
Printed in Great Britain Advance Access publication on 31 July 2018

Semi-Exact Control Functionals From Sard's Method

By L. F. South
School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD 4000, Australia.
l1.south@qut.edu.au

T. KARVONEN
The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK.
tkarvonen@turing.ac.uk

C. NEMETH
Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.
c.nemeth@lancaster.ac.uk

M. GIROLAMI
Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K.
mag92@eng.cam.ac.uk

and C. J. OATES
School of Mathematics, Statistics & Physics, Newcastle University, Newcastle NE1 7RU, U.K.
chris.oates@ncl.ac.uk

Summary

A novel control variate technique is proposed for post-processing of Markov chain Monte Carlo output, based both on Stein's method and an approach to numerical integration due to Sard. The resulting estimators of posterior expected quantities of interest are proven to be polynomially exact in the Gaussian context, while empirical results suggest the estimators approximate a Gaussian cubature method near the Bernstein-von-Mises limit. The main theoretical result establishes a bias-correction property in settings where the Markov chain does not leave the posterior invariant. Empirical results are presented across a selection of Bayesian inference tasks. All methods used in this paper are available in the `R` package `ZVCV`.

Some key words: Control Variate; Stein Operator; Variance Reduction.

1. Introduction

This paper focuses on the numerical approximation of integrals of the form

$$I(f) = \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

where f is a function of interest and p is a positive and continuously differentiable probability density on $\mathbb{R}^d$, under the restriction that p and its gradient can only be evaluated pointwise up to an intractable normalisation constant. The standard approach to computing $I(f)$ in this context is to simulate the first n steps of a p-invariant Markov chain $(\boldsymbol{x}^{(i)})_{i=1}^{\infty}$, possibly after an initial burn-in period, and to take the average along the sample path as an approximation to the integral:

$$I(f) \approx I_{\mathrm{MC}}(f) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}^{(i)}). \tag{1}$$

See Chapters 6–10 of Robert & Casella (2013) for background. In this paper $\mathbb{E}$, $\mathbb{V}$ and $\mathbb{C}$ respectively denote expectation, variance and covariance with respect to the law $\mathbb{P}$ of the Markov chain. Under regularity conditions on p that ensure the Markov chain $(\boldsymbol{x}^{(i)})_{i=1}^{\infty}$ is aperiodic, irreducible and reversible, the convergence of $I_{\mathrm{MC}}(f)$ to $I(f)$ as $n \to \infty$ is described by a central limit theorem

$$\sqrt{n}(I_{\mathrm{MC}}(f) - I(f)) \to \mathcal{N}(0, \sigma(f)^2) \tag{2}$$

where convergence occurs in distribution and, if the chain starts in stationarity,

$$\sigma(f)^2 = \mathbb{V}[f(\boldsymbol{x}^{(1)})] + 2\sum_{i=2}^{\infty} \mathbb{C}[f(\boldsymbol{x}^{(1)}), f(\boldsymbol{x}^{(i)})]$$

is the asymptotic variance of f along the sample path. See Theorem 4.7.7 of Robert & Casella (2013) and more generally Meyn & Tweedie (2012) for theoretical background. Note that for all but the most trivial function f we have $\sigma(f)^2 > 0$ and hence, to achieve an approximation error of $O_P(\epsilon)$, a potentially large number $O(\epsilon^{-2})$ of calls to f and p are required.

One approach to reduce the computational cost is to employ control variates (Hammersley & Handscomb, 1964; Ripley, 1987), which involves finding an approximation f_n to f that can be exactly integrated under p, such that $\sigma(f - f_n)^2 \ll \sigma(f)^2$. Given a choice of f_n, the standard estimator (1) is replaced with

$$I_{\mathrm{CV}}(f) = \frac{1}{n}\sum_{i=1}^{n}\{f(\boldsymbol{x}^{(i)}) - f_n(\boldsymbol{x}^{(i)})\} + \underbrace{\int f_n(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}_{(*)}, \tag{3}$$

where $(*)$ is exactly computed. This last requirement makes it challenging to develop control variates for general use, particularly in Bayesian statistics where often the density p can only be accessed in a form that is un-normalised. In the Bayesian context, Assaraf & Caffarel (1999); Mira et al. (2013) and Oates et al. (2017) addressed this challenge by using $f_n = c_n + \mathcal{L}g_n$ where $c_n \in \mathbb{R}$, g_n is a user-chosen parametric or non-parametric function and $\mathcal{L}$ is an operator, for example the Langevin Stein operator (Stein, 1972; Gorham & Mackey, 2015), that depends on p through its gradient and satisfies $\int(\mathcal{L}g_n)(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ under regularity conditions (see Lemma 1). Convergence of

$I_{\mathrm{CV}}(f)$ to $I(f)$ has been studied under (strong) regularity conditions and, in particular (i) if g_n is chosen parametrically, then in general $\liminf \sigma(f - f_n)^2 > 0$ so that, even if asymptotic variance is reduced, convergence rates are unaffected; (ii) if g_n is chosen in an appropriate non-parametric manner then $\limsup \sigma(f - f_n)^2 = 0$ and a smaller number $O(\epsilon^{-2+\delta})$, $0 < \delta < 2$, of calls to f, p and its gradient are required to achieve an approximation error of $O_P(\epsilon)$ for the integral (see Oates et al., 2019; Mijatović & Vogrinc, 2018; Barp et al., 2018; Belomestny et al., 2017, 2019, 2020). In the parametric case $\mathcal{L}g_n$ is called a *control variate* while in the non-parametric case it is called a *control functional*.

Practical parametric approaches to the choice of g_n have been well-studied in the Bayesian context, typically based on polynomial regression models (Assaraf & Caffarel, 1999; Mira et al., 2013; Papamarkou et al., 2014; Oates et al., 2016; Brosse et al., 2019), but neural networks have also been proposed recently (Wan et al., 2019; Si et al., 2020). In particular, existing control variates based on polynomial regression have the attractive property of being *semi-exact*, meaning that there is a well-characterized set of functions $f \in \mathcal{F}$ for which f_n can be shown to exactly equal f after a finite number of samples n have been obtained. For the control variates of Assaraf & Caffarel (1999) and Mira et al. (2013) the set $\mathcal{F}$ contains certain low order polynomials when p is a Gaussian distribution on $\mathbb{R}^d$. Those authors term their control variates *zero variance*, but we prefer the term *semi-exact* since a general integrand f will not be an element of $\mathcal{F}$. Regardless of terminology, semi-exactness of the control variate is an appealing property because it implies that the approximation $I_{\mathrm{CV}}(f)$ to $I(f)$ is exact on $\mathcal{F}$. Intuitively, the performance of the control variate method is related to the richness of the set $\mathcal{F}$ on which it is exact. For example, polynomial exactness of cubature rules is used to establish their high order convergence rates using a Taylor expansion argument (e.g. Hildebrand, 1987, Chapter 8).

The development of non-parametric approaches to the choice of g_n has to-date focused on kernel methods (Oates et al., 2017; Barp et al., 2018), piecewise constant approximations (Mijatović & Vogrinc, 2018) and non-linear approximations based on selecting basis functions from a dictionary (Belomestny et al., 2017; South et al., 2019). Theoretical analysis of non-parametric control variates was provided in the papers cited above, but compared to parametric methods, practical implementations of non-parametric methods are less well-developed.

In this paper we propose a semi-exact control functional method. This constitutes the best of both worlds, where at small n the semi-exactness property promotes stability and robustness of the estimator $I_{\mathrm{CV}}(f)$, while at large n the non-parametric regression component can be used to accelerate the convergence of $I_{\mathrm{CV}}(f)$ to $I(f)$. In particular we argue that, in the Bernstein-von-Mises limit, the set $\mathcal{F}$ on which our method is exact is precisely the set of low order polynomials, so that our method can be considered as an approximately polynomially-exact cubature rule developed for the Bayesian context. Furthermore, we establish a bias-correcting property, which guarantees the approximations produced using our method are consistent in certain settings where the Markov chain is not p-invariant.

Our motivation comes from the approach to numerical integration due to Sard (1949). Many numerical integration methods are based on constructing an approximation f_n to the integrand f that can be exactly integrated. In this case the integral $I(f)$ is approximated using $(*)$ in (3). In Gaussian and related cubatures, the function f_n is chosen in such a way that polynomial exactness is guaranteed (Gautschi, 2004, Section 1.4). On the other hand, in kernel cubature and related approaches, f_n is an element of a reproduc-

ing kernel Hilbert space chosen such that an error criterion is minimised (Larkin, 1970). The contribution of Sard was to combine these two concepts in numerical integration by choosing f_n to enforce exactness on a low-dimensional space $\mathcal{F}$ of functions and use the remaining degrees of freedom to find a minimum-norm interpolant to the integrand.

2. Methods

2.1. *Sard's Method*

Many popular methods for numerical integration are based on either (i) enforcing *exactness* of the integral estimator on a finite-dimensional set of functions $\mathcal{F}$, typically a linear space of polynomials, or on (ii) integration of a *minimum-norm interpolant* selected from an infinite-dimensional set of functions $\mathcal{H}$. In each case, the result is a cubature method of the form

$$I_{\mathrm{NI}}(f) = \sum_{i=1}^{n} w_i f(\boldsymbol{x}^{(i)}) \tag{4}$$

for weights $\{w_i\}_{i=1}^n \subset \mathbb{R}$ and points $\{\boldsymbol{x}^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$. Classical examples of methods in the former category include univariate Gaussian quadrature rules (Gautschi, 2004, Section 1.4), which are determined by the unique $\{(w_i, \boldsymbol{x}^{(i)})\}_{i=1}^n \subset \mathbb{R} \times \mathbb{R}^d$ such that $I_{\mathrm{NI}}(f) = I(f)$ whenever f is a polynomial of order at most $2n-1$, and Clenshaw–Curtis rules (Clenshaw & Curtis, 1960). Methods in the minimum-norm interpolant category specify a suitable normed space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ of functions, construct an interpolant $f_n \in \mathcal{H}$ such that

$$f_n \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \left\{ \|h\|_{\mathcal{H}} \,:\, h(\boldsymbol{x}^{(i)}) = f(\boldsymbol{x}^{(i)}) \text{ for } i = 1, \ldots, n \right\} \tag{5}$$

and use the integral of f_n to approximate the true integral. Specific examples include splines (Wahba, 1990) and kernel or Gaussian process based methods (Larkin, 1970; O'Hagan, 1991; Briol et al., 2019).

If the set of points $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ is fixed, the cubature method in (4) has n degrees of freedom corresponding to the choice of the weights $\{w_i\}_{i=1}^n$. The approach proposed by Sard (1949) is a hybrid of the two classical approaches just described, calling for $m \leq n$ of these degrees of freedom to be used to ensure that $I_{\mathrm{NI}}(f)$ is exact for f in a given m-dimensional linear function space $\mathcal{F}$ and, if $m < n$, allocating the remaining $n - m$ degrees of freedom to select a minimal norm interpolant from a large class of functions $\mathcal{H}$. The approach of Sard is therefore exact for functions in the finite-dimensional set $\mathcal{F}$ and, at the same time, suitable for the integration of functions in the infinite-dimensional set $\mathcal{H}$. Further background on Sard's method can be found in Larkin (1974) and Karvonen et al. (2018).

However, it is difficult to implement Sard's method, or indeed any of the classical approaches just discussed, in the Bayesian context, since

1. the density p can be evaluated pointwise only up to an intractable normalization constant;
2. to construct weights one needs to evaluate the integrals of basis functions of $\mathcal{F}$ and of the interpolant f_n, which can be as difficult as evaluating the original integral.

To circumvent these issues, in this paper we propose to combine Sard's approach to integration with Stein operators (Stein, 1972; Gorham & Mackey, 2015), thus eliminating the need to access normalization constants and to exactly evaluate integrals.

2.2. *Stein Operators*

Let $\cdot$ denote the dot product $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^\top \boldsymbol{b}$, $\nabla_{\boldsymbol{x}}$ denote the gradient $\nabla_{\boldsymbol{x}} = [\partial_{x_1}, \ldots, \partial_{x_d}]^\top$ and $\Delta_{\boldsymbol{x}}$ denote the Laplacian $\Delta_{\boldsymbol{x}} = \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}}$. Let $\|\boldsymbol{x}\| = (\boldsymbol{x} \cdot \boldsymbol{x})^{1/2}$ denote the Euclidean norm on $\mathbb{R}^d$. The construction that enables us to realize Sard's method in the Bayesian context is the Langevin Stein operator $\mathcal{L}$ (Gorham & Mackey, 2015) on $\mathbb{R}^d$, defined for sufficiently regular g and p as

$$(\mathcal{L}g)(\boldsymbol{x}) = \Delta_{\boldsymbol{x}} g(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}). \tag{6}$$

We refer to $\mathcal{L}$ as a Stein operator due to the use of equations of the form (6) (up to a simple substitution) in the method of Stein (1972) for assessing convergence in distribution and due to its property of producing functions whose integrals with respect to p are zero under suitable conditions such as those described in Lemma 1.

LEMMA 1. *If $g\colon \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, $\log p : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and $\|\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\| \leq C\|\boldsymbol{x}\|^{-\delta} p(\boldsymbol{x})^{-1}$ is satisfied for some $C \in \mathbb{R}$ and $\delta > d - 1$, then*

$$\int (\mathcal{L}g)(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0,$$

where $\mathcal{L}$ is the Stein operator in (6).

The proof is provided in Appendix 1. Although our attention is limited to (6), the choice of Stein operator is not unique and other Stein operators can be derived using the generator method of Barbour (1988) or using Schrödinger Hamiltonians (Assaraf & Caffarel, 1999). Contrary to the standard requirements for a Stein operator, the operator $\mathcal{L}$ in control functionals does not need to fully characterize convergence and, as a consequence, a broader class of functions g can be considered than in more traditional applications of Stein's method (Stein, 1972).

It follows that, if the conditions of Lemma 1 are satisfied by $g_n : \mathbb{R}^d \to \mathbb{R}$, the integral of a function of the form $f_n = c_n + \mathcal{L}g_n$ is simply c_n, the constant. The main challenge in developing control variates, or functionals, based on Stein operators is therefore to find a function g_n such that the asymptotic variance $\sigma(f - f_n)^2$ is small. To explicitly minimize asymptotic variance, Mijatović & Vogrinc (2018); Belomestny et al. (2020) and Brosse et al. (2019) restricted attention to particular Metropolis–Hastings or Langevin samplers for which asymptotic variance can be explicitly characterized. The minimization of empirical variance has also been proposed and studied in the case where samples are independent (Belomestny et al., 2017) and dependent (Belomestny et al., 2020, 2019). For an approach that is not tied to a particular Markov kernel, authors such as Assaraf & Caffarel (1999) and Mira et al. (2013) proposed to minimize mean squared error along the sample path, which corresponds to the case of an independent sampling method. In a similar spirit, the constructions in Oates et al. (2017, 2019) and Barp et al. (2018) were based on a minimum-norm interpolant, where the choice of norm is decoupled from the mechanism from where the points are sampled.

2.3. *The Proposed Method*

In this section we first construct an infinite-dimensional space $\mathcal{H}$ and a finite-dimensional space $\mathcal{F}$ of functions; these will underpin the proposed semi-exact control functional method.

For the infinite-dimensional component, let $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive-definite *kernel*, meaning that (i) k is symmetric, with $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, and (ii) the *kernel matrix* $[\boldsymbol{K}]_{i,j} = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ is positive-definite for any distinct points $\{\boldsymbol{x}^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$ and any $n \in \mathbb{N}$. Recall that such a k induces a unique *reproducing kernel Hilbert space* $\mathcal{H}(k)$. This is a Hilbert space that consists of functions $g\colon \mathbb{R}^d \to \mathbb{R}$ and is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(k)}$. The kernel k is such that $k(\cdot, \boldsymbol{x}) \in \mathcal{H}(k)$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and it is *reproducing* in the sense that $\langle g, k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}(k)} = g(\boldsymbol{x})$ for any $g \in \mathcal{H}(k)$ and $\boldsymbol{x} \in \mathbb{R}^d$. For $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ the multi-index notation $\boldsymbol{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$ will be used. If k is twice continuously differentiable in the sense of Steinwart & Christmann (2008, Definition 4.35), meaning that the derivatives

$$\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} \partial_{\boldsymbol{y}}^{\boldsymbol{\alpha}} k(\boldsymbol{x}, \boldsymbol{y}) = \frac{\partial^{2|\boldsymbol{\alpha}|}}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}} \partial \boldsymbol{y}^{\boldsymbol{\alpha}}} k(\boldsymbol{x}, \boldsymbol{y})$$

exist and are continuous for every multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \leq 2$, then

$$k_0(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}), \tag{7}$$

where $\mathcal{L}_{\boldsymbol{x}}$ stands for application of the Stein operator defined in (6) with respect to variable $\boldsymbol{x}$, is a well-defined and positive-definite kernel (Steinwart & Christmann, 2008, Lemma 4.34). The kernel in (7) can be written as

$$\begin{aligned} k_0(\boldsymbol{x}, \boldsymbol{y}) = {} & \Delta_{\boldsymbol{x}} \Delta_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{u}(\boldsymbol{x})^\top \nabla_{\boldsymbol{x}} \Delta_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}) \\ & + \boldsymbol{u}(\boldsymbol{y})^\top \nabla_{\boldsymbol{y}} \Delta_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{u}(\boldsymbol{x})^\top \big\{ \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}}^\top k(\boldsymbol{x}, \boldsymbol{y}) \big\} \boldsymbol{u}(\boldsymbol{y}), \end{aligned} \tag{8}$$

where $\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}}^\top k(\boldsymbol{x}, \boldsymbol{y})$ is the $d \times d$ matrix with entries $[\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}}^\top k(\boldsymbol{x}, \boldsymbol{y})]_{i,j} = \partial_{x_i} \partial_{y_j} k(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{u}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$. If k is radial then (8) can be simplified; see Appendix 2. Lemma 2 establishes conditions under which the functions $\boldsymbol{x} \mapsto k_0(\boldsymbol{x}, \boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{R}^d$, and hence elements of the Hilbert space $\mathcal{H}(k_0)$ reproduced by k_0, have zero integral. Let $\|\boldsymbol{M}\|_{\mathrm{OP}} = \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{M}\boldsymbol{x}\|$ denote the operator norm of a matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$.

LEMMA 2. *If $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable in each argument, $\log p : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, $\|\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}}^\top k(\boldsymbol{x}, \boldsymbol{y})\|_{\mathrm{OP}} \leq C(\boldsymbol{y}) \|\boldsymbol{x}\|^{-\delta} p(\boldsymbol{x})^{-1}$ and $\|\nabla_{\boldsymbol{x}} \Delta_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y})\| \leq C(\boldsymbol{y}) \|\boldsymbol{x}\|^{-\delta} p(\boldsymbol{x})^{-1}$ are satisfied for some $C : \mathbb{R}^d \to (0, \infty)$, and $\delta > d - 1$, then*

$$\int k_0(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 0 \tag{9}$$

for every $\boldsymbol{y} \in \mathbb{R}^d$, where k_0 is defined in (7).

The proof is provided in Appendix 4. The infinite-dimensional space $\mathcal{H}$ used in this work is exactly the reproducing kernel Hilbert space $\mathcal{H}(k_0)$. The basic mathematical properties of k_0 and the Hilbert space it reproduces are contained in Appendix 3 and these can be used to inform the selection of an appropriate kernel.

For the finite-dimensional component, let Φ be a linear space of twice-continuously differentiable functions with dimension $m - 1$, $m \in \mathbb{N}$, and a basis $\{\phi_i\}_{i=1}^{m-1}$. Define

then the space obtained by applying the differential operator (6) to Φ as $\mathcal{L}\Phi = \text{span}\{\mathcal{L}\phi_1, \ldots, \mathcal{L}\phi_{m-1}\}$. If the pre-conditions of Lemma 1 are satisfied for each basis function $g = \phi_i$ then linearity of the Stein operator implies that $\int(\mathcal{L}\phi)\mathrm{d}p = 0$ for every $\phi \in \Phi$. Typically we will select $\Phi = \mathcal{P}^r$ as the polynomial space $\mathcal{P}^r = \text{span}\{\boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{N}_0^d, 0 < |\boldsymbol{\alpha}| \leq r\}$ for some non-negative integer r. Note that constant functions are excluded from $\mathcal{P}^r$ since they are in the null space of $\mathcal{L}$; when required we let $\mathcal{P}_0^r = \text{span}\{1\} \oplus \mathcal{P}^r$ denote the larger space with the constant functions included. The finite-dimensional space $\mathcal{F}$ is then taken to be $\mathcal{F} = \text{span}\{1\} \oplus \mathcal{L}\Phi = \text{span}\{1, \mathcal{L}\phi_1, \ldots \mathcal{L}\phi_{m-1}\}$.

It is now possible to state the proposed method. Following Sard, we approximate the integrand f with a function f_n that interpolates f at the locations $\boldsymbol{x}^{(i)}$, is exact on the m-dimensional linear space $\mathcal{F}$, and minimises a particular (semi-)norm subject to the first two constraints. It will occasionally be useful to emphasise the dependence of f_n on f using the notation $f_n(\cdot) = f_n(\cdot; f)$. The proposed interpolant takes the form

$$f_n(\boldsymbol{x}) = b_1 + \sum_{i=1}^{m-1} b_{i+1}(\mathcal{L}\phi_i)(\boldsymbol{x}) + \sum_{i=1}^{n} a_i k_0(\boldsymbol{x}, \boldsymbol{x}^{(i)}), \tag{10}$$

where the coefficients $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ are selected such that the following two conditions hold:

1. $f_n(\boldsymbol{x}^{(i)}; f) = f(\boldsymbol{x}^{(i)})$ for $i = 1, \ldots, n$ (interpolation);
2. $f_n(\cdot; f) = f(\cdot)$ whenever $f \in \mathcal{F}$ (semi-exactness).

Since $\mathcal{F}$ is m-dimensional, these requirements correspond to the total of $n + m$ constraints. Under weak conditions, discussed in Section 2.5, the total number of degrees of freedom due to selection of $\boldsymbol{a}$ and $\boldsymbol{b}$ is equal to $n + m$ and the above constraints can be satisfied. Furthermore, the corresponding function f_n can be shown to minimise a particular (semi-)norm on a larger space of functions, subject to the interpolation and exactness constraints (to limit scope, we do not discuss this characterisation further but the semi-norm is defined in (17) and the reader can find full details in Wendland, 2004, Theorem 13.1). Figure 1 illustrates one such interpolant. The proposed estimator of the integral is then

$$I_{\text{SECF}}(f) = \int f_n(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{11}$$

a special case of (3) (the interpolation condition causes the first term in (3) to vanish) that we call a *semi-exact control functional*. The following is immediate from (10) and (11):

Corollary 1. *Under the hypotheses of Lemma* 1 *for each* $g = \phi_i$, $i = 1, \ldots, m-1$, *and Lemma* 2, *it holds that, whenever the estimator* $I_{\text{SECF}}(f)$ *is well-defined,* $I_{\text{SECF}}(f) = b_1$, *where* b_1 *is the constant term in* (10).

The earlier work of Assaraf & Caffarel (1999) and Mira et al. (2013) corresponds to $\boldsymbol{a} = \boldsymbol{0}$ and $\boldsymbol{b} \neq \boldsymbol{0}$, while setting $\boldsymbol{b} = \boldsymbol{0}$ in (10) and ignoring the semi-exactness requirement recovers the unique minimum-norm interpolant in the Hilbert space $\mathcal{H}(k_0)$ where k_0 is reproducing, in the sense of (5). The work of Oates et al. (2017) corresponds to $b_i = 0$ for $i = 2, \ldots, m$. It is therefore clear that the proposed approach is a strict generalization of existing work and can be seen as a compromise between semi-exactness and minimum-norm interpolation.

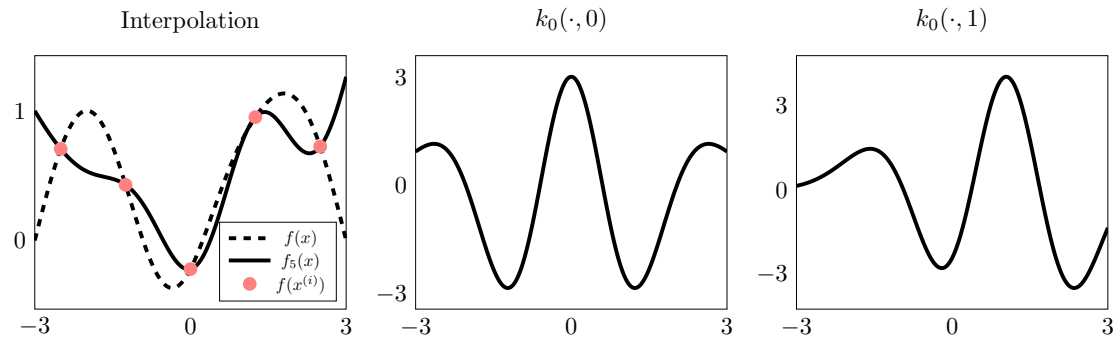

Fig. 1: *Left*: The interpolant f_n from (10) at $n = 5$ points to the function $f(x) = \sin\{0.5\pi(x-1)\} + \exp\{-(x-0.5)^2\}$ for the Gaussian density $p(x) = \mathcal{N}(x; 0, 1)$. The interpolant uses the Gaussian kernel $k(x, y) = \exp\{-(x-y)^2\}$ and a polynomial parametric basis with $r = 2$. *Center & right*: Two translates $k_0(\cdot, y)$, $y \in \{0, 1\}$, of the kernel (7).

2.4. *Polynomial Exactness in the Bernstein-von-Mises Limit*

A central motivation for our approach is the prototypical case where p is the density of a posterior distribution $P_{\boldsymbol{x}|y_1,\ldots,y_N}$ for a latent variable $\boldsymbol{x}$ given independent and identically distributed data $y_1, \ldots, y_N \sim P_{y_1,\ldots,y_N|\boldsymbol{x}}$. Under regularity conditions discussed in Section 10.2 of van der Vaart (1998), the Bernstein-von-Mises theorem states that

$$\left\| P_{\boldsymbol{x}|y_1,\ldots,y_N} - \mathcal{N}\big(\hat{\boldsymbol{x}}_N, N^{-1} I(\hat{\boldsymbol{x}}_N)^{-1}\big) \right\|_{\mathrm{TV}} \to 0$$

where $\hat{\boldsymbol{x}}_N$ is a maximum likelihood estimate for $\boldsymbol{x}$, $I(\boldsymbol{x})$ is the Fisher information matrix evaluated at $\boldsymbol{x}$, $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm and convergence is in probability as $N \to \infty$ with respect to the law $P_{y_1,\ldots,y_N|\boldsymbol{x}}$ of the dataset. In this limit, polynomial exactness of the proposed method can be established. Indeed, for a Gaussian density p with mean $\hat{\boldsymbol{x}}_N \in \mathbb{R}^d$ and precision $NI(\hat{\boldsymbol{x}}_N)$, if $\phi(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}}$ for a multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, then

$$(\mathcal{L}\phi)(\boldsymbol{x}) = \sum_{i=1}^{d} \alpha_i \left\{ (\alpha_i - 1) x_i^{\alpha_i - 2} - \frac{N}{2} P_i(\boldsymbol{x}) x_i^{\alpha_i - 1} \right\} \prod_{j \neq i} x_j^{\alpha_j},$$

where $P_i(\boldsymbol{x}) = 2\boldsymbol{e}_i^\top I(\hat{\boldsymbol{x}}_N)(\boldsymbol{x} - \hat{\boldsymbol{x}}_N)$ and $\boldsymbol{e}_i$ is the ith coordinate vector in $\mathbb{R}^d$. This allows us to obtain the following result, whose proof is provided in Appendix 5:

LEMMA 3. *Consider the Bernstein-von-Mises limit and suppose that the Fisher information matrix $I(\hat{\boldsymbol{x}}_N)$ is non-singular. Then, for the choice $\Phi = \mathcal{P}^r$, $r \in \mathbb{N}$, the estimator I_{SECF} is exact on $\mathcal{F} = \mathcal{P}_0^r$.*

Thus the proposed estimator is polynomially exact up to order r in the Bernstein-von-Mises limit. We believe this property can confer robustness of the estimator in a broad range of applied contexts. At finite N, when the limit has not been reached, the above argument can only be expected to approximately hold.

2.5. *Computation for the Proposed Method*

Define the $n \times m$ matrix

$$\boldsymbol{P} = \begin{bmatrix} 1 & \mathcal{L}\phi_1(\boldsymbol{x}^{(1)}) & \cdots & \mathcal{L}\phi_{m-1}(\boldsymbol{x}^{(1)}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathcal{L}\phi_1(\boldsymbol{x}^{(n)}) & \cdots & \mathcal{L}\phi_{m-1}(\boldsymbol{x}^{(n)}) \end{bmatrix}, \tag{12}$$

which is sometimes called a *Vandermonde* (or *alternant*) matrix corresponding to the linear space $\mathcal{F}$. Let $\boldsymbol{K}_0$ be the $n \times n$ matrix with entries $[\boldsymbol{K}_0]_{i,j} = k_0(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ and let $\boldsymbol{f}$ be the n-dimensional column vector with entries $[\boldsymbol{f}]_i = f(\boldsymbol{x}^{(i)})$.

LEMMA 4. *Let the $n \geq m$ points $\boldsymbol{x}^{(i)}$ be distinct and $\mathcal{F}$-*unisolvent*, meaning that the matrix $\boldsymbol{P}$ in* (12) *has full rank. Let k_0 be a positive-definite kernel for which* (9) *is satisfied. Then $I_{\mathrm{SECF}}(f)$ is well-defined and the coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$ are given by the solution of the linear system*

$$\begin{bmatrix} \boldsymbol{K}_0 & \boldsymbol{P} \\ \boldsymbol{P}^\top & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{0} \end{bmatrix}. \tag{13}$$

In particular,

$$I_{\mathrm{SECF}}(f) = \boldsymbol{e}_1^\top (\boldsymbol{P}^\top \boldsymbol{K}_0^{-1} \boldsymbol{P})^{-1} \boldsymbol{P}^\top \boldsymbol{K}_0^{-1} \boldsymbol{f}. \tag{14}$$

The proof is provided in Appendix 6. Notice that (14) is a linear combination of the values in $\boldsymbol{f}$ and therefore the proposed estimator is recognized as a cubature method of the form (4) with weights

$$\boldsymbol{w} = \boldsymbol{K}_0^{-1} \boldsymbol{P} (\boldsymbol{P}^\top \boldsymbol{K}_0^{-1} \boldsymbol{P})^{-1} \boldsymbol{e}_1. \tag{15}$$

The requirement in Lemma 4 for the $\boldsymbol{x}^{(i)}$ to be distinct precludes, for example, the direct use of Metropolis–Hastings output. However, as emphasized in Oates et al. (2017) for control functionals and studied further in Liu & Lee (2017); Hodgkinson et al. (2020), the consistency of I_{SECF} does *not* require that the Markov chain is p-invariant. It is therefore trivial to, for example, filter out duplicate states from Metropolis–Hastings output.

The solution of linear systems of equations defined by an $n \times n$ matrix $\boldsymbol{K}_0$ and an $m \times m$ matrix $\boldsymbol{P}^\top \boldsymbol{K}_0^{-1} \boldsymbol{P}$ entails a computational cost of $O(n^3 + m^3)$. In some situations this cost may yet be smaller than the cost associated with evaluation of f and p, but in general this computational requirement limits the applicability of the method just described. In Appendix 7 we therefore propose a computationally efficient approximation, I_{ASECF}, to the full method, based on a combination of the Nyström approximation (Williams & Seeger, 2001) and the well-known conjugate gradient method, inspired by the recent work of Rudi et al. (2017). All proposed methods are implemented in the `R` package `ZVCV` (South, 2020).

3. Empirical Assessment

3.1. *Experiment Setup*

A detailed comparison of existing and proposed control variate and control functional techniques was performed. Three examples were considered; Section 3.2 considers a Gaussian target, representing the Bernstein-von-Mises limit; Section 3.3 considers a setting

where non-parametric control functional methods perform well; Section 3.4 considers a setting where parametric control variate methods are known to be successful. In each case we determine whether or not the proposed semi-exact control functional method is competitive with the state-of-the-art.

Specifically, we compared the following estimators, which are all instances of I_{CV} in (3) for a particular choice of f_n, which may or may not be an interpolant:

- Standard Monte Carlo integration, (1), based on Markov chain output.
- The control functional estimator recommended in Oates et al. (2017), $I_{\mathrm{CF}}(f) = (\mathbf{1}^\top \boldsymbol{K}_0^{-1}\mathbf{1})^{-1}\mathbf{1}^\top \boldsymbol{K}_0^{-1}\boldsymbol{f}$.
- The *zero variance* polynomial control variate method of Assaraf & Caffarel (1999) and Mira et al. (2013), $I_{\mathrm{ZV}}(f) = \boldsymbol{e}_1^\top(\boldsymbol{P}^\top\boldsymbol{P})^{-1}\boldsymbol{P}^\top\boldsymbol{f}$.
- The *auto zero variance* approach of South et al. (2019), which uses 5-fold cross validation to automatically select (a) between the ordinary least squares solution I_{ZV} and an ℓ_1-penalised alternative (where the penalisation strength is itself selected using 10-fold cross-validation within the test dataset), and (b) the polynomial order.
- The proposed semi-exact control functional estimator, (14).
- An approximation, I_{ASECF}, of (14) based on the Nyström approximation and the conjugate gradient method, described in Appendix 7.

Open-source software for implementing all of the above methods is available in the `R` package `ZVCV` (South, 2020). The same sets of n samples were used for all estimators, in both the construction of f_n and the evaluation of I_{CV}. For methods where there is a fixed polynomial basis we considered only orders $r = 1$ and $r = 2$, following the recommendation of Mira et al. (2013). For kernel-based methods, duplicate values of $\boldsymbol{x}_i$ were removed (as discussed in Section 2.5) and Frobenius regularization was employed whenever the condition number of the kernel matrix $\boldsymbol{K}_0$ was close to machine precision (Higham, 1988). Several choices of kernel were considered, but for brevity in the main text we focus on the rational quadratic kernel $k(\boldsymbol{x}, \boldsymbol{y}; \lambda) = \{1 + \lambda^{-2}\|\boldsymbol{x} - \boldsymbol{y}\|^2\}^{-1}$. This kernel was found to provide the best performance across a range of experiments; a comparison to the Matérn and Gaussian kernels is provided in Appendix 8. The parameter λ was selected using 5-fold cross-validation, based again on performance across a spectrum of experiments; a comparison to the median heuristic (Garreau et al., 2017) is presented in Appendix 8.

To ensure that our assessment is practically relevant, the estimators were compared on the basis of both statistical and computational efficiency relative to the standard Monte Carlo estimator. Statistical efficiency $\mathcal{E}(I_{\mathrm{CV}})$ and computational efficiency $\mathcal{C}(I_{\mathrm{CV}})$ of an estimator I_{CV} of the integral I are defined as

$$\mathcal{E}(I_{\mathrm{CV}}) = \frac{\mathbb{E}\left[(I_{\mathrm{MC}} - I)^2\right]}{\mathbb{E}\left[(I_{\mathrm{CV}} - I)^2\right]}, \qquad \mathcal{C}(I_{\mathrm{CV}}) = \mathcal{E}(I_{\mathrm{CV}})\frac{T_{\mathrm{MC}}}{T_{\mathrm{CV}}}$$

where T_{CV} denotes the combined wall time for sampling the $\boldsymbol{x}^{(i)}$ and computing the estimator I_{CV}. For the results reported below, $\mathcal{E}$ and $\mathcal{C}$ were approximated using averages $\hat{\mathcal{E}}$ and $\hat{\mathcal{C}}$ over 100 realizations of the Markov chain output.

3.2. *Gaussian Illustration*

Here we consider a Gaussian integral that serves as an analytically tractable caricature of a posterior near to the Bernstein-von-Mises limit. This enables us to assess

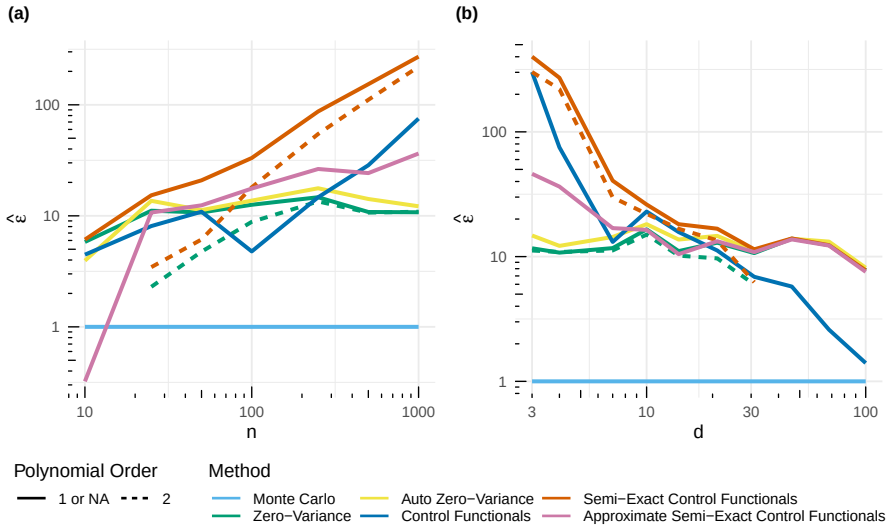

Fig. 2: Gaussian example (a) estimated *statistical efficiency* with $d = 4$ and (b) estimated *statistical efficiency* with $n = 1000$ for integrand (16).

the effect of the sample size n and dimension d on each estimator, in a setting that is not confounded by the idiosyncrasies of any particular MCMC method. Specifically, we set $p(\boldsymbol{x}) = (2\pi)^{-d/2} \exp(-\|\boldsymbol{x}\|^2/2)$ where $\boldsymbol{x} \in \mathbb{R}^d$. For the parametric component we set $\Phi = \mathcal{P}^r$, so that (from Lemma 3) I_{SECF} is exact on polynomials of order at most r; this holds also for I_{ZV}. For the integrand $f : \mathbb{R}^d \to \mathbb{R}$, $d \geq 3$, we took

$$f(\boldsymbol{x}) = 1 + x_2 + 0.1 x_1 x_2 x_3 + \sin(x_1) \exp\{-(x_2 x_3)^2\} \tag{16}$$

in order that the integral is analytically tractable ($I(f) = 1$) and that no method will be exact.

Figure 2 displays the statistical efficiency of each estimator for $10 \leq n \leq 1000$ and $3 \leq d \leq 100$. Computational efficiency is not shown since exact sampling from p in this example is trivial. The proposed semi-exact control functional method performs consistently well compared to its competitors for this non-polynomial integrand. Unsurprisingly, the best improvements are for high n and small d, where the proposed method results in a statistical efficiency over 100 times better than the baseline estimator and up to 5 times better than the next best method.

3.3. *Capture-Recapture Example*

The two remaining examples, here and in Section 3.4, are applications of Bayesian statistics described in South et al. (2019). In each case the aim is to estimate expectations with respect to a posterior distribution $P_{\boldsymbol{x}|\boldsymbol{y}}$ of the parameters $\boldsymbol{x}$ of a statistical model based on $\boldsymbol{y}$, an observed dataset. Samples $\boldsymbol{x}^{(i)}$ were obtained using the Metropolis-adjusted Langevin algorithm (Roberts & Tweedie, 1996), which is a Metropolis-Hastings algorithm with proposal $\mathcal{N}(\boldsymbol{x}^{(i-1)} + h^2 \frac{1}{2} \boldsymbol{\Sigma} \nabla_{\boldsymbol{x}} \log P_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}^{(i-1)} \mid \boldsymbol{y}), h^2 \boldsymbol{\Sigma})$. Step sizes of $h = 0.72$ for the capture-recapture example and $h = 0.3$ for the sonar example (see Section 3.4) were selected and an empirical approximation of the posterior covariance matrix was used as the pre-conditioner $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. Since the proposed method does not rely on the Markov chain being $P_{\boldsymbol{x}|\boldsymbol{y}}$-invariant we also repeated these experiments using the un-

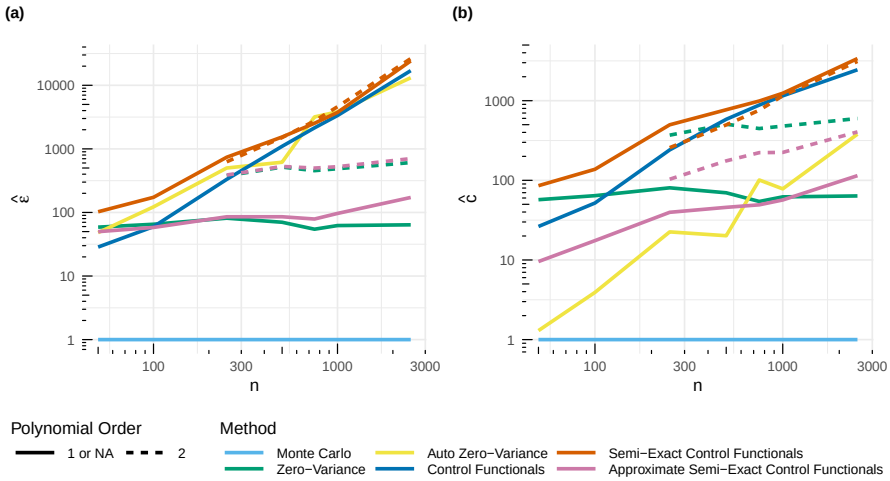

Fig. 3: Capture-recapture example (a) estimated *statistical efficiency* and (b) estimated *computational efficiency*. Efficiency here is reported as an average over the 11 expectations of interest.

adjusted Langevin algorithm (Parisi, 1981; Ermak, 1975), with similar results reported in Appendix 9.

In this first example, a Cormack–Jolly–Seber capture-recapture model (Lebreton et al., 1992) is used to model data on the capture and recapture of the bird species *Cinclus Cinclus* (Marzolin, 1988). The integrands of interest are the marginal posterior means $f_i(\boldsymbol{x}) = x_i$ for $i = 1, \ldots, 11$, where $\boldsymbol{x} = (\phi_1, \ldots, \phi_5, p_2, \ldots, p_6, \phi_6 p_7)$, ϕ_j is the probability of survival from year j to $j+1$ and p_j is the probability of being captured in year j. The likelihood is

$$\ell(\boldsymbol{y}|\boldsymbol{x}) \propto \prod_{i=1}^{6} \chi_i^{d_i} \prod_{k=i+1}^{7} \left\{ \phi_i p_k \prod_{m=i+1}^{k-1} \phi_m (1-p_m) \right\}^{y_{ik}},$$

where $d_i = D_i - \sum_{k=i+1}^{7} y_{ik}$, $\chi_i = 1 - \sum_{k=i+1}^{7} \phi_i p_k \prod_{m=i+1}^{k-1} \phi_m (1-p_m)$ and the data $\boldsymbol{y}$ consists of D_i, the number of birds released in year i, and y_{ik}, the number of animals caught in year k out of the number released in year i, for $i = 1, \ldots, 6$ and $k = 2, \ldots, 7$. Following South et al. (2019), parameters are transformed to the real line using $\tilde{x}_j = \log\{x_j/(1-x_j)\}$ and the adjusted prior density for $\tilde{x}_j$ is $\exp(\tilde{x}_j)/\{1+\exp(\tilde{x}_j)\}^2$, for $j = 1, \ldots, 11$.

South et al. (2019) found that non-parametric methods outperform standard parametric methods for this 11-dimensional example. The estimator I_{SECF} combines elements of both approaches, so there is interest in determining how the method performs. It is clear from Figure 3 that all variance reduction approaches are helpful in improving upon the vanilla Monte Carlo estimator in this example. The best improvement in terms of statistical and computational efficiency is offered by I_{SECF}, which also has similar performance to I_{CF}.

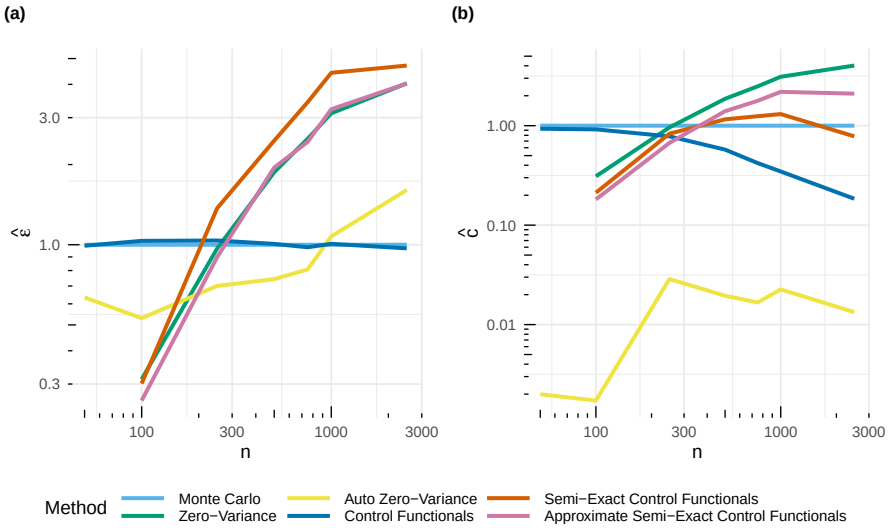

Fig. 4: Sonar example (a) estimated *statistical efficiency* and (b) estimated *computational efficiency*.

3.4. *Sonar Example*

Our final application is a 61-dimensional logistic regression example using data from Gorman & Sejnowski (1988) and Dheeru & Karra Taniskidou (2017). To use standard regression notation, the parameters are denoted $\boldsymbol{\beta} \in \mathbb{R}^{61}$, the matrix of covariates in the logistic regression model is denoted $\boldsymbol{X} \in \mathbb{R}^{208\times 61}$ where the first column is all 1's to fit an intercept and the response is denoted $\boldsymbol{y} \in \mathbb{R}^{208}$. In this application, $\boldsymbol{X}$ contains information related to energy frequencies reflected from either a metal cylinder ($y = 1$) or a rock ($y = 0$). The log likelihood for this model is

$$\log \ell(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{\beta}) = \sum_{i=1}^{208} \left[y_i \boldsymbol{X}_{i,\cdot}\boldsymbol{\beta} - \log\{1 + \exp(\boldsymbol{X}_{i,\cdot}\boldsymbol{\beta})\} \right].$$

We use a $\mathcal{N}(0, 5^2)$ prior for the predictors (after standardising to have standard deviation of 0.5) and $\mathcal{N}(0, 20^2)$ prior for the intercept, following South et al. (2019); Chopin & Ridgway (2017), but we focus on estimating the more challenging integrand $f(\boldsymbol{\beta}) = \{1 + \exp(-\tilde{\boldsymbol{X}}\boldsymbol{\beta})\}^{-1}$, which can be interpreted as the probability that observed covariates $\tilde{\boldsymbol{X}}$ emanate from a metal cylinder. The gold standard of $I \approx 0.4971$ was obtained from a 10 million iteration Metropolis-Hastings (Hastings, 1970) run with multivariate normal random walk proposal.

Figure 4 illustrates the statistical and computational efficiency of estimators for various n in this example. It is interesting to note that I_{SECF} and I_{ASECF} offer similar statistical efficiency to I_{ZV}, especially given the poor relative performance of I_{CF}. Since it is inexpensive to obtain the m samples using the Metropolis-adjusted Langevin algorithm in this example, I_{ZV} and I_{ASECF} are the only approaches which offer improvements in computational efficiency over the baseline estimator for the majority of n values considered, and even in these instances the improvements are marginal.

4. Theoretical Properties and Convergence Assessment

4.1. *Finite Sample Error and a Practical Diagnostic*

The performance of the proposed method can be monitored using the finite sample error bound provided in Proposition 1. Proposition 1 makes use of the semi-norm

$$|f|_{k_0,\mathcal{F}} = \inf_{\substack{f=h+g \\ h\in\mathcal{F},\, g\in\mathcal{H}(k_0)}} \|g\|_{\mathcal{H}(k_0)}, \tag{17}$$

which is well-defined when the infimum is taken over a non-empty set, otherwise $|f|_{k_0,\mathcal{F}} := \infty$.

Proposition 1. *Let the hypotheses of Corollary* 1 *hold. Then the integration error satisfies the bound*

$$|I(f) - I_{\mathrm{SECF}}(f)| \leq |f|_{k_0,\mathcal{F}} \, (\boldsymbol{w}^\top \boldsymbol{K}_0 \boldsymbol{w})^{1/2} \tag{18}$$

where the weights $\boldsymbol{w}$*, defined in* (15)*, satisfy*

$$\boldsymbol{w} = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathbb{R}^n} (\boldsymbol{v}^\top \boldsymbol{K}_0 \boldsymbol{v})^{1/2} \quad \textit{s.t.} \quad \sum_{i=1}^{n} v_i h(\boldsymbol{x}^{(i)}) = \int h(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \textit{for every } h \in \mathcal{F}.$$

The proof is provided in Appendix 11. The first quantity $|f|_{k_0,\mathcal{F}}$ in (18) can be approximated by $|f_n|_{k_0,\mathcal{F}}$ when f_n is a reasonable approximation for f and this can in turn can be bounded as $|f_n|_{k_0,\mathcal{F}} \leq (\boldsymbol{a}^\top \boldsymbol{K}_0 \boldsymbol{a})^{1/2}$. The finiteness of $|f|_{k_0,\mathcal{F}}$ ensures the existence of a solution to the Stein equation, sufficient conditions for which are discussed in Mackey & Gorham (2016); Si et al. (2020). The second quantity $(\boldsymbol{w}^\top \boldsymbol{K}_0 \boldsymbol{w})^{1/2}$ in (18) is computable and is recognized as a *kernel Stein discrepancy* between the empirical measure $\sum_{i=1}^n w_i \delta(\boldsymbol{x}^{(i)})$ and the distribution whose density is p, based on the Stein operator $\mathcal{L}$ (Chwialkowski et al., 2016; Liu et al., 2016). Note that our choice of Stein operator differs to that in Chwialkowski et al. (2016) and Liu et al. (2016). There has been substantial recent research into the use of kernel Stein discrepancies for assessing algorithm performance in the Bayesian computational context (Gorham & Mackey, 2017; Chen et al., 2018, 2019; Singhal et al., 2019; Hodgkinson et al., 2020) and one can also exploit this discrepancy as a diagnostic for the performance of the semi-exact control functional. The diagnostic that we propose to monitor is the product $(\boldsymbol{w}^\top \boldsymbol{K}_0 \boldsymbol{w})^{1/2} (\boldsymbol{a}^\top \boldsymbol{K}_0 \boldsymbol{a})^{1/2}$. This approach to error estimation was also suggested (outside the Bayesian context) in Section 5.1 of Fasshauer (2011).

Empirical results in Figure 5 suggest that this diagnostic provides a conservative approximation of the actual error. Further work is required to establish whether this diagnostic detects convergence and non-convergence in general.

4.2. *Consistency of the Estimator*

In what follows we consider an increasing number n of samples $\boldsymbol{x}^{(i)}$ whilst the finite-dimensional space Φ, with basis $\{\phi_1, \ldots, \phi_{m-1}\}$, is held fixed. The samples $\boldsymbol{x}^{(i)}$ will be assumed to arise from a V-uniformly ergodic Markov chain; the reader is referred to Chapter 16 of Meyn & Tweedie (2012) for the relevant background. Recall that the points $(\boldsymbol{x}^{(i)})_{i=1}^n$ are called $\mathcal{F}$-*unisolvent* if the matrix in (12) has full rank. It will be convenient to introduce an inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_n = \boldsymbol{u}^\top \boldsymbol{K}_0^{-1} \boldsymbol{v}$ and associated norm $\|\boldsymbol{u}\|_n = \langle \boldsymbol{u}, \boldsymbol{u} \rangle_n^{1/2}$. Let Π be the matrix that projects orthogonally onto the columns of $[\boldsymbol{\Psi}]_{i,j} := \mathcal{L}\phi_j(\boldsymbol{x}^{(i)})$ with respect to the $\langle \cdot, \cdot \rangle_n$ inner product.

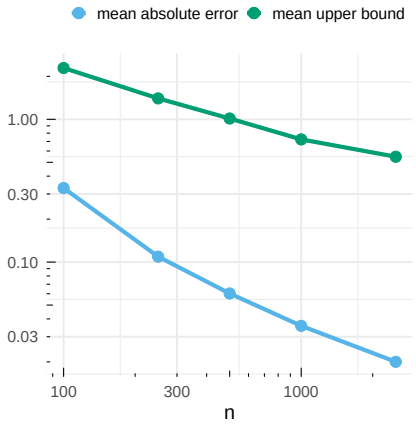

Fig. 5: The mean absolute error and mean of the approximate upper bound $(\boldsymbol{w}^\top \boldsymbol{K}_0 \boldsymbol{w})^{1/2}(\boldsymbol{a}^\top \boldsymbol{K}_0 \boldsymbol{a})^{1/2}$, for different values of n in the sonar example of Section 3.4. Both are based on the semi-exact control functional method with $\Phi = \mathcal{P}^1$.

THEOREM 1. *Let the hypotheses of Corollary* 1 *hold and let* f *be any function for which* $|f|_{k_0,\mathcal{F}} < \infty$. *Let* q *be a probability density with* $p/q > 0$ *on* $\mathbb{R}^d$ *and consider a* q*-invariant Markov chain* $(\boldsymbol{x}^{(i)})_{i=1}^n$, *assumed to be* V*-uniformly ergodic for some* $V : \mathbb{R}^d \to [1, \infty)$, *such that*

A1. $\sup_{\boldsymbol{x}\in\mathbb{R}^d} V(\boldsymbol{x})^{-r} \{p(\boldsymbol{x})/q(\boldsymbol{x})\}^4 k_0(\boldsymbol{x},\boldsymbol{x})^2 < \infty$ *for some* $0 < r < 1$;
A2. the points $(\boldsymbol{x}^{(i)})_{i=1}^n$ *are almost surely distinct and* $\mathcal{F}$*-unisolvent;*
A3. $\limsup_{n\to\infty} \|\Pi\mathbf{1}\|_n / \|\mathbf{1}\|_n < 1$ *almost surely.*

Then $|I_{\mathrm{SECF}}(f) - I(f)| = O_P(n^{1/2})$.

This demonstrates that, even in the biased sampling setting, the proposed estimator is consistent. The proof is provided in Appendix 12 and exploits a recent theoretical contribution in Hodgkinson et al. (2020). Assumption A1 serves to ensure that q is similar enough to p that a q-invariant Markov chain will also explore the high probability regions of p, as discussed in Hodgkinson et al. (2020). Sufficient conditions for V-uniform ergodicity are necessarily Markov chain dependent. The case of the Metropolis-adjusted Langevin algorithm is discussed in Roberts & Tweedie (1996); Chen et al. (2019) and, in particular, Theorem 9 of Chen et al. (2019) provides sufficient conditions for V-uniform ergodicity with $V(\boldsymbol{x}) = \exp(s\|\boldsymbol{x}\|)$ for all $s > 0$. Under these conditions, and with the rational quadratic kernel k considered in Section 3, we have $k_0(\boldsymbol{x},\boldsymbol{x}) = O(\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\|^2)$ and therefore A1 is satisfied whenever $\{p(\boldsymbol{x})/q(\boldsymbol{x})\}\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})\| = O(\exp(t\|\boldsymbol{x}\|))$ for some $t > 0$; a weak requirement. Assumption A2 ensures that the finite sample size bound (18) is almost surely well-defined. Assumption A3 ensures the points in the sequence $(\boldsymbol{x}^{(i)})_{i=1}^n$ distinguish (asymptotically) the constant function from the functions $\{\phi_i\}_{i=1}^{m-1}$, which is a weak technical requirement.

5. Discussion

Several possible extensions of the proposed method can be considered. For example, the parametric component Φ could be adapted to the particular f and p using a dimensionality reduction method. Likewise, extending cross-validation to encompass the choice of kernel and even the choice of control variate or control functional estimator may be useful. The potential for alternatives to the Nyström approximation to further improve scalability of the method can also be explored. In terms of the points $\boldsymbol{x}^{(i)}$ on which the estimator is defined, these could be optimally selected to minimize the error bound in (18), for example following the approaches of Chen et al. (2018, 2019). Finally, we highlight a possible extension to the case where only stochastic gradient information is available, following Friel et al. (2016) in the parametric context.

Acknowledgement

CJO is grateful to Yvik Swan for discussion of Stein's method. TK was supported by the Aalto ELEC Doctoral School and the Vilho, Yrjö and Kalle Väisälä Foundation. MG was supported by a Royal Academy of Engineering Research Chair and by EPSRC grants EP/T000414/1, EP/R018413/2, EP/P020720/2, EP/R034710/1, EP/R004889/1. TK, MG and CJO were supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. CN and LFS were supported by EPSRC grants EP/S00159X/1 and EP/V022636/1. The authors are grateful for feedback from three anonymous Reviewers, an Associate Editor and an Editor.

Supplementary material

Supplementary material available at *Biometrika* online includes further technical details, additional simulation results and the proofs of results stated in the main text.

References

Assaraf, R. & Caffarel, M. (1999). Zero-variance principle for Monte Carlo algorithms. *Physical Review Letters* **83**, 4682–4685.

Barbour, A. D. (1988). Stein's method and Poisson process convergence. *Journal of Applied Probability* **25**, 175–184.

Barp, A., Oates, C. J., Porcu, E. & Girolami, M. (2018). A Riemann-Stein kernel method. *arXiv preprint arXiv:1810.04946* .

Belomestny, D., Iosipoi, L., Moulines, E., Naumov, A. & Samsonov, S. (2020). Variance reduction for Markov chains with application to MCMC. *Statistics and Computing* **30**, 973–997.

Belomestny, D., Iosipoi, L. & Zhivotovskiy, N. (2017). Variance reduction via empirical variance minimization: convergence and complexity. *arXiv preprint arXiv:1712.04667* .

Belomestny, D., Moulines, E., Shagadatov, N. & Urusov, M. (2019). Variance reduction for MCMC methods via martingale representations. *arXiv preprint arXiv:1903.07373* .

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A. & Sejdinovic, D. (2019). Probabilistic integration: A role in statistical computation? (with discussion and rejoinder). *Statistical Science* **34**, 1–22.

Brosse, N., Durmus, A., Meyn, S., Moulines, É. & Radhakrishnan, A. (2019). Diffusion approximations and control variates for MCMC. *arXiv preprint arXiv:1808.01665* .

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L. & Oates, C. (2019). Stein point Markov chain Monte Carlo. In *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri & R. Salakhutdinov, eds., vol. 97 of *Proceedings of Machine Learning Research*. PMLR.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X. & Oates, C. J. (2018). Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, J. Dy & A. Krause, eds., vol. 80 of *Proceedings of Machine Learning Research*. PMLR.

Chopin, N. & Ridgway, J. (2017). Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Statistical Science* **32**, 64–87.

Chwialkowski, K., Strathmann, H. & Gretton, A. (2016). A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan & K. Q. Weinberger, eds., vol. 48 of *Proceedings of Machine Learning Research*. New York, New York, USA: PMLR.

Clenshaw, C. W. & Curtis, A. R. (1960). A method for numerical integration on an automatic computer. *Numerische Mathematik* **2**, 197–205.

Dheeru, D. & Karra Taniskidou, E. (2017). UCI machine learning repository.

Ermak, D. L. (1975). A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *The Journal of Chemical Physics* **62**, 4189–4196.

Fasshauer, G. E. (2011). Positive-definite kernels: Past, present and future. *Dolomites Research Notes on Approximation* **4**, 21–63.

Friel, N., Mira, A. & Oates, C. J. (2016). Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. *Bayesian Analysis* **11**, 215–245.

Garreau, D., Jitkrittum, W. & Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269* .

Gautschi, W. (2004). *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press.

Gorham, J. & Mackey, L. (2015). Measuring sample quality with Stein's method. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press.

Gorham, J. & Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, D. Precup & Y. W. Teh, eds., vol. 70 of *Proceedings of Machine Learning Research*. PMLR.

Gorman, R. P. & Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks* **1**, 75–89.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo Methods*. Chapman & Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and Its Applications* **103**, 103–118.

Hildebrand, F. B. (1987). *Introduction to Numerical Analysis*. Courier Corporation.

Hodgkinson, L., Salomone, R. & Roosta, F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266* .

Karvonen, T., Oates, C. J. & Särkkä, S. (2018). A Bayes–Sard cubature method. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, vol. 31.

Larkin, F. M. (1970). Optimal approximation in Hilbert spaces with reproducing kernel functions. *Mathematics of Computation* **24**, 911–921.

Larkin, F. M. (1974). Probabilistic error estimates in spline interpolation and quadrature. In *Information Processing 74: Proceedings of IFIP Congress 74*. North-Holland.

Lebreton, J. D., Burnham, K. P., Clobert, J. & Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs* **61**, 67–118.

Liu, Q. & Lee, J. (2017). Black-box importance sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh & J. Zhu, eds., vol. 54 of *Proceedings of Machine Learning Research*. Fort Lauderdale, FL, USA: PMLR.

Liu, Q., Lee, J. & Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan & K. Q. Weinberger, eds., vol. 48 of *Proceedings of Machine Learning Research*. New York, New York, USA: PMLR.

Mackey, L. & Gorham, J. (2016). Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability* **21**.

Marzolin, G. (1988). Polygynie du cincle plongeur (cinclus cinclus) dans le côtes de Loraine. *Oiseau et la Revue Francaise d'Ornithologie* **58**, 277–286.

Meyn, S. P. & Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Mijatović, A. & Vogrinc, J. (2018). On the Poisson equation for Metropolis–Hastings chains. *Bernoulli* **24**, 2401–2428.

Mira, A., Solgi, R. & Imparato, D. (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing* **23**, 653–662.

Oates, C. J., Cockayne, J., Briol, F.-X. & Girolami, M. (2019). Convergence rates for a class of estimators based on Stein's method. *Bernoulli* **25**, 1141–1159.

Oates, C. J., Girolami, M. & Chopin, N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 695–718.

Oates, C. J., Papamarkou, T. & Girolami, M. (2016). The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association* **111**, 634–645.

O'Hagan, A. (1991). Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference* **29**, 245–260.

Papamarkou, T., Mira, A. & Girolami, M. (2014). Zero variance differential geometric Markov chain Monte Carlo algorithms. *Bayesian Analysis* **9**, 97–128.

Parisi, G. (1981). Correlation functions and computer simulations. *Nuclear Physics B* **180**, 378–384.

Ripley, B. (1987). *Stochastic Simulation*. John Wiley & Sons.

Robert, C. & Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Roberts, G. O. & Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363.

Rudi, A., Carratino, L. & Rosasco, L. (2017). FALKON: An optimal large scale kernel method. In *Proceedings of the 31st Conference on Neural Information Processing Systems*. Curran Associates Inc.

Sard, A. (1949). Best approximate integration formulas; best approximation formulas. *American Journal of Mathematics* **71**, 80–91.

Si, S., Oates, C., Duncan, A. B., Carin, L. & Briol, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. *arXiv preprint arXiv:2006.07487* .

Singhal, R., Lahlou, S. & Ranganath, R. (2019). Kernelized complete conditional Stein discrepancy. *arXiv preprint arXiv:1904.04478* .

South, L. F. (2020). *ZVCV: Zero-Variance Control Variates*. R package version 2.1.0.

South, L. F., Oates, C. J., Mira, A. & Drovandi, C. (2019). Regularised zero-variance control variates for high-dimensional variance reduction. *arXiv preprint arXiv:1811.05073* .

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2. University of California Press.

Steinwart, I. & Christmann, A. (2008). *Support Vector Machines*. Information Science and Statistics. Springer.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press.

Wahba, G. (1990). *Spline Models for Observational Data*. No. 59 in CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wan, R., Zhong, M., Xiong, H. & Zhu, Z. (2019). Neural control variates for variance reduction. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Wendland, H. (2004). *Scattered data approximation*, vol. 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press.

Williams, C. K. I. & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich & V. Tresp, eds., vol. 13. MIT Press.

[*Received on* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]