Control functionals for Monte Carlo integration

Chris J. Oates[1]*, Mark Girolami[1] and Nicolas Chopin[2]
[1]University of Warwick, Coventry, UK
[2]CREST-LS and ENSAE, Paris, France

October 9, 2014

Abstract

This paper introduces a novel class of estimators for Monte Carlo integration, that leverage gradient information in order to provide the improved estimator performance demanded by contemporary statistical applications. The proposed estimators, called "control functionals", achieve sub-root-n convergence and often require orders of magnitude fewer simulations, compared with existing approaches, in order to achieve a fixed level of precision. We focus on a particular sub-class of estimators that permit an elegant analytic form and study their properties, both theoretically and empirically. Results are presented on Bayes-Hermite quadrature, hierarchical Gaussian process models and non-linear ordinary differential equation models, where in each case our estimators are shown to offer state of the art performance.

Keywords: variance reduction, control variates, gradient methods, non-parametric regression

1 Introduction

1.1 Objective

Statistical methods are increasingly employed to analyse complex models of physical phenomena. Analytic intractability of such models has inspired the development of increasingly sophisticated Monte Carlo methodologies to facilitate estimation. Many of these approaches ultimately rely on Monte Carlo integration. In its most basic form, the resulting estimators converge as the reciprocal of root-n where n is the number of Monte Carlo samples. For complex models it may only be feasible to obtain a limited number of Monte Carlo samples. In these situations, root-n convergence is too slow and can lead in practice to high-variance estimation (Caflisch, 1998).

The focus of this paper is to estimate an expectation $\mu = \mathbb{E}[f(\boldsymbol{X})]$ using Monte Carlo methods, where $\boldsymbol{X}$ is a random variable and f is a real-valued function. Our work is motivated by the slow convergence of the familiar arithmetic mean estimator

$$\overline{\mu}_n(\mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \tag{1}$$

* *Address for correspondence:* Dept. Statistics, Zeeman Building, University of Warwick, Coventry, CV4 7AL, UK. E-mail: c.oates@warwick.ac.uk

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

based on n independent and identically distributed (IID) realisations $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ of the random variable. Provided that $f(\boldsymbol{X})$ has finite variance, the arithmetic mean $\overline{\mu}_n$ satisfies the central limit theorem and we say that $\overline{\mu}_n$ "converges to μ" at the rate $O_P(n^{-1/2})$, or simply at "root-n". When we are working with complex computer models, root-n convergence can be problematic as highlighted in e.g. Oakley and O'Hagan (2002). Specifically, we consider a computer model to be "complex" when either (i) $\boldsymbol{X}$ is expensive to simulate, or (ii) f is expensive to evaluate, where in each case the cost is defined relative to the required estimator precision. Both situations are prevalent in scientific and engineering applications (e.g. in climate forecasting or simulations of molecular dynamics; Slingo *et al.*, 2009; Angelikopoulos *et al.*, 2012). In this paper we introduce a class of estimators for Monte Carlo integration that converge to μ at a rate that is sub-root-n; that is, the estimator variance vanishes asymptotically with respect to the variance of $\overline{\mu}_n$.

1.2 Comparison with other variance reduction techniques

Generic approaches to reduction of variance are well-known in the literature on Monte Carlo integration. These include (i) importance sampling, (ii) stratified sampling and related techniques such as systematic sampling, (iii) antithetic variables (Green and Han, 1992) and more generally (randomised) quasi-Monte Carlo (QMC/RQMC; Niederreiter, 1978), (iv) Rao-Blackwellisation (Robert and Casella, 2004), (v) Riemann sums (Philippe, 1997), (vi) control variates (Glasserman, 2004), (vii) multi-level Monte Carlo and other application-specific techniques (e.g. Heinrich, 2001; Giles, 2008), and (viii) a plethora of sophisticated Markov chain Monte Carlo sampling schemes. Reviews of many of the above techniques can be found in Glasserman (2004) as well as Rubinstein and Kroese (2011). Fig. 1, adapted from Glasserman (2004), illustrates the trade-off between the "effectiveness" of variance-reduction techniques and their "complexity", in terms of the knowledge required to implement these techniques. Among these techniques we draw three key distinctions: (I) Firstly, Rao-Blackwellisation, Riemann sums and control variates can all be conceived as post-processing schemes that can be applied after samples have been obtained, whereas the remaining methods typically 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. (II) Secondly, within the post-processing schemes, a further distinction can be drawn between Rao-Blackwellisation on one hand and Riemann sums and control variates on the other. In the former, the form of the estimator depends upon the method used to generate samples, whereas in the latter it does not. Consequently Riemann sums and control variates offer increased freedom for the design of sampling schemes. (III) The final distinction that we draw is that Riemann sums can achieve sub-root-n convergence rates (at least when f is bounded), whereas the control variates proposed in existing literature cannot.

Despite their generality and excellent asymptotic properties, Riemann sums are rarely used in practice due to (i) the requirement that f be bounded, (ii) a significant increase in methodological complexity when $d > 1$, and perhaps most influentially (iii) the fact that estimators are *not unbiased* at finite sample sizes. Our methodology, called "control functionals", enjoys the same advantages as Riemann sums, i.e. it is a sub-root-n post-processing approach that is sampling scheme independent, but does not suffer from any of these three serious drawbacks.

Control functionals can be loosely described as a non-parametric generalisation of classical control variates. In control variate schemes one seeks statistics $U_1(\boldsymbol{X}), \ldots, U_e(\boldsymbol{X})$ that have expectation zero. Then a surrogate function $\tilde{f} = f - a_1 U_1 - \cdots - a_e U_e$ is constructed such that $\tilde{f}(\boldsymbol{X})$ has the same expectation but, for suitably chosen $a_1, \ldots, a_e \in \mathbb{R}$, has a reduced variance compared to

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

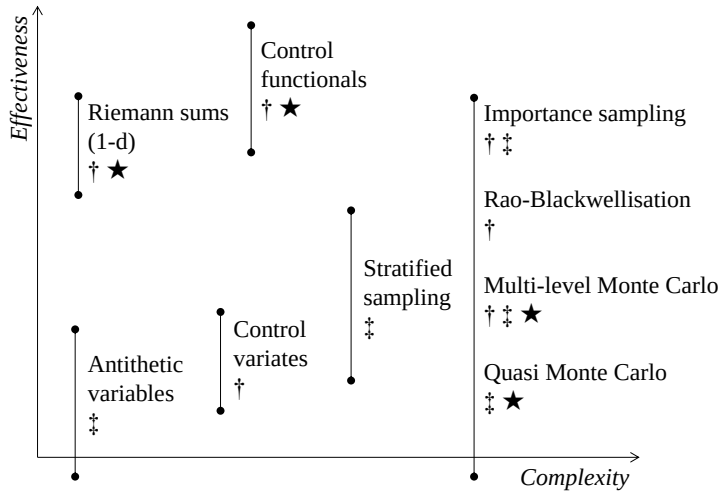


Figure 1: A subjective comparison of techniques for variance reduction in Monte Carlo. The "effectiveness" of a method is the efficiency improvement that it brings and "complexity" includes the level of effort and detailed knowledge required for implementation. [Techniques marked with a † require the modification of code for post-processing of samples, relative to the standard arithmetic mean estimator. Similarly those marked with a ‡ require modification of code used for sampling. Here $\star$ indicates estimators that have been shown to achieve sub-root-n convergence.]

$f(\boldsymbol{X})$ (see e.g. Rubinstein and Marcus, 1985). The random variables $U_i(\boldsymbol{X})$ are known as "control variates" and it can be shown that the variance of $\tilde{f}(\boldsymbol{X})$ can be reduced to zero when there is perfect canonical correlation between the set of control variates and $f(\boldsymbol{X})$. This has motivated several efforts to construct effective control variates. For estimation based on Markov chains, statistics relating to the chain itself can be used as control variates (e.g. Andradóttir *et al.*, 1993; Mira *et al.*, 2003; Hammer and Tjelmeland, 2008; Dellaportas and Kontoyiannis, 2012). Recently, a general methodology for Monte Carlo estimation was proposed by Mira *et al.* (2013) that, in the simplest case, uses the score function as a control variate. Whilst effective in particular situations, control variates can only achieve a constant factor reduction in estimator variance, so that the asymptotic convergence remains gated by root-n. We show below that the score-based control variates of Mira *et al.* (2013) are a (degenerate) case of a more powerful control functional framework that is based on gradient information and permits sub-root-n convergence.

1.3 Outline of the paper

The paper begins by showing how gradient information can be exploited to construct a surrogate function $\tilde{f}_{\mathcal{D}}$, based on samples $\mathcal{D}$, such that $\tilde{f}_{\mathcal{D}}(\boldsymbol{X})$ has expectation equal to μ and a variance that vanishes as the number n of samples tends to infinity. This surrogate function then forms the basis for the control functional estimator, which is shown to converge to μ at a rate that is sub-root-n. To realise our methodology we focus on the popular framework of linear smoothers, where we can derive analytic expressions for our estimators. Extensive empirical support is provided in favour of the proposed methodology, including applications to Bayes-Hermite quadrature (O'Hagan, 1991), hierarchical Gaussian process models and non-linear ordinary differential equation models. These

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

last two examples illustrate the combined use of control functionals with both RQMC and with gradient-based population MCMC. In each case we achieve state-of-the-art estimation performance.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from *www.warwick.ac.uk/chrisoates/control_functionals*.

2 Methods

2.1 Set-up and notation

We assume that the random variable $\boldsymbol{X}$ takes values in a space $\mathcal{X} \subseteq \mathbb{R}^d$ and admits a density $\pi(\boldsymbol{x})$ with respect to d-dimensional Lebesgue measure. In addition we suppose that the associated score function $\boldsymbol{u}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x})$ exists and can, in principle, be evaluated at any given point $\boldsymbol{x} \in \mathcal{X}$. Here $\nabla_{\boldsymbol{x}} := [\partial/\partial x_1, \ldots, \partial/\partial x_d]^T$. The function $f : \mathcal{X} \to \mathbb{R}$ is assumed to be well-defined and can, in principle, be evaluated at any given point $\boldsymbol{x} \in \mathcal{X}$. Write $\mathcal{L}^2(\pi) = \{g : \mathcal{X} \to \mathbb{R} : \int |g|^2 \pi(d\boldsymbol{x}) < \infty\}$.

The samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$, the corresponding function values $\{f_i\}_{i=1}^n$ and the corresponding scores $\{\boldsymbol{u}_i\}_{i=1}^n$ are assumed to have been pre-computed and cached. The methodology that we develop below does not then require any additional recourse to the statistical model π, nor any further evaluations of the function f, and is in this sense a widely-applicable post-processing scheme.

2.2 Control functionals

The score statistic has previously been used as a control variate (see e.g. Philippe and Robert, 2001). More recently Mira *et al.* (2013) constructed score-based control variates, using expectation-preserving transforms of the score that were based on polynomials. This paper goes further, by taking the innovative step of considering non-parametric expectation-preserving transformations of the score statistic. Specifically, this paper studies the class of surrogate functions $\tilde{f}_{\boldsymbol{\phi}} : \mathcal{X} \to \mathbb{R}$ of the form

$$\tilde{f}_{\boldsymbol{\phi}}(\boldsymbol{x}) := f(\boldsymbol{x}) - \psi_{\boldsymbol{\phi}}(\boldsymbol{x}) \tag{2}$$

$$\psi_{\boldsymbol{\phi}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}) \tag{3}$$

where $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}$ is a differentiable function to be specified. This class is motivated by the observation that if

$$[\pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})]_{\boldsymbol{x} \in \partial\mathcal{X}} = \boldsymbol{0} \tag{4}$$

then it follows from integration by parts that $\mathbb{E}[\psi_{\boldsymbol{\phi}}(\boldsymbol{X})] = 0$ and so the function $\tilde{f}_{\boldsymbol{\phi}}(\boldsymbol{X})$ shares the same expectation as $f(\boldsymbol{X})$. Here $\partial\mathcal{X}$ denotes the boundary of the state space $\mathcal{X}$; when the state space is unbounded we interpret this condition (via the divergence theorem; see e.g. Kendall and Bourne, 1992) as $\int_{S_r \cap \mathcal{X}} \pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) d\boldsymbol{x} \to 0$ where S_r is the sphere of radius r centred at the origin and $\boldsymbol{n}(\boldsymbol{x})$ is the unit normal to the surface of S_r. (Note that the expression for $\tilde{f}_{\boldsymbol{\phi}}(\boldsymbol{x})$ above is more general than in Mira *et al.* (2013). In their paper, $\boldsymbol{\phi}$ is the gradient of a certain function, which imposes constraints on the derivatives of $\boldsymbol{\phi}$, but the reasoning above shows that these restrictions are not necessary.)

In classical literature, $\psi_{\boldsymbol{\phi}}(\boldsymbol{X})$ would be recognised as a control variate, i.e. a random variable with expectation zero. The approach that we pursue here is fundamentally different in that we treat $\psi_{\boldsymbol{\phi}}$ as a functional $\mathcal{X} \to \mathbb{R}$ that is itself to be estimated. To acknowledge this distinction

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

we will refer to $\psi_{\boldsymbol{\phi}}$ as a "control functional". Our aim is to specify $\boldsymbol{\phi}$ (and hence the control functional $\psi_{\boldsymbol{\phi}}$) in such a way that the corresponding surrogate function $\tilde{f}_{\boldsymbol{\phi}}$ will minimise the variance $\mathbb{V}[\tilde{f}_{\boldsymbol{\phi}}(\boldsymbol{X})] = \int (\tilde{f}_{\boldsymbol{\phi}} - \mu)^2 \pi(d\boldsymbol{x})$. This corresponds exactly to fitting a functional regression of the form

$$f(\boldsymbol{x}) = \mu + \psi_{\boldsymbol{\phi}}(\boldsymbol{x}) + \epsilon_{\boldsymbol{\phi}}(\boldsymbol{x}) \tag{5}$$

where we aim to minimise the mean square error $\int \epsilon_{\boldsymbol{\phi}}^2 \pi(d\boldsymbol{x})$. A connection between classical control variates and linear regression is well known (e.g. Rubinstein and Marcus, 1985). The connection in Eqn. 5, between control functionals and functional regression, offers more general insight and the opportunity to exploit established techniques from non-parametric statistics. We show below how this generalisation can enable dramatic variance reductions at finite sample sizes compared to existing methodology, as well as offer superior asymptotic rates of convergence.

2.3 Convergence at sub-root-n

The proposed approach begins by splitting samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ 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 the size of both subsets is assumed to increase linearly as $n \to \infty$. The first subset $\mathcal{D}_0$ is used to estimate a surrogate function $\tilde{f}_{\mathcal{D}_0} : \mathcal{X} \to \mathbb{R}$, such that $\tilde{f}_{\mathcal{D}_0}(\boldsymbol{X})$ shares the same expectation as $f(\boldsymbol{X})$ but has variance that vanishes as $m \to \infty$. Then the second subset $\mathcal{D}_1$ is used to evaluate an arithmetic mean based on the values $\tilde{f}_{\mathcal{D}_0}(\boldsymbol{x}_i)$ where $i = m+1, \ldots, n$. This dichotomy of the samples is required to ensure that the estimator we construct below is unbiased. For simplicity we continue to assume that the samples $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ arose independently from π. This can be relaxed as we describe later, so that $\mathcal{D}_0$ can instead be IID samples from some other distribution π_0, with only $\mathcal{D}_1$ arising from π.

To begin, notice that for a given function $\boldsymbol{\phi}$ we can estimate μ using the arithmetic mean

$$\hat{\mu}_{\boldsymbol{\phi}}(\mathcal{D}_1) := \frac{1}{n-m} \sum_{i=m+1}^{n} \tilde{f}_{\boldsymbol{\phi}}(\boldsymbol{x}_i). \tag{6}$$

Here unbiasedness, i.e. $\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}_{\boldsymbol{\phi}}(\mathcal{D}_1)] = \mu$, is an immediate consequence of $\mathbb{E}[\tilde{f}_{\boldsymbol{\phi}}(\boldsymbol{X})] = \mu$, where the first expectation here is with respect to the $n-m$ random variables that constitute $\mathcal{D}_1$. The insight required to achieve sub-root-n convergence is that we can leverage $\mathcal{D}_0$ to estimate an "optimal" function $\boldsymbol{\phi}$ based on the functional regression in Eqn. 5.

An important observation is that we are not restricted to single point estimates for $\boldsymbol{\phi}$. Indeed, consider application of a $\mathcal{D}_0$-dependent linear operator $\mathcal{E}_{\mathcal{D}_0}$ that acts on $\boldsymbol{\phi}$ as follows:

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) := \mathcal{E}_{\mathcal{D}_0}[\hat{\mu}_{\boldsymbol{\phi}}(\mathcal{D}_1)]. \tag{7}$$

This corresponds to an arithmetic mean estimator based on the surrogate function $\tilde{f}_{\mathcal{D}_0} = \mathcal{E}_{\mathcal{D}_0}[\tilde{f}_{\boldsymbol{\phi}}]$. The resulting estimator $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$ will be unbiased since

$$\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)] = \mathbb{E}_{\mathcal{D}_1}[\mathcal{E}_{\mathcal{D}_0}[\hat{\mu}_{\boldsymbol{\phi}}(\mathcal{D}_1)]] = \mathcal{E}_{\mathcal{D}_0}[\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}_{\boldsymbol{\phi}}(\mathcal{D}_1)]] = \mathcal{E}_{\mathcal{D}_0}[\mu] = \mu, \tag{8}$$

where the interchange of operators is justified by linearity and the independence of samples $\mathcal{D}_0$ and $\mathcal{D}_1$. Candidate choices for the operator $\mathcal{E}_{\mathcal{D}_0}$ determine the final form of our control functional estimator and a variety of choices are discussed in the next section.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

In this more general setting, define the expected predictive square error

$$\sigma^2_{\mathcal{D}_0} := \int (\mathcal{E}_{\mathcal{D}_0}[\epsilon_{\boldsymbol{\phi}}])^2 \pi(d\boldsymbol{x}), \tag{9}$$

that is, the expected square error in our approximation $\tilde{f}_{\mathcal{D}_0}(\boldsymbol{X})$ to μ when $\boldsymbol{X}$ is generated from π. We say the operator $\mathcal{E}_{\mathcal{D}_0}$ is "prediction consistent" (or "presistent"; Wasserman, 2013) whenever the average predictive square error vanishes as $m \to \infty$; i.e. whenever $\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}] \to 0$ as $m \to \infty$.

Theorem 1. *Suppose that $\mathcal{E}_{\mathcal{D}_0}$ is a presistent linear operator and that the boundary condition $[\pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})]_{\boldsymbol{x}\in\partial\mathcal{X}} = \mathbf{0}$ holds almost surely. Then*

$$\frac{\mathbb{V}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)]}{\mathbb{V}_{\mathcal{D}}[\overline{\mu}_n(\mathcal{D})]} \to 0 \quad as \quad n \to \infty, \tag{10}$$

where $\overline{\mu}_n(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)$ is the usual arithmetic mean, so that we have sub-root-n convergence of the control functional estimator.

All proofs are reserved for Appendix A. We note that presistency can hold irrespective of the requirement that $\pi_0 = \pi$. Indeed, the proof of Theorem 1 requires that $\mathcal{D}_0$ and $\mathcal{D}_1$ are independent but does not require that $\mathcal{D}_0$ arise from π.

In the next section we demonstrate how previous literature on score-based control variates can be cast as particular (degenerate) choices of the linear operator $\mathcal{E}_{\mathcal{D}_0}$. We then show how this new perspective enables the construction of more efficient approaches to Monte Carlo integration.

2.4 Choosing a linear operator $\mathcal{E}_{\mathcal{D}_0}$

2.4.1 An optimal choice

From the functional regression problem (Eqn. 5) we see that an optimal function $\boldsymbol{\phi}$ is the solution of the first order linear partial differential equation (PDE) obtained by setting $\epsilon_{\boldsymbol{\phi}}(\boldsymbol{x}) = 0$, namely

$$\nabla_{\boldsymbol{x}} \cdot [\pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})] = [f(\boldsymbol{x}) - \mu]\pi(\boldsymbol{x}). \tag{11}$$

Write $\Phi^*(f, \pi)$ for the set of all solutions to Eqn. 11, so that for any function $\boldsymbol{\phi}^* \in \Phi^*(f, \pi)$, the estimator $\hat{\mu}_{\boldsymbol{\phi}^*}(\mathcal{D}_1) = \mu$ is exact and so has exactly zero variance. An optimal choice of the linear operator $\mathcal{E}_{\mathcal{D}_0}$ would therefore be a Dirac delta (or "atom") $\mathcal{E}_{\mathcal{D}_0} = \delta_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}$ on one such function, if such a function $\boldsymbol{\phi}^*$ exists. When π is everywhere positive, it is directly verified that the solution space for Eqn. 11 contains (but is not limited to) the family of functions $\boldsymbol{\phi}^*$ of the form

$$\phi^*_i(\boldsymbol{x}) = \frac{1}{d\pi(\boldsymbol{x})}\left[\int_{-\infty}^{x_i} \begin{array}{r}[f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_d) - \mu] \\ \times\pi(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_d)dx'_i + c_i\end{array}\right] \tag{12}$$

indexed by arbitrary constants $c_1, \ldots, c_d \in \mathbb{R}$. Thus in this case we have $\Phi^*(f, \pi) \neq \emptyset$. Note that whilst $\Phi^*(f, \pi)$ is infinite, each member will produce the same control functional (i.e. $\boldsymbol{\phi}, \boldsymbol{\phi}' \in \Phi^*(f, \pi) \implies \psi_{\boldsymbol{\phi}} = \psi_{\boldsymbol{\phi}'}$ a.s.).

Unfortunately Eqn. 12 does not provide a direct strategy to compute an optimal function $\boldsymbol{\phi}^* \in \Phi^*(f, \pi)$, since it depends on the unknown of interest μ (as indeed would any solution of Eqn. 11). In practice an optimal function can instead be estimated using samples $\mathcal{D}_0$. This is the strategy that we will develop in this paper.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

2.4.2 Parametric point estimates

The use of the score statistic as a control variate corresponds to the simplest possible case where $\mathcal{E}_{\mathcal{D}_0}$ is an atom on a constant function $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{c}$ for some $\boldsymbol{c} \in \mathbb{R}^d$ and for all $\boldsymbol{x} \in \mathcal{X}$. More generally $\mathcal{E}_{\mathcal{D}_0}$ could be an atom on a parametric function $\boldsymbol{\phi}_{\boldsymbol{\theta}}(x)$, indexed by a parameter $\boldsymbol{\theta} \in \Theta$ so that the set of possible choices for $\boldsymbol{\phi}$ is $\Phi = \{\boldsymbol{\phi}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. This is essentially the approach considered by Mira *et al.* (2013), called "Zero Variance" (ZV) control variates. That work focused on polynomial functions, in particular the first order scheme $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}$ and the second order scheme $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2^T \boldsymbol{x}$ with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The maximum polynomial degree is fixed in advance and this defines the set Φ. Setting this procedure within our general framework, we see that sub-root-n convergence could be possible if both (i) $\Phi \cap \Phi^*(f, \pi) \neq \emptyset$, so that $\boldsymbol{\phi}_{\boldsymbol{\theta}^*} \in \Phi(f, \pi)$ for some $\boldsymbol{\theta}^* \in \Theta$ and (ii) it is possible to construct a consistent estimator $\hat{\boldsymbol{\theta}}(\mathcal{D}_0) \to \boldsymbol{\theta}^*$ as $m \to \infty$. For ZV control variates this can occur in exceptional circumstances; indeed it was shown in Papamarkou *et al.* (2014b) that second-order ZV estimators are essentially exact when estimating the first and second moments of a Gaussian distribution (hence the "zero variance" nomenclature). However in most applications the sets Φ and $\Phi^*(f, \pi)$ will be disjoint and convergence will generally be limited to root-n.

2.4.3 Non-parametric point estimates

Our control functional framework suggests that a candidate function $\boldsymbol{\phi}$ can be estimated using techniques from functional regression. Specifically, given samples $\mathcal{D}_0$, corresponding function values $f_i = f(\boldsymbol{x}_i)$ and score values $\boldsymbol{u}_i = \boldsymbol{u}(\boldsymbol{x}_i)$, $i = 1, \ldots, m$, we fit the functional regression model

$$f_i = \mu + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{u}_i + \epsilon_i. \tag{13}$$

Fitting here requires that an estimate $\hat{\boldsymbol{\phi}}(\mathcal{D}_0)$ is selected from a pre-defined set Φ with the goal of minimising the average prediction error over all functions in Φ. However Φ need not be a parametric family, for example Φ could be the closure of the family of polynomials, producing a non-parametric generalisation of Mira *et al.* (2013) that does not require the maximum polynomial degree to be restricted. The linear operator $\mathcal{E}_{\mathcal{D}_0}$ is then taken to be an atom on this estimated function. As before, sub-root-n convergence could be possible if both (i) $\boldsymbol{\phi}^* \in \Phi^*(f, \pi)$ for some $\boldsymbol{\phi}^* \in \Phi$ and (ii) the estimator $\hat{\boldsymbol{\phi}}(\mathcal{D}_0)$ converges to this function $\hat{\boldsymbol{\phi}}(\mathcal{D}_0) \to \boldsymbol{\phi}^*$ (in an appropriate sense) as $m \to \infty$. In principle any functional regression methodology could be used, but care must be taken when using basis functions that are not differentiable. This perspective offers the possibility to design powerful estimators based on established non-parametric regression techniques; see Appendix B for such an approach.

2.4.4 Bayesian posterior expectation

The above approaches all take the linear operator $\mathcal{E}_{\mathcal{D}_0}$ to be an atom. More broadly, a particularly natural choice for $\mathcal{E}_{\mathcal{D}_0}$ would be a Bayesian posterior expectation $\mathcal{E}_{\mathcal{D}_0}[\cdot] = \mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\cdot]$. This approach places a prior $p(\boldsymbol{\phi})$ over all possible functions $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$ and then computes the posterior distribution $p(\boldsymbol{\phi}|\mathcal{D}_0)$ according to Bayes' rule. The result is an estimator

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) = \mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\hat{\mu}_{\boldsymbol{\phi}}(\mathcal{D}_1)] = \frac{1}{n-m} \mathbf{1}_{1 \times n-m} \left(\boldsymbol{f}_1 - \mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\boldsymbol{\psi}_1] \right) \tag{14}$$

where $\boldsymbol{f}_1 = [f_{m+1}, \ldots, f_n]^T$, $\boldsymbol{\psi}_1 = [\psi_{\boldsymbol{\phi}}(\boldsymbol{x}_{m+1}), \ldots, \psi_{\boldsymbol{\phi}}(\boldsymbol{x}_n)]^T$ and $\mathbf{1}_{1 \times n-m}$ is a $1 \times n - m$ vector of ones. This Bayesian approach can avoid difficulties pertaining to non-differentiability of basis

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

functions through the specification of a prior $p(\boldsymbol{\phi})$ that has support contained in the space of differentiable functions. Rather excitingly, this approach combines with linear smoothers to lead to elegant analytic expressions for $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$. This is the direction that we pursue below. We emphasise that the estimators we derive in this way are justified from an entirely frequentist perspective.

2.5 Gaussian process control functionals

To proceed with the approach outlined in Sec. 2.4.4 we require that a prior distribution is specified for the function $\boldsymbol{\phi}$. A theoretically and practically convenient choice is a Gaussian probability measure (i.e. a Borel probability measure whose finite-dimensional marginals are all Gaussian in distribution). This approach has been explored theoretically (e.g. Stuart, 2010) and has received widespread usage in computer experiment emulation (Oakley and O'Hagan, 2002) and machine learning applications (e.g. Rasmussen and Williams, 2006).

2.5.1 An analytic expression for $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$

A Gaussian process (GP) is a linear smoother (Hastie and Tibshirani, 1990) whose conjugacy properties enable the derivation of an analytic expression for the control functional estimator $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$. For each component function $\phi_i : \mathcal{X} \to \mathbb{R}$ we assume an independent GP prior

$$\phi_i(\boldsymbol{x}) \sim \mathcal{GP}(0, k_0(\boldsymbol{x}, \boldsymbol{x}')), \quad i = 1, \ldots, d, \tag{15}$$

where $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a twice differentiable covariance function that must be specified (see Rasmussen and Williams (2006) for further details on GP priors).

Lemma 1. *Under Eqn. 15 the control functional $\psi_{\boldsymbol{\phi}}$ is also a GP, satisfying*

$$\psi_{\boldsymbol{\phi}}(\boldsymbol{x}) \sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}')) \tag{16}$$

with $k(\boldsymbol{x}, \boldsymbol{x}') = \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} k_0(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{u}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}'} k_0(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{u}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} k_0(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{u}(\boldsymbol{x}') k_0(\boldsymbol{x}, \boldsymbol{x}')$. Moreover, taking a uniform prior $p(\mu) \propto 1$ produces

$$\mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\psi_1] = \boldsymbol{K}_{1,0} \left[\boldsymbol{I}_{m \times m} - \frac{\boldsymbol{K}_0^{-1} \mathbf{1}_{m \times 1} \mathbf{1}_{1 \times m}}{\mathbf{1}_{1 \times m} \boldsymbol{K}_0^{-1} \mathbf{1}_{m \times 1}} \right] \boldsymbol{K}_0^{-1} \boldsymbol{f}_0 \tag{17}$$

where $(\boldsymbol{K}_0)_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $(\boldsymbol{K}_{1,0})_{i,j} = k(\boldsymbol{x}_{m+i}, \boldsymbol{x}_j)$ and $\boldsymbol{f}_0 = [f_1, \ldots, f_m]^T$. Here the expectation is with respect to the posterior marginal $p(\boldsymbol{\phi}|\mathcal{D}_0) = \int p(\boldsymbol{\phi}, \mu|\mathcal{D}_0) d\mu$.

Substituting Eqn. 17 into Eqn. 14 we obtain the (unbiased) GP control functional (GPCF) estimator

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) = \frac{1}{n-m} \mathbf{1}_{1 \times n-m} \left\{ \boldsymbol{f}_1 - \boldsymbol{K}_{1,0} \left[\boldsymbol{I}_{m \times m} - \frac{\boldsymbol{K}_0^{-1} \mathbf{1}_{m \times 1} \mathbf{1}_{1 \times m}}{\mathbf{1}_{1 \times m} \boldsymbol{K}_0^{-1} \mathbf{1}_{m \times 1}} \right] \boldsymbol{K}_0^{-1} \boldsymbol{f}_0 \right\} \tag{18}$$

which, since GPs are linear smoothers, is simply a weighted combination of function values $\boldsymbol{f} = [\boldsymbol{f}_0^T, \boldsymbol{f}_1^T]^T$. Thus our proposed GPCF estimator is readily obtained using standard matrix algebra. Moreover the weights are independent of the function f and can be re-used to estimate multiple expectations $\mu_j = \mathbb{E}[f_j(\boldsymbol{X})]$. Note that the arithmetic mean estimator is related to Eqn. 24 by taking $\boldsymbol{K}_0$ to be the identity matrix and both $\boldsymbol{K}_{0,1}$, $\boldsymbol{K}_{1,0}$ to be matrices of zeros. This would

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

occur when the GP prior k_0 is a Dirac delta function, so that the function values f_i are treated as independent conditional upon the $\boldsymbol{x}_i$.

To gain intuition for the GPCF estimator in Eqn. 18, we rewrite the expression as

$$\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) = \underbrace{\frac{1}{n-m}\mathbf{1}_{1\times n-m}(\boldsymbol{f}_1 - \hat{\boldsymbol{f}}_1)}_{(*)} + \frac{\mathbf{1}_{1\times m}\boldsymbol{K}_0^{-1}\boldsymbol{f}_0}{\mathbf{1}_{1\times m}\boldsymbol{K}_0^{-1}\mathbf{1}_{m\times 1}} \tag{19}$$

where

$$\hat{\boldsymbol{f}}_1 = \boldsymbol{K}_{1,0}\boldsymbol{K}_0^{-1}\boldsymbol{f}_0 + \frac{\boldsymbol{R}^T\mathbf{1}_{1\times m}\boldsymbol{K}_0^{-1}\boldsymbol{f}_0}{\mathbf{1}_{1\times m}\boldsymbol{K}_0^{-1}\mathbf{1}_{m\times 1}} \tag{20}$$

is the posterior mean of $\boldsymbol{f}_1$, $\boldsymbol{R} = \mathbf{1}_{1\times n-m} - \mathbf{1}_{1\times m}\boldsymbol{K}_0^{-1}\boldsymbol{K}_{0,1}$ and $\boldsymbol{K}_{0,1} = \boldsymbol{K}_{1,0}^T$. Notice that the samples $\mathcal{D}_1$ enter only through the term $(*)$ in Eqn. 19. Presistency of the functional regression implies that $(*)$ vanishes almost surely as $m \to \infty$. In the GP literature, $\sigma^2_{\mathcal{D}_0}$ from Eqn. 9 is known as the "expected generalisation error" (EGE; Williams and Vivarelli, 2000). Presistency for GPs is therefore equivalent to requiring that the EGE vanishes almost surely as $m \to \infty$. This effectively removes any randomness in $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$ that is due to $\mathcal{D}_1$ and explains the source of sub-root-n convergence of the estimator. We will see in Section 3.2 that the remaining term in Eqn. 19 can be interpreted as the posterior mean of the intercept parameter μ in the functional regression (Eqn. 13) when inference is based on $\mathcal{D}_0$ and an improper prior $p(\mu) \propto 1$ is employed. Note that it is possible to use this second term in isolation as a (biased but) consistent estimator for μ; we return to this point shortly.

The computational complexity of this GP implementation is $O(m^3)$ due to the inversion of an $m \times m$ covariance matrix. In situations where either π is expensive to sample or f is expensive to evaluate, such as in complex computer models, then m will not be prohibitively large and this additional computation will be negligible with respect to simulation from the model. Nevertheless, in situations where m may also be large, approximation schemes for efficient (but biased) computation in GPs may be directly used here (Quiñonero-Candela *et al.*, 2007).

2.5.2 Theoretical considerations

We now state regularity assumptions that together imply sub-root-n convergence of the GPCF estimator.

(A1) $f \in \mathcal{L}^2(\pi)$.

(A2) u_j exists and $u_j \in \mathcal{L}^2(\pi)$ for each $j = 1, \ldots, d$.

Assumption 2 is satisfied whenever the Fisher information matrix $-\mathbb{E}_{\boldsymbol{X}}[\nabla_{x_i}\nabla_{x_j}\log\pi(\boldsymbol{X})]$ exists and the diagonal elements are finite. This is therefore an extremely weak requirement.

(A3) The first and second-order partial derivatives of k_0 exist.

(A4) Both k_0 and its partial derivatives are bounded on $\mathcal{X} \times \mathcal{X}$.

Assumptions 3 and 4 can be satisfied by construction with an appropriate choice for the covariance function k_0. The selection of a covariance function is discussed in the next section.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

(A5) The boundary condition $[\pi(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})]_{\boldsymbol{x}\in\partial\mathcal{X}} = \mathbf{0}$ holds almost surely.

Assumption 5 can typically be satisfied by construction by re-parametrising, via a suitable diffeomorphism, so that π vanishes on the boundary $\partial\mathcal{X}$. (e.g. If $X \sim \text{Exp}(\lambda)$ then X has a density π_X on $(0,\infty)$ that satisfies $\pi_X(0) = \lambda > 0$ whereas $Y = \log X$ has a density π_Y on $(-\infty,\infty)$ that satisfies $\pi_Y(y) \to 0$ as $|y| \to \infty$).

(A6) The linear operator $\mathcal{E}_{\mathcal{D}_0}$ is presistent.

Theorem 2. *Assume (A1)-(A5). Then for any fixed $\mathcal{D}_0$ we have $\tilde{f}_{\mathcal{D}_0} \in \mathcal{L}^2(\pi)$, so that $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$ has finite variance when $\mathcal{D}_1$ arise from π. If we additionally assume (A6) then the estimator $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$ will converge to its expectation μ at a rate that is sub-root-n.*

Theoretical results on presistency for GPs (and more general linear smoothers) have focussed on compact state spaces $\mathcal{X}$, densities π that are bounded away from zero and particular forms of the covariance function k for which concentration inequalities are available (see e.g. the recent work of van der Vaart and van Zanten, 2011). Unfortunately this is not our setting and a more general result in the theory of GP regression seems elusive at present, since convergence rates will depend on the tail behaviour of π which may be complicated (Aad van der Vaart, personal communication). Indeed, even the case where $\mathcal{D}_0$ is non-random poses significant theoretical challenges (see also Williams and Vivarelli, 2000). For these reasons, here we state a general presistency condition that in future may be shown to be satisfied under additional regularity assumptions on f, π, $\boldsymbol{u}$ and k_0. In Appendix B we argue that presistency is a reasonable assumption that can often be met in practice. Moreover, it is straightforward to empirically diagnose a violation of presistency, as we show in the next section.

2.5.3 Choosing a covariance function

The GPCF methodology relies on eliciting an appropriate GP covariance function k_0. Intuitively, for sufficiently regular f, π, we want to choose k_0 such that an element of $\Phi^*(f,\pi)$ can be well-approximated by functions in the reproducing kernel Hilbert space (RKHS) Φ of k_0, particularly in the high-probability regions for π. A natural approach would be to induce a covariance function k_0 that is consistent with solutions of the PDE in Eqn. 11. For simplicity consider the one-dimensional case ($d = 1$). Following recent work in this direction by Lawrence *et al.* (2007); Wheeler *et al.* (2014), we rewrite Eqn. 12 as

$$\phi^*(x) = \int G(x,x')[f(x') - \mu]dx' \tag{21}$$

where the Green's function $G(x,x') = \pi(x')/\pi(x)$ for $x' < x$ and $G(x,x') = 0$ otherwise is simply a ratio of densities. An initial GP prior $f - \mu \sim \mathcal{GP}(m_f, k_f)$ can then be transformed into a secondary GP prior $\phi^* \sim \mathcal{GP}(m_0, k_0)$ by exploiting linearity of the integral in Eqn. 21, so that $m_0(x) = \int G(x,x')m_f(x')dx'$ and $k_0(x,x') = \int\int G(x,z)G(x',z')k_f(z,z')dzdz'$. However, whilst this approach provides theoretical insight, its practical utility is limited since in most settings these integrals will not posses an analytic form.

In this paper, to limit scope, we restrict attention to a single default choice; the squared-error covariance $k_0(\boldsymbol{x},\boldsymbol{x}') = \exp(-\frac{1}{2\ell^2}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2)$ with a single length-scale hyper-parameter ℓ. This covariance function has five important properties that make it well-suited to our purposes: (i) The

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

partial derivatives of k_0 exist. (ii) Both k_0 and its partial derivatives are bounded on $\mathcal{X} \times \mathcal{X}$. (iii) The prior support is contained in the space of differentiable functions (up to a null set; see e.g. Rasmussen and Williams, 2006, for further details). (iv) The boundary condition in Eqn. 4 is satisfied with probability one when $\mathcal{X} = \mathbb{R}^d$ (since π vanishes in the tails and a sample $\boldsymbol{\phi}$ from the GP is bounded with probability one). (v) There is only one hyper-parameter (ℓ) that must be specified. Note, however, that the squared-error covariance function can lead to slow rates of posterior concentration when the true process is not smooth. In this case we would need to consider alternative covariance functions, such as the Matérn class, that produce sample paths with restricted differentiability properties.

2.5.4 Choosing the covariance hyper-parameter(s)

For an arbitrary covariance function $k_0 \equiv k_0(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$, denote by $\boldsymbol{\theta}$ any hyper-parameters required for its specification. Elicitation of $\boldsymbol{\theta}$ and empirical assessment of presistency can proceed under the additional assumption that $\mathcal{D}_0$ arose as independent samples from π (i.e. $\pi_0 = \pi$). Specifically, we randomly split the samples $\mathcal{D}_0$ into r training samples $\mathcal{D}_{0,0}$ and $m - r$ test samples $\mathcal{D}_{0,1}$. Then we propose to select $\boldsymbol{\theta}$ in order to minimise $\|\boldsymbol{f}_{(0,1)} - \hat{\boldsymbol{f}}_{(0,1)}\|_2$ where $\boldsymbol{f}_{(0,1)}$ is a vector of values $f(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{D}_{0,1}$, and $\hat{\boldsymbol{f}}_{(0,1)} = \mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_{(0,0)}}[\boldsymbol{f}_{(0,1)}]$ is the corresponding posterior predictive mean. In this way we are targeting the predictive square error $\sigma^2_{\mathcal{D}_{0,0}} := \int (\mathcal{E}_{\mathcal{D}_{0,0}}[\epsilon_{\boldsymbol{\phi}}])^2 \pi(d\boldsymbol{x})$ that reflects the variance $\sigma^2_{\mathcal{D}_0}$ of our Monte Carlo estimator. Moreover the full posterior predictive distribution can be used to diagnose whether the presistency assumption is violated. Thus all decisions relating to the control functional framework, including whether or not to use it at all, can be made on the basis of $\mathcal{D}_0$-dependant cross validation. We emphasise that the cross-validated estimator will remain unbiased (in the sense of Eqn. 8) when cross-validation is performed only using $\mathcal{D}_0$.

2.5.5 Sample-splitting and a simplified estimator

In the above exposition we selected a fixed split S of the samples $\mathcal{D}$ into subsets $\mathcal{D}_0$ and $\mathcal{D}_1$. The relative size of $\mathcal{D}_0$ and $\mathcal{D}_1$ should be chosen in order to minimise the variance of the control functional estimator $\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)$.

Lemma 2. *Suppose the operator $\mathcal{E}_{\mathcal{D}_0}$ is presistent at a rate $\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}] = O(m^{-\gamma})$ for some $\gamma > 0$. Then the variance $\mathbb{V}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)]$ of the control functional estimator is minimised by taking a split of size $\frac{m}{n} \approx \frac{\gamma}{1+\gamma}$.*

Thus root-n presistency in the functional regression (i.e. $\gamma = 1$) implies that we should select an approximately equal partition of the samples ($m \approx n/2$). If presistency occurs more slowly than root-n then we should favour $m < n/2$, which in the extreme case simply recovers the arithmetic mean estimator. (For the proof-of-principle estimator in Appendix B we have $\frac{m}{n} \approx \frac{2\beta}{4\beta+1}$.)

The selection of the split S itself should be independent of $\mathcal{D}$. Randomness in this process will introduce additional estimator variance that is clearly undesirable. We therefore propose to marginalise over several different splits S of size m, thereby obtaining an estimator with reduced variance. Recent work by Meinshausen and Bühlmann (2010) termed this procedure "multi-splitting". We note that a multi-splitting estimator remains unbiased. In practice it will not be possible to explore all ${}_nC_m$ partitions of the samples, but only a random subset of these partitions.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

As an alternative to multi-splitting, for applications where consistency suffices and unbiased estimation is not essential, we also propose a "simplified" estimator

$$\hat{\mu}(\mathcal{D}) := \frac{\mathbf{1}_{1\times n}\boldsymbol{K}^{-1}\boldsymbol{f}}{\mathbf{1}_{1\times n}\boldsymbol{K}^{-1}\mathbf{1}_{n\times 1}} \tag{22}$$

that assumes $m = n$, that is, assumes all samples $\mathcal{D}$ are used to form $\mathcal{D}_0$ and the second set $\mathcal{D}_1$ is empty. (For clarity we have dropped the subscript zeros in Eqn. 22). Since ${}_nC_n = 1$, this conveniently avoids the need to perform multi-splitting, at the cost of introducing bias into the estimator. Empirical results below show that this bias is typically negligible and, to pre-empt our conclusions, we recommend this simplified estimator for use in applications where unbiased estimation is not essential, due to its reduced variance. To gain intuition for Eqn. 22, notice that it can also be derived by setting $\mathcal{D}_0 = \mathcal{D}_1$ in Eqn. 19. Thus we are in effect "predicting values that we have already seen" and this leads to increased stability.

3 Applications

The methodology is initially illustrated using simple tractable examples. We then present three sophisticated examples that, together, encompass many of the challenges associated with contemporary applications, e.g. complex computer models. Firstly we consider the topical field of uncertainty quantification and, specifically, the method of Bayes-Hermite quadrature for Monte Carlo integration. Here we show how the control functional methodology elegantly solves three problems that are well-known in this literature. Secondly we consider marginalisation of hyper-parameters in hierarchical models, focussing on a 21-dimensional GP prediction problem. Here evaluation of f forms a computational bottleneck due to the required inversion of a large covariance matrix. For this problem, control functionals are shown to offer significant computational savings. Moreover we show how control functionals can be combined with another variance reduction scheme (RQMC) to deliver enhanced performance. Thirdly we consider computation of normalising constants for models based on non-linear ordinary differential equations (ODEs). Here evaluation of the likelihood function requires numerical integration of a system of ODEs and dominates computational expenditure in both sampling from π and evaluation of f. We show how control functionals 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 Analytically tractable case studies

To illustrate the methodology we consider the toy problem of estimating the expectation of the sine function

$$f(\boldsymbol{X}) = \sin\left(\frac{\pi}{d}\sum_{i=1}^{d} X_i\right) \tag{23}$$

evaluated on a d-dimensional standard Gaussian random variable $\boldsymbol{X}$, based on $n = 50$ independent samples from π. By symmetry the true expectation is $\mu = 0$. Initially we consider the scalar case ($d = 1$). For this example one can check that (i) the score function is $u(x) = -x$, (ii) the GP covariance is $k(x, x') = a(x, x')k_0(x, x')$ where $a(x, x') = \ell^{-2} + xx' - (\ell^{-2} + \ell^{-4})(x - x')^2$, and (iii) the assumptions (A1)-(A5) are satisfied, so that $\tilde{f}_{\mathcal{D}_0} \in \mathcal{L}^2(\pi)$. The problem is therefore well-posed.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

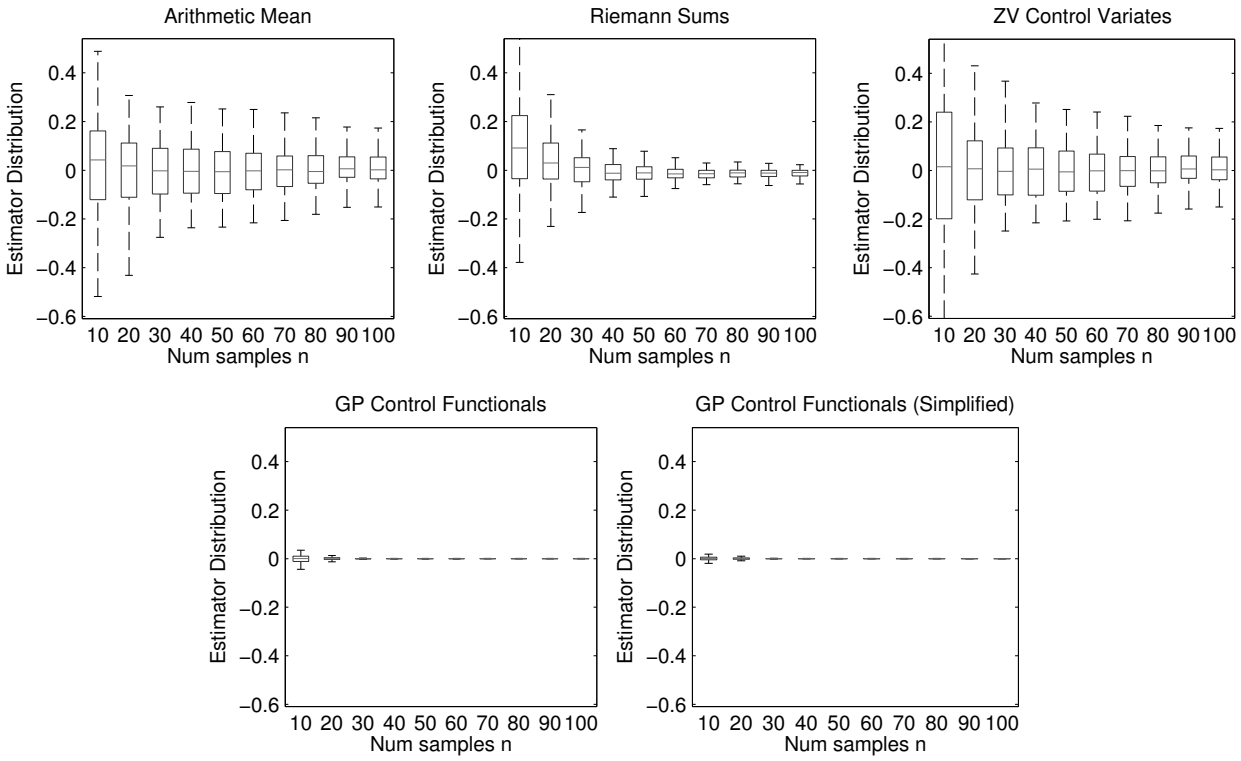


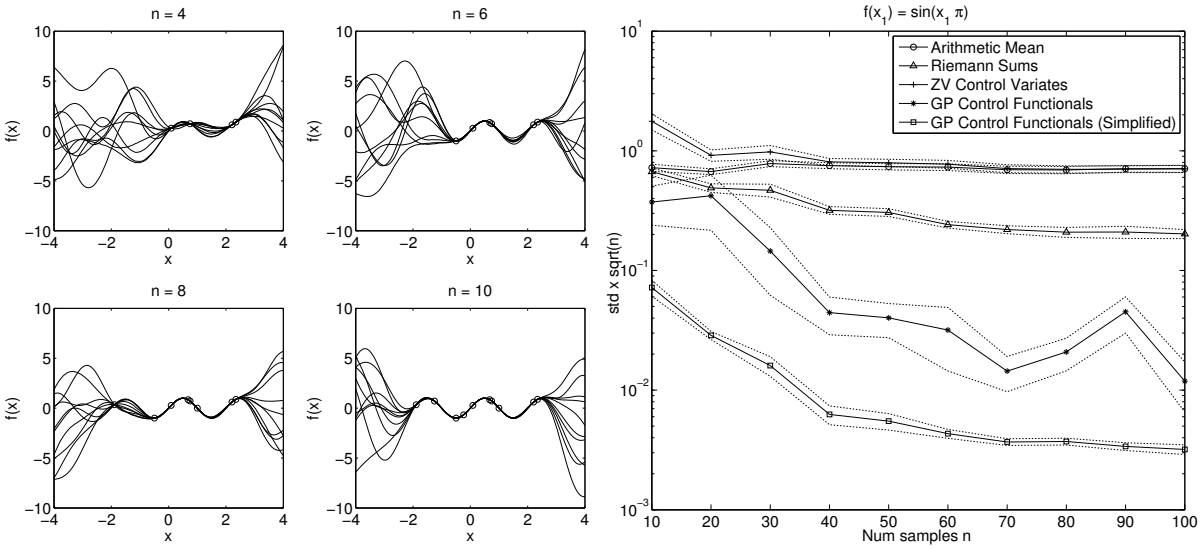
Figure 2: Simple sine/Gaussian case study. Here we display the empirical sampling distribution of Monte Carlo estimators, based on $n = 50$ samples and 100 independent realisations. [The settings for all methods were as described in the main text.]

Moreover both f and π are smooth so that we would expect (A6) to hold when the squared-error covariance function is used.

Our GPCF methodology was calibrated using cross-validation, in line with the implementation that we recommend above. Specifically: (i) We selected the covariance hyper-parameter $\ell = 1$ on the basis that this approximately minimised the cross-validation error (Supp. Fig. 1a). (ii) We found that estimator variance due to sample-splitting was minimised when most of the samples were allocated to $\mathcal{D}_0$ (Supp. Fig. 1b). We therefore set $m/n \approx 0.8$. (iii) Empirical results showed that little additional variance reduction occurs from employing multiple splits (Supp. Fig. 1c). (iv) Finally, we found that the bias of the simplified estimator was negligible ($\sim 10^{-4}$) compared to Monte Carlo error ($\sim 10^{-3}$) (Supp. Fig. 1d). 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 GPCF estimators as a function of the total number of samples n. The alternative approaches of the arithmetic mean, Riemann sums and ZV control variates are also shown, the latter being based on quadratic polynomials since these were most extensively studied by Papamarkou *et al.* (2014a). It is visually apparent that GPCFs enjoy the lowest variance at all samples sizes considered. To understand the superior performance of GPCFs we plot samples from the (simplified) GP posterior predictive distribution $\boldsymbol{f}_1|\mathcal{D}_0$ for increasing numbers of samples n. Intuitively, variance reduction

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism



(a) Draws from the GP posterior

(b) Relative variance of estimators ($d = 1$)

Figure 3: Simple sine/Gaussian case study (continued). (a) Ten draws were obtained from the (simplified) GP posterior for sample sizes of $n = 4$, 6, 8, 10, shown here as curves. Samples $\mathcal{D}$ are denoted by circles. [Here we used a GP length-scale hyper-parameter $\ell = 1$.] (b) Empirical assessment of asymptotic properties. Scalar case ($d = 1$).

requires that we have enough samples to build up a good approximation to f. Results in Fig. 3a suggest that this number n can in practice be quite small. We note at this point that, in this toy example where there are essentially no computational restrictions, the GPCF framework is unnecessary and gains in precision come with comparable gains in computational cost. However we emphasise that, in the serious applications that follow, the GPCF calculations requires negligible computational resources in comparison to simulation from the model.

Since the performance of GPCF is so remarkable, in order to more clearly visualise the results for all sample sizes, in Fig. 3b we plot the estimator standard deviation multiplied by $n^{1/2}$, so that root-n convergence corresponds to a horizontal line. Empirical results here are consistent with theory, showing that the arithmetic mean and ZV control variates all achieve a constant factor variance reduction, whereas Riemann sums and GPCFs sub-root-n convergence. In this example GPCFs significantly outperformed Riemann sums. We plot results for both the sample-splitting GPCF estimator and the simplified GPCF estimator, observing that the latter is more stable.

To assess the generality of our conclusions we considered going beyond the scalar case to higher-dimensional examples with $d = 3$ and $d = 5$. The analogous results in Supp. Figs. 2a, 2b (available online) show that, whilst the performance of all methods become increasingly similar in higher dimensions, simplified GPCF continues to out-perform the alternatives. Going further we considered a variety of alternative problems, varying both the function f and the underlying distribution π. These include several pathological cases, with results summarised in Table 1. The results marked (b) echo the conclusions of Mira *et al.* (2013), that ZV control variates are effective in special cases where $f(\boldsymbol{x})$ is well-approximated by a low-degree polynomial and π is a Gaussian or a Gamma density. However, when $f(\boldsymbol{x})$ is not well-approximated by a low-degree polynomial or when π takes a different form, as in cases marked (c), ZV control variates do not offer much variance reduction

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

	Problem		Mean Square Error				Notes
	$f(\boldsymbol{x})$	π	Arithmetic Mean	Riemann Sums	ZV Control Variates	GPCF (Simplified)	
GPCF ✓	$\sin(\pi x)$	$N(0,1)$	0.011 ± 0.0014	0.0014 ± 0.00026	0.011 ± 0.0014	$\mathbf{4.0e-07 \pm 7.9e-08}$	
	$\sin\left(\frac{(x_1+x_2+x_3)\pi}{3}\right)$	$N(\mathbf{0}_{3\times1}, \boldsymbol{I}_{3\times3})$	0.0084 ± 0.0011	–	0.013 ± 0.0023	$\mathbf{0.00067 \pm 0.00015}$	
	$\sin\left(\frac{(x_1+\cdots+x_5)\pi}{5}\right)$	$N(\mathbf{0}_{5\times1}, \boldsymbol{I}_{5\times5})$	0.0098 ± 0.0015	–	0.012 ± 0.0018	$\mathbf{0.0052 \pm 0.00091}$	
	$\sin\left(\frac{(x_1+\cdots+x_{10})\pi}{10}\right)$	$N(\mathbf{0}_{10\times1}, \boldsymbol{I}_{10\times10})$	0.0081 ± 0.0010	–	5.2 ± 2.6	$\mathbf{0.0078 \pm 0.00099}$	(a)
	1	$N(0,1)$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$1.6e-33 \pm 4.2e-34$	(b)
	x	$N(0,1)$	0.023 ± 0.0026	0.0073 ± 0.0011	$\mathbf{7.8e-34 \pm 1.7e-34}$	$1.4e-06 \pm 3.3e-07$	
	x_1	$N(\mathbf{0}_{3\times1}, \boldsymbol{I}_{3\times3})$	0.024 ± 0.0034	–	$\mathbf{4.0e-33 \pm 8.6e-34}$	0.0028 ± 0.00051	
	x^2	$N(0,1)$	0.040 ± 0.0059	0.050 ± 0.0041	$\mathbf{4.8e-33 \pm 1.4e-33}$	$4.5e-05 \pm 5.8e-06$	
	x	$\Gamma(5,1)$	0.091 ± 0.013	0.043 ± 0.0075	$\mathbf{3.2e-32 \pm 1.6e-32}$	0.019 ± 0.0020	
	$\exp(x)$	$N(0,1)$	0.090 ± 0.011	0.049 ± 0.0063	0.016 ± 0.0037	$\mathbf{0.00019 \pm 3.7e-05}$	(c)
	$x^2\sin(\pi x)$	$N(0,1)$	0.028 ± 0.0041	0.014 ± 0.0034	0.026 ± 0.0033	$\mathbf{0.00014 \pm 4.8e-05}$	
	x	$\beta(2,2)$	0.00094 ± 0.00012	$0.00013 \pm 1.6e-05$	$0.00086 \pm 9.9e-05$	$\mathbf{2.1e-11 \pm 4.5e-12}$	
	$\sin(2\pi x)$	$\beta(2,2)$	0.012 ± 0.0015	0.00093 ± 0.00015	0.021 ± 0.0035	$\mathbf{2.5e-06 \pm 4.2e-07}$	
GPCF ×	$\sin(\pi x)$	Cauchy$(0,1)$	0.0095 ± 0.0015	0.033 ± 0.0080	$\mathbf{0.0086 \pm 0.0013}$	1.8 ± 0.25	(d)
	x	Exp(1)	0.014 ± 0.0021	$\mathbf{0.0094 \pm 0.0014}$	0.014 ± 0.0021	0.81 ± 0.076	(e)
	x	Non-differentiable	0.00092 ± 0.00011	$0.00013 \pm 2.2e-05$	0.0079 ± 0.0012	$\mathbf{6.1e-05 \pm 1.4e-05}$	(f)
	$1_{x>0}$	$N(0,1)$	0.0050 ± 0.00079	$\mathbf{0.0016 \pm 0.00019}$	0.0022 ± 0.00026	$2.7e+04 \pm 1.1e+04$	(g)

Table 1: A broad range of examples, some of which are compatible with the GPCF methodology (GPCF ✓) and some of which are not (GPCF ×). In each case the true mean μ is known analytically and we report the mean square error with respect to μ, computed over 100 independent realisations. [Here the number of samples was fixed at $n = 50$. For all examples we employed (simplified) GPCF with the square error covariance function and hyper-parameter $\ell = 1$. Riemann sums were not used in the multi-dimensional problems due to their non-trivial implementation.] Notes: (a) ZV control variates perform poorly since their implementation here requires us to estimate a 10×10 covariance matrix from only $n = 50$ samples. On the other hand, GPCF offers stability here that is comparable with the usual arithmetic mean. (b) For these low-dimensional polynomial examples, ZV control variates are essentially exact. (c) In each of these examples GPCF offers the most precise estimates. (d) Here $f \notin \mathcal{L}^2(\pi)$, violating (A1). (e) Here $[\pi(x)\phi(x)]_{x=0}$ is not almost surely equal to zero, violating (A5). (f) Here π is not differentiable, violating (A2). This "non-differentiable" example uses a triangular distribution on $[0, 1]$. (g) Here the GP operator $\mathcal{E}_{\mathcal{D}_0}$, based on the squared-error covariance function, is inappropriate since f is not continuous and therefore does not belong to the RKHS of k, likely violating (A6).

whereas GPCFs retain the ability to decrease variance, in some cases dramatically. We then investigated how GPCFs can fail when the assumptions of Theorem 2 are violated (see examples marked "GPCF ×"). As expected, violation of (A1), (A5) and (A6) in (d), (e), (g) respectively led to poor performance of the GPCF estimator. Interestingly, violation of (A2) in example (f) did not lead to poor estimation, though this may be because π is only non-differentiable at a single point.

3.2 Bayes-Hermite quadrature, revisited

Our first serious application of the control functional methodology is to reveal a new and powerful approach to Bayesian uncertainty quantification (BUQ) for Monte Carlo integration. In brief, BUQ attempts to model the numerical error that arises from implementing various mathematical and statistical procedures on a computer (see e.g. Diaconis, 1988; O'Hagan, 1992). We briefly review BUQ in the context of Monte Carlo integration and demonstrate how control functionals provide an elegant solution to three well-known problems in this area.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

3.2.1 BUQ for Monte Carlo integration

Consider a simple scenario that involves a random variable X, a sufficiently "regular" function f and $n = 3$ samples x_1, x_2, x_3 that are generated independently from a distribution π. It might happen that x_2 and x_3 lie very close together due to chance, i.e. not simply due to the form of π itself. Then $f(x_2) \approx f(x_3)$, so that having computed $f(x_2)$ we learn almost nothing about μ by then computing $f(x_3)$. In this situation O'Hagan (1987) argued that it is inappropriate to estimate μ using the arithmetic mean, which essentially places two-thirds of the total weight on $f(x_2)$ and just one third on $f(x_1)$. At a fundamental level, the use of the arithmetic mean contradicts the conditionality principle, that inferences should be made conditional on the actual values of the x_1, x_2, x_3 that were observed (Birnbaum, 1962). A Bayesian solution was provided by O'Hagan (1991), under the name "Bayes-Hermite Quadrature", later rediscovered by Rasmussen and Ghahramani (2003) under the name "Bayesian Monte Carlo". More recent work in this direction includes Osborne *et al.* (2012). We do not review this work in detail here, but note that in each case the authors proceeded by placing a GP prior directly on the function f. This has three major drawbacks that have not yet been resolved: (i) The resulting estimators for μ are biased, with bias depending on the prior information used. (ii) The main analytically tractable case requires that π is a Gaussian density, which is a serious constraint. Rasmussen and Ghahramani (2003) proposed importance sampling to transform non-Gaussian problems, but this introduces additional computational difficulties. (iii) There is, at present, no theoretical argument to support sub-root-n convergence. Together these three factors may explain why BUQ methods are not, at present, widely used in Monte Carlo applications.

3.2.2 An elegant solution using control functionals

We adopt a Bayesian perspective on solving Eqn. 5 for the intercept parameter μ under a prior $\mu \sim N(\mu_0, \sigma_\mu^2)$, where σ_μ may be infinite, and the GP framework outlined above. For clarity we present the simplified estimator, but the sample-splitting approach proceeds analogously. The posterior mean of the intercept term in the semi-parametric regression problem of Eqn. 13 is given by

$$\hat{\mu}(\mathcal{D}) = \frac{\mu_0 \sigma_\mu^{-2} + \mathbf{1}_{1\times n} \boldsymbol{K}^{-1} \boldsymbol{f}}{\sigma_\mu^{-2} + \mathbf{1}_{1\times n} \boldsymbol{K}^{-1} \mathbf{1}_{n\times 1}}. \tag{24}$$

Moreover the posterior for μ is normally distributed with precision

$$\mathrm{prec}(\hat{\mu}(\mathcal{D})) = \sigma_\mu^{-2} + \mathbf{1}_{1\times n} \boldsymbol{K}^{-1} \mathbf{1}_{n\times 1}. \tag{25}$$

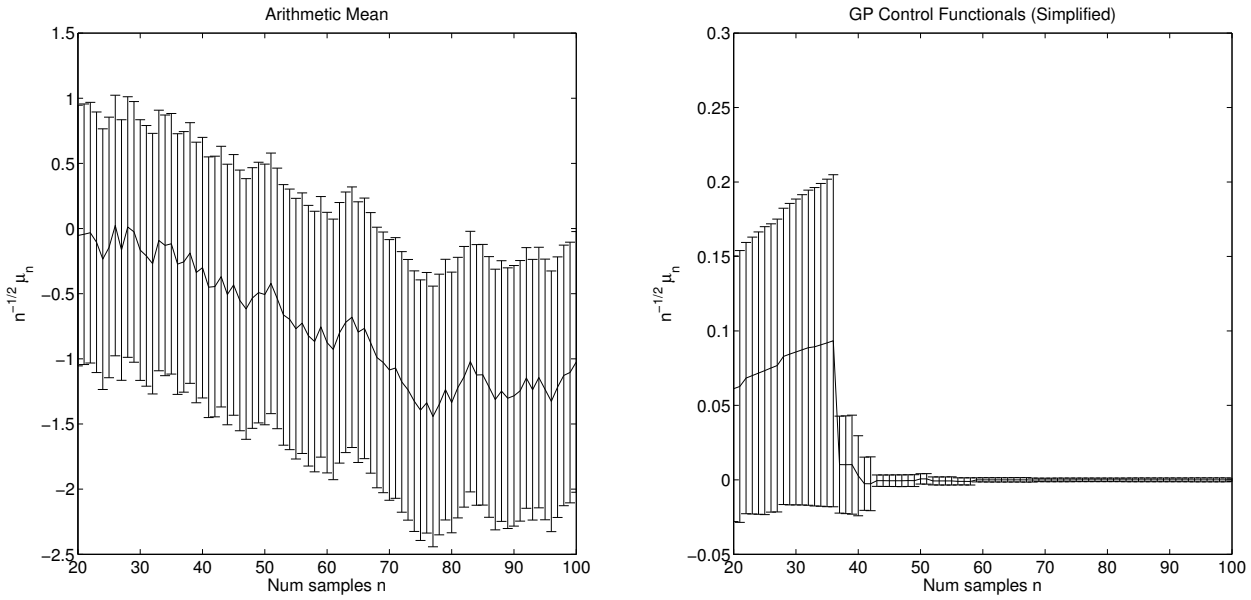
Thus we are able to construct e.g. credible intervals

$$\mu \in [\hat{\mu}(\mathcal{D}) - \alpha(\sigma_\mu^{-2} + \mathbf{1}_{1\times n} \boldsymbol{K}^{-1} \mathbf{1}_{n\times 1})^{-1/2}, \hat{\mu}(\mathcal{D}) + \alpha(\sigma_\mu^{-2} + \mathbf{1}_{1\times n} \boldsymbol{K}^{-1} \mathbf{1}_{n\times 1})^{-1/2}] \tag{26}$$

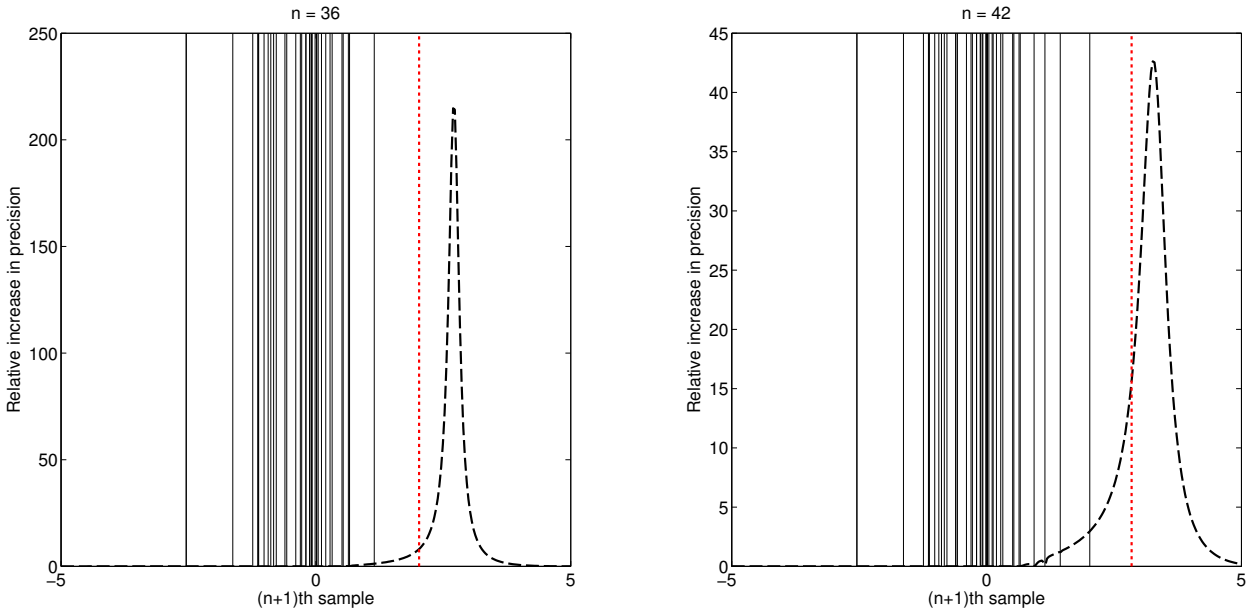
with posterior probability $1 - 2\Phi(-\alpha)$ where Φ is the cumulative distribution function of the standard normal distribution.

This methodology clearly addresses the three drawbacks of earlier BUQ proposals; (i) it can provide an unbiased estimator via sample-splitting; (ii) π is not required to be Gaussian; (iii) the estimator can converge at a sub-root-n rate. These extra advantages seem related to the fact that we employ prior information on the joint smoothness properties of both f and the density π, rather than on the marginal properties of f.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism



(a)



(b)

Figure 4: Bayesian uncertainty quantification using control functionals. (a) Here we contrast the classical $O_P(n^{-1/2})$ confidence interval for the arithmetic mean (left) against the Bayesian posterior credible interval for GP control functionals (GPCF; right). On the y-axis we plot the estimator scaled by $n^{1/2}$, so that the classical confidence interval is represented with error-bars that have constant height, whereas the GPCF error-bars will decrease in height with n. The true expectation in this sine/Gaussian example is $\mu = 0$. (b) Here we examine the location of the $(n+1)$st samples x_{n+1} for $n = 36$ and $n = 42$, which each cause the credible interval in (a; right) to contract sharply. In each case we display the relative increase in posterior precision that would result from observing a hypothetical new sample x^*_{n+1} (dashed line). The $(n+1)$st sample that was actually observed (dotted line) lies in a region that has not been adequately explored by samples $x_1, \ldots, x_n$.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

3.2.3 A simple case study

The methodology is illustrated using the sine/Gaussian problem considered earlier (Eqn. 23, $d = 1$). In Fig. 4a we contrast classical confidence intervals for the arithmetic mean and Bayesian credible intervals for the control functional estimator. Notice that the width of the GPCF credible interval vanishes at sub-root-n, whereas the width of the familiar confidence interval vanishes at root-n. The Bayesian posterior precision differs from frequentist precision both fundamentally, in its interpretation, and behaviourally, in the sense that it is adaptive, depending on the actual location of the samples $\mathcal{D}$ in the state space $\mathcal{X}$. The sudden increases in precision of the GPCF estimator at $n = 36$ and $n = 42$ in this realisation occur when the new sample x_{n+1} lies in a region of the state space $\mathcal{X}$ that has not yet been explored by the previous samples $x_1, \ldots, x_n$. This is because we gain a large amount of new information in these circumstances. Fig. 4b illustrates this behaviour, where we plot the relative increase in posterior precision that would occur from observing a hypothetical point x^*_{n+1} (dashed line) along with the location of the actual sampled value x_{n+1} (dotted line). In each case the new sample x_{n+1} lies in the informative region, where the increase in posterior precision is large.

The methodology presented here suggests further applications beyond BUQ: (i) A Bayesian solution to the experimental design problem for numerical quadrature, based on selecting a design point x_{n+1} in order to maximise the posterior precision over all hypothetical samples x^*_{n+1}. (ii) Using posterior credible intervals to construct a Bayesian analogue to Kendall *et al.* (2007), who produced diagnostic tools for Monte Carlo estimators based on MCMC examples using "confidence bands" derived from a functional central limit theorem.

3.3 Marginalisation in hierarchical models

Our second application addresses a frequently encountered problem in hierarchical modelling; the setting of hyper-parameters. The fully-Bayesian solution is to marginalise over hyper-parameters, but this often entails a prohibitive level of computation. Here we explore whether control functionals can confer a gain in computational efficiency in this setting.

3.3.1 A hierarchical GP model

The marginalisation of uncertain hyper-parameters in hierarchical models is a widely occurring problem, in particular in spatial statistics (Besag and Green, 1993). 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 $Y_i = g_i + \epsilon_i$ 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

$$c(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\theta}) = \theta_1 \exp\left(-\frac{(\boldsymbol{z} - \boldsymbol{z}')^T(\boldsymbol{z} - \boldsymbol{z}')}{2\theta_2^2}\right). \tag{27}$$

Here $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are unknown 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 independent priors, say $\theta_1 \sim \Gamma(\alpha, \beta)$, $\theta_2 \sim \Gamma(\gamma, \delta)$ in the shape/scale parametrisation, which we write jointly as $\pi(\boldsymbol{\theta})$.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

3.3.2 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

$$\hat{Y}_* := \mathbb{E}[Y_*|\boldsymbol{y}] = \int \mathbb{E}[Y_*|\boldsymbol{y}, \boldsymbol{\theta}]\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{28}$$

where we implicitly condition on the covariates $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N, \boldsymbol{z}_*$. Eqn. 28 is unavailable in closed form and we therefore construct a Monte Carlo estimate by sampling $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ independently from the prior $\pi(\boldsymbol{\theta})$. Phrasing in terms of our previous notation, the function of interest is

$$f(\boldsymbol{\theta}) = \mathbb{E}[Y_*|\boldsymbol{y}, \boldsymbol{\theta}] = \boldsymbol{C}_{*,N}(\boldsymbol{C}_N + \sigma^2 \boldsymbol{I}_{N\times N})^{-1}\boldsymbol{y} \tag{29}$$

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

$$f(\boldsymbol{\theta}) \approx \boldsymbol{C}_{*,N'}(\boldsymbol{C}_{N',N}\boldsymbol{C}_{N,N'} + \sigma^2 \boldsymbol{C}_{N'})^{-1}\boldsymbol{C}_{N',N}\boldsymbol{y} \tag{30}$$

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. 30 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.

3.3.3 SARCOS robot arm

We used the hierarchical GP model in Sec. 3.3.2 to learn the inverse dynamics of a seven degrees-of-freedom SARCOS anthropomorphic robot arm. The task, as described in Rasmussen and Williams (2006), is to map from a 21-dimensional input space (7 positions, 7 velocities, 7 accelerations) to the corresponding 7 joint torques. 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 linearly rescaled 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 θ_i has a prior mean of 1 and a prior standard deviation of 0.2. Moreover these values imply that the prior $\pi(\boldsymbol{\theta})$ satisfies the boundary conditions of Eqn. 2, so that (A1)-(A5) hold and the GPRF estimator is well-defined.

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 GPCF we took a default hyper-parameter $\ell = 1$ that reflects 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. 5. 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,

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

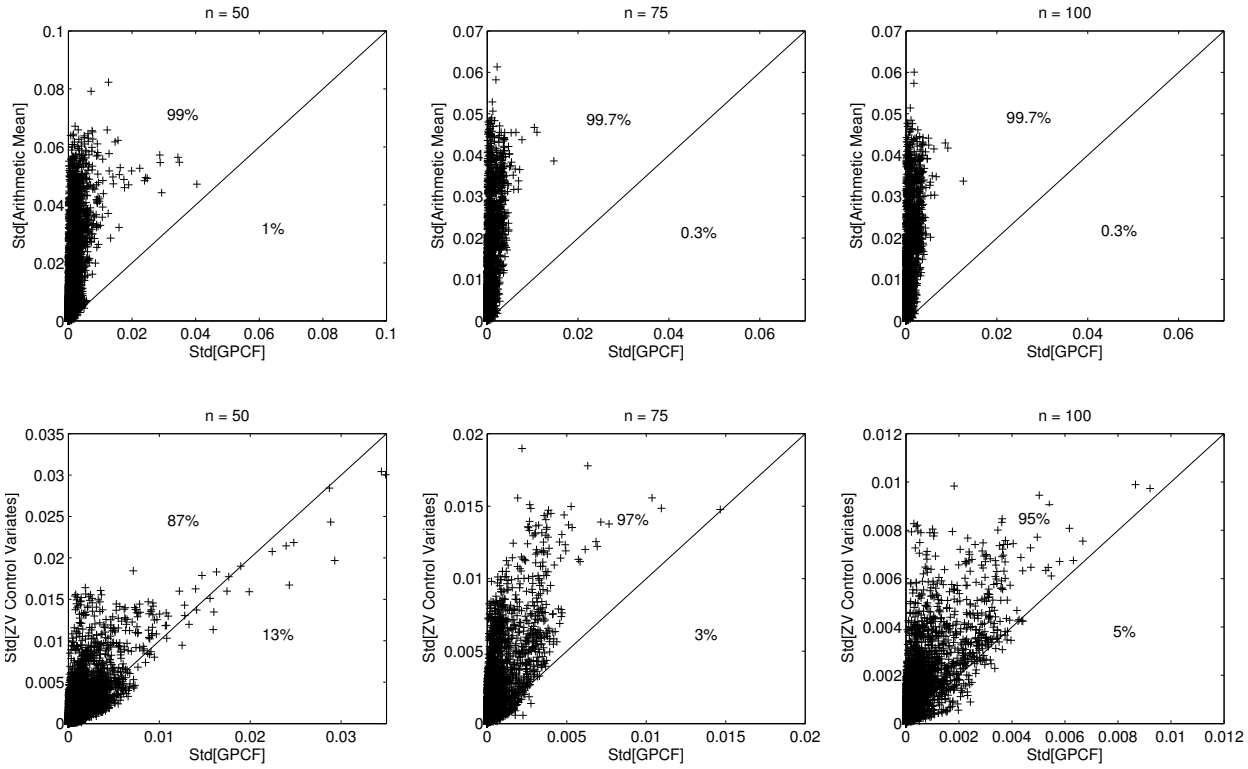


Figure 5: Marginalisation of hyper-parameters in hierarchical models. [Here we display the sampling standard deviation of Monte Carlo estimators for the posterior predictive mean $\mathbb{E}[Y_*|\boldsymbol{y}]$ in the SARCOS robot arm example, computed over 10 independent realisations. Each point, representing one Monte Carlo integration problem, is represented by a cross. Percentages indicate the number of crosses above/below the diagonal.]

for the vast majority of integration problems, GPCF achieves a lower estimator variance compared with both the arithmetic mean estimator and ZV control variates. Thus GPCF requires fewer evaluations of the function f whilst delivering a precision that is equivalent to existing methods. Note that the cost of post-processing the Monte Carlo samples (using either ZV control variates or GPCF) will be negligible in comparison to the cost of evaluating the function f, even once. Indeed, GPCF requires that we invert a $n \times n$ matrix once, where n is typically much smaller than N'.

Going further, we investigated whether control functionals were still able to confer these gains when employed alongside state-of-the-art (R)QMC techniques. Since (R)QMC is a sampling methodology, it is in a sense orthogonal to post-processing techniques including control functionals and it is interesting to investigate whether a combined approach achieves superior estimation performance. Note, however, that samples from (R)QMC are not independent and thus the theoretical framework that we described above does not directly apply - this motivates the empirical investigation below. Specifically, we focus on a RQMC Halton scheme with scrambling that has been shown to achieve $O_P(n^{-3/2+\epsilon})$ convergence for any $\epsilon > 0$ (Owen, 1997). Results in Supp. Fig. 3 show that at low sample sizes ($n = 50$) the combined approach of GPCF + RQMC achieves lower-variance estimation compared to GPCF alone in 82% of the SARCOS integration problems.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

At larger sample sizes ($n = 75$, $n = 100$), whilst the performance of GPCF + RQMC was significantly better than GPCF alone, the actual variance reduction achieved was not substantial. This suggests a shared origin for the superior performance of GPCF and RQMC, so that the combined algorithm does not offer much advantage over either individually. Indeed, both approaches are motivated by exploiting a metric on $\mathcal{X}$; GPCF essentially re-weights the samples $\boldsymbol{x}_i$ according to their separation in $\mathcal{X}$ (defined implicitly by the covariance function), whilst RQMC ensures that samples $\boldsymbol{x}_i$ are well-separated in $\mathcal{X}$ *ab initio*.

3.4 Normalising constants for non-linear ODE models

Our final application concerns the estimation of normalising constants for non-linear ODE models; a problem that is known to be extremely challenging (Calderhead and Girolami, 2009). Recent empirical investigations recommend thermodynamic integration (TI) as one of the best-performing approaches to the estimation of normalising constants in complex models (Vyshemirsky and Girolami, 2008; Friel and Wyse, 2012). The control variate methodology of Mira *et al.* (2003) was recently applied to TI by Oates *et al.* (2014), 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. Below we study the application of control functionals to TI in this setting where CTI is less effective. We note that control functionals can apply also to other estimators of normalising constants, for example based on sequential Monte Carlo (Zhou *et al.*, 2013).

3.4.1 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

$$\log p(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y},t}[\log p(\boldsymbol{y}|\boldsymbol{\theta})]dt \tag{31}$$

where the expectation in the integrand is with respect to the power posterior whose density is given above. The correctness of Eqn. 31 is established in e.g. Friel and Pettitt (2008). In TI, the one-dimensional integral in Eqn. 31 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):

$$\hat{\mu} := \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2}(\hat{\mu}_i + \hat{\mu}_{i+1}) - \frac{(t_{i+1} - t_i)^2}{12}(\hat{\nu}_{i+1} - \hat{\nu}_i), \tag{32}$$

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 $\boldsymbol{\theta}|\boldsymbol{y}, t_i$. In complex models, $\log p(\boldsymbol{y}|\boldsymbol{\theta})$ will be poorly approximated by a low-degree polynomial and $\boldsymbol{\theta}|\boldsymbol{y}, t$ will be non-Gaussian; this explains the mediocre performance of CTI in these cases. In contrast, control functionals should still be able to deliver gains in estimation.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

3.4.2 Non-linear ODE models

We consider first-order non-linear dynamical systems of the form

$$\frac{d\boldsymbol{x}}{ds} = \boldsymbol{F}(\boldsymbol{x}, s; \boldsymbol{\theta}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{33}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^p$ is assumed known, $p < \infty$, and $\boldsymbol{F} : \mathbb{R}^p \times [0, \infty) \to \mathbb{R}^p$ is non-linear. We assume that only a subset of the state variables are observed under noise, so that $\boldsymbol{x} = [\boldsymbol{x}_a, \boldsymbol{x}_b]$ and $\boldsymbol{y}$ is a $d \times n$ matrix of observations of the coordinates $\boldsymbol{x}_a$. Write $s_1 < s_2 < \cdots < s_n$ for the times at which observations are obtained, such that $\boldsymbol{y}(s_j) = \boldsymbol{y}_{\bullet,j}$ where $\boldsymbol{y}_{\bullet,j}$ is the jth column of the data matrix $\boldsymbol{y}$. We consider a Gaussian observation process with likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x}_0, \sigma) = \prod_{j=1}^{n} \mathcal{N}(\boldsymbol{y}(s_j)|\boldsymbol{x}_a(s_j; \boldsymbol{\theta}, \boldsymbol{x}_0), \sigma^2 \boldsymbol{I}_{d\times d}) \tag{34}$$

where $\boldsymbol{x}_a(s_j; \boldsymbol{\theta}, \boldsymbol{x}_0)$ denotes the deterministic solution of the system in Eqn. 33 and $\sigma > 0$ is assumed known. For the Gaussian observation model it can be shown that boundary conditions of Eqn. 4 are implied by the parameter prior density $p(\boldsymbol{\theta})$ vanishing when $\boldsymbol{\theta} \to \infty$ in any norm.

Assuming the boundary condition holds, we have

$$u_i(\boldsymbol{\theta}) = \nabla_{\theta_i} \log p(\boldsymbol{\theta}) + \frac{t}{\sigma^2} \sum_{j=1}^{n} \boldsymbol{S}^i_{j,1:d}(\boldsymbol{y}(s_j) - \boldsymbol{x}_a(s_j; \boldsymbol{\theta}, \boldsymbol{x}_0)) \tag{35}$$

where $\boldsymbol{S}^i$ is a matrix of sensitivities with entries $S^i_{j,k} = \frac{\partial x_k}{\partial \theta_i}(s_j)$. Note that in Eqn. 35, $\boldsymbol{S}^k_{j,k}$ ranges over indices $1 \leq k \leq d$ corresponding only to the observed variables. In general the sensitivities $\boldsymbol{S}^i$ will be unavailable in closed form, but may be computed numerically by augmenting the system of ordinary differential equations, as

$$\dot{S}^i_{j,k} = \frac{\partial F_k}{\partial \theta_i} + \sum_{l=1}^{p} \frac{\partial F_k}{\partial x_l} S^i_{j,l} \tag{36}$$

where $\frac{\partial x_k}{\partial \theta_i} = 0$ at $s = 0$. Indeed, these sensitivities are already computed when differential-geometric sampling schemes are employed, so that the evaluation of Eqn. 35 incurs negligible additional computational cost (Girolami and Calderhead, 2011; Papamarkou *et al.*, 2014a).

3.4.3 The van der Pol oscillator

The methodology is illustrated on the van der Pol oscillator (van der Pol, 1926), a non-conservative oscillator with non-linear damping that has classical modelling applications in fields ranging from neuroscience (FitzHugh, 1961) to seismology (Cartwright *et al.*, 1999). Here a position $x(s) \in \mathbb{R}$ evolves in time s according to the second order differential equation

$$\frac{d^2x}{dt^2} - \theta(1 - x^2)\frac{dx}{dt} + x = 0 \tag{37}$$

where $\theta \in \mathbb{R}$ is an unknown parameter indicating the non-linearity and the strength of the damping. Letting $x_1 := x$ and $x_2 := dx/dt$ we can formulate the oscillator as the first-order system

$$\boldsymbol{F}(\boldsymbol{x}, s; \theta) = \begin{cases} x_2 \\ \theta(1 - x_1^2)x_2 - x_1 \end{cases} \tag{38}$$

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

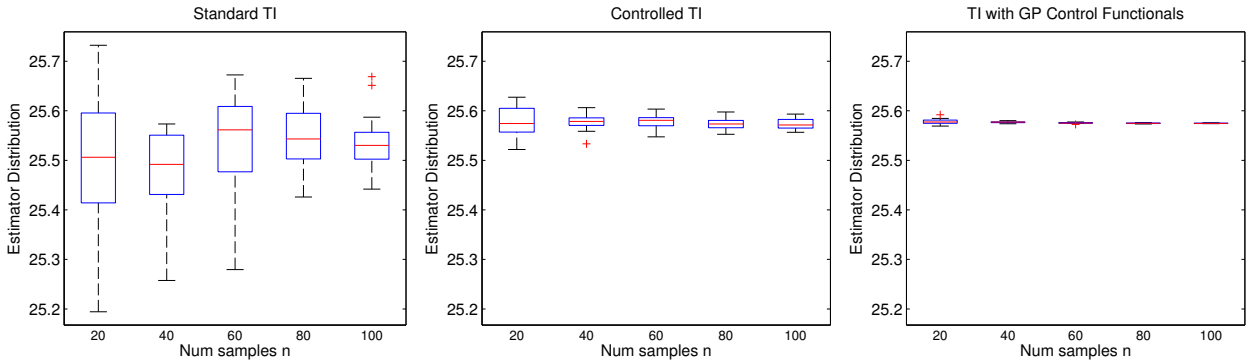


Figure 6: Estimation of normalising constants for non-linear ordinary differential equations using thermodynamic integration (TI); van der Pol oscillator example. [Here we show the distribution of 100 independent realisations of each estimator. "Standard TI" is based on arithmetic means. "Controlled TI", proposed in Oates *et al.* (2014), is based on ZV control variates.]

where only the first component x_1 is observed. This system was solved numerically using $\theta = 1$, $\boldsymbol{x}_0 = [0, 2]$. Observations were made once every second, up to 10 seconds, and Gaussian measurement noise of standard deviation $\sigma = 0.1$ was added. A log-normal prior was placed on θ such that $\log(\theta) \sim N(0, 0.25)$. This prior vanishes at the origin $\theta = 0$ and has decaying tails as $\theta \to \infty$, so that the boundary condition in Eqn. 4 is satisfied and GPCF can be used.

For the model-based computation, 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 $\boldsymbol{\theta}|\boldsymbol{y}, t_i$ 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). 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).

Results in Fig. 6 show that the CTI estimator improves upon the standard TI estimator, but a much more substantial reduction in estimator variance results from using the GPCF methodology. For the GPCF computation we have used the simplified but biased GPCF estimator, since TI in any case produces a biased estimate for the normalising constant due to numerical quadrature. The squared-error covariance function was used and the hyper-parameter $\ell = 3$ was selected manually on the basis of cross-validation. The additional cost of using GPCF is essentially zero relative to running the population MCMC sampler, the latter requiring repeated solution of the ODE system. These results demonstrate that GPCF requires far fewer samples n compared with CTI in practice, in order to achieve a desired estimator precision.

4 Discussion

In this paper we developed an approach to Monte Carlo integration that can achieve sub-root-n convergence. Methods that aim to reduce Monte Carlo variance can play an important role in modern applications of statistical methodology, for example in applications involving large-scale spatial

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

models or involving complex computer models. An important feature of the approach presented here is that variance reduction is formulated as a post-processing 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 for e.g. the GP covariance function can be made *after* performing the simulations.

Through exploitation of Gaussian processes we were able to realise our general framework and construct estimators with an analytic form. Theoretical results for presistency of GPs is an active area of research and currently GPs are widely used for both inference and prediction in settings where theoretical guarentees are not yet established. A violation of presistency could lead to high-variance estimation and we therefore would still seek theoretical guarantees that this will be a low-probability occurrence. We presented a cross-validation diagnostic approach that can assist in detecting these violations. Empirical results evidenced the practical utility of control functional estimators in settings where the score function is available and the dimensionality of the problem is not too large (e.g. ≤ 10).

Our work suggests several interesting research directions: (i) The score function and hence the control functional methodology is not parameterisation-invariant. It would be interesting to elicit an effective parametrisation as an additional post-processing step. (ii) Recent work by Friel *et al.* (2014) showed how an unbiased estimate for the score statistic can be obtained for a wide class of (doubly) intractable likelihood models. Though we did not discuss it here, it is straight-forward to extend the control functional methodology to this setting, with an appropriate modification to the covariance function to account for uncertainty in an estimate for the score. (iii) Exploring in detail the non-IID setting, where samples arise from a Markov chain, for example. (iv) Extending the methodology to non-differentiable models using suitable alternatives to the classical derivative (e.g. Pereyra, 2013). (v) We sketched in Appendix B how orthogonal basis expansions can be used to estimate control functionals. A natural direction to pursue is the extension of the control functional methodology to this and related non-parametric techniques.

Acknowledgements: The authors are extremely grateful to Taeryon Choi, Christian Robert, Gareth Roberts, Antonietta Mira, Daniel Simpson and Aad van der Vaart for helpful discussions during the preparation of this manuscript. CJO was supported by the EPSRC grant "Centre for Research in Statistical Methodology" [EP/D002060/1]. MG was supported by EPSRC [EP/J016934/1], the EU grant "Analysing and Striking the Sensitivities of Embryonal Tumours" [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

Lemma 3. *For a real-valued function $g(\boldsymbol{X}, \boldsymbol{Y})$ of two random variables $\boldsymbol{X}$, $\boldsymbol{Y}$ such that $\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[g(\boldsymbol{X}, \boldsymbol{Y})] = 0$ we have $\mathbb{V}_{\boldsymbol{X},\boldsymbol{Y}}[g(\boldsymbol{X}, \boldsymbol{Y})] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{V}_{\boldsymbol{Y}|\boldsymbol{X}}[g(\boldsymbol{X}, \boldsymbol{Y})]]$.* □

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

Proof of Theorem 1. Since $\mathbb{E}_{\mathcal{D}_1|\mathcal{D}_0}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) - \mu] = 0$ we have from Lemma 3 that

$$\begin{aligned}
\mathbb{V}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)] = \mathbb{V}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) - \mu] &= \mathbb{E}_{\mathcal{D}_0}[\mathbb{V}_{\mathcal{D}_1}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1) - \mu]] \\
&= \mathbb{E}_{\mathcal{D}_0}\left[\mathbb{V}_{\mathcal{D}_1}\left[\frac{1}{n-m}\sum_{i=m+1}^{n} \tilde{f}_{\mathcal{D}_0}(\boldsymbol{x}_i) - \mu\right]\right] \\
&= \mathbb{E}_{\mathcal{D}_0}\left[\frac{1}{n-m}\mathbb{V}_{\boldsymbol{X}}\left[\tilde{f}_{\mathcal{D}_0}(\boldsymbol{X}) - \mu\right]\right] \\
&= \frac{1}{n-m}\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}] \qquad (39)
\end{aligned}$$

since the $\tilde{f}_{\mathcal{D}_0}(\boldsymbol{X}_i)$ are independent and identically distributed for $i = m+1, \ldots, n$ given $\mathcal{D}_0$. Letting $\sigma^2 := \mathbb{V}[f(\boldsymbol{X})]$ we have

$$\frac{\mathbb{V}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)]}{\mathbb{V}_{\mathcal{D}}[\overline{\mu}_n(\mathcal{D})]} = \frac{\frac{1}{n-m}\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}]}{\frac{1}{n}\sigma^2} = \frac{n}{n-m}\frac{\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}]}{\sigma^2}. \qquad (40)$$

It is clear that if the size of both $\mathcal{D}_0$ and $\mathcal{D}_1$ tend linearly to infinity then $n/(n-m)$ tends to a positive constant and thus Eqn. 40 vanishes as $n \to \infty$ under the presistency hypothesis. □

Proof of Lemma 1. When k_0 is twice differentiable, the derivative of a GP is itself a GP that satisfies

$$\begin{bmatrix} \phi_i(\boldsymbol{x}) \\ \nabla\phi_i(\boldsymbol{x}) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} k_0(\boldsymbol{x},\boldsymbol{x}') & \nabla_{\boldsymbol{x}'}k_0(\boldsymbol{x},\boldsymbol{x}') \\ \nabla_{\boldsymbol{x}}k_0(\boldsymbol{x},\boldsymbol{x}') & \nabla_{\boldsymbol{x}}\nabla_{\boldsymbol{x}'}k_0(\boldsymbol{x},\boldsymbol{x}') \end{bmatrix}\right). \qquad (41)$$

Similarly, the product of a GP and a deterministic function $u_i(\boldsymbol{x})$ is itself a GP with

$$\phi_i(\boldsymbol{x})u_i(\boldsymbol{x}) \sim \mathcal{GP}(0, u_i(\boldsymbol{x})u_i(\boldsymbol{x}')k_0(\boldsymbol{x},\boldsymbol{x}')). \qquad (42)$$

It follows that

$$k(\boldsymbol{x},\boldsymbol{x}') = \sum_{i=1}^{d} \begin{array}{l} \nabla_{x_i}\nabla_{x_i'}k_0(\boldsymbol{x},\boldsymbol{x}') + u_i(\boldsymbol{x})\nabla_{x_i'}k_0(\boldsymbol{x},\boldsymbol{x}') \\ \quad +u_i(\boldsymbol{x}')\nabla_{x_i}k_0(\boldsymbol{x},\boldsymbol{x}') + u_i(\boldsymbol{x})u_i(\boldsymbol{x}')k_0(\boldsymbol{x},\boldsymbol{x}') \end{array}. \qquad (43)$$

The final claim in the Lemma follows from standard results for conditioning on components of Gaussian distributions (see e.g. sec. 2.7 of Rasmussen and Williams, 2006). □

Proof of Theorem 2. Since $f \in \mathcal{L}^2(\pi)$ and $\tilde{f}_{\mathcal{D}_0} = f - \mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\psi_{\boldsymbol{\phi}}]$, it is sufficient to show that $\mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\psi_{\boldsymbol{\phi}}(\boldsymbol{X})] \in \mathcal{L}^2(\pi)$, since sums of $\mathcal{L}^2(\pi)$ functions are themselves $\mathcal{L}^2(\pi)$ functions. From Lemma 1 we have

$$\mathbb{E}_{\boldsymbol{\phi}|\mathcal{D}_0}[\boldsymbol{\psi}_{\boldsymbol{\phi}}(\boldsymbol{x})] = \boldsymbol{K}(\boldsymbol{x},\mathcal{D}_0)\underbrace{\left[\boldsymbol{I}_{m\times m} - \frac{\boldsymbol{K}_0^{-1}\mathbf{1}_{m\times 1}\mathbf{1}_{1\times m}}{\mathbf{1}_{1\times m}\boldsymbol{K}_0^{-1}\mathbf{1}_{m\times 1}}\right]\boldsymbol{K}_0^{-1}\boldsymbol{f}_0}_{(*)} \qquad (44)$$

where $\boldsymbol{K}(\boldsymbol{x},\mathcal{D}_0)$ is a $1 \times m$ vector with ith entry $k(\boldsymbol{x},\boldsymbol{x}_i)$. Since $(*)$ is independent of $\boldsymbol{x}$ it suffices to show that each component of $\boldsymbol{K}(\boldsymbol{x},\mathcal{D}_0)$ belongs to $\mathcal{L}^2(\pi)$.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

For generic $\boldsymbol{x}' \in \mathcal{D}_0$ we have from Eqn. 43 that $k(\boldsymbol{x}, \boldsymbol{x}')$ is a sum of terms of the form $\nabla_{x_i}\nabla_{x_i'}k_0(\boldsymbol{x}, \boldsymbol{x}')$, $u_i(\boldsymbol{x})\nabla_{x_i'}k_0(\boldsymbol{x}, \boldsymbol{x}')$, $u_i(\boldsymbol{x}')\nabla_{x_i}k_0(\boldsymbol{x}, \boldsymbol{x}')$ and $u_i(\boldsymbol{x})u_i(\boldsymbol{x}')k_0(\boldsymbol{x}, \boldsymbol{x}')$. Since $u_i(\boldsymbol{x}')$ is constant in $\boldsymbol{x}$, it suffices to show that the terms $\nabla_{x_i}\nabla_{x_i'}k_0(\boldsymbol{x}, \boldsymbol{x}')$, $u_i(\boldsymbol{x})\nabla_{x_i'}k_0(\boldsymbol{x}, \boldsymbol{x}')$, $\nabla_{x_i}k_0(\boldsymbol{x}, \boldsymbol{x}')$ and $u_i(\boldsymbol{x})k_0(\boldsymbol{x}, \boldsymbol{x}')$ are in $\mathcal{L}^2(\pi)$. But this is clearly satisfied when k_0 and its partial derivatives are bounded and the score function has components $u_i(\boldsymbol{x}) \in \mathcal{L}^2(\pi)$, which is the case by hypothesis.

Sub-root-n convergence is an immediate corollary of Theorem 1. □

Proof of Lemma 2. From Eqn. 40 we have that

$$\frac{\mathbb{V}_{\mathcal{D}}[\hat{\mu}_{\mathcal{D}_0}(\mathcal{D}_1)]}{\mathbb{V}_{\mathcal{D}}[\overline{\mu}_n(\mathcal{D})]} = \frac{n}{n-m}\frac{\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}]}{\sigma^2}. \tag{45}$$

Under our asymptotic assumption $\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}] = O(m^{-\alpha})$, we thus seek to minimise $m^{-\alpha}(n-m)^{-1}$ over $m \in \{1, \ldots, n\}$. Approximating m by a continuous variable, elementary calculus shows that this minimum occurs at $m = \alpha(1+\alpha)^{-1}n$. □

B Presistency: A proof of principle

In this Appendix we investigate, in a simplified framework, whether the presistency assumption is reasonable; that is, whether is possible to estimate a control functional $\psi_{\boldsymbol{\phi}}$ such that $\mathbb{E}_{\mathcal{D}_0}\left[\sigma^2_{\mathcal{D}_0}\right] \to 0$ as $m \to \infty$. In doing so, we gain some insight into the exact nature of the non-linear regression problem (Eqn. 5). We focus here on the scalar case $\mathcal{X} \subseteq \mathbb{R}$.

Consider a generic function $\phi : \mathcal{X} \to \mathbb{R}$ that may be decomposed using a suitable orthogonal basis as $\phi(x) = \sum_{k=0}^{\infty} a_k e_k(x)$. Plugging this expression into Eqn. 5 leads to the regression problem

$$f(x) = \mu + \sum_{k=0}^{\infty} a_k \left\{e_k'(x) + e_k(x)u(x)\right\} + \epsilon_\phi(x) \tag{46}$$

where we know from Sec. 2.4.1 that there exists coefficients $a_k \in \mathbb{R}$ such that $\epsilon_\phi(x) = 0$ exactly. A naive strategy suggested by this expression is to truncate the above sum to a fixed number of terms l and perform a linear regression of the f_i on the $l+1$ covariates $\{e_k'(x_i) + e_k(x_i)u(x_i) : 0 \leq k \leq l\}$. This is exactly equivalent to classical control variates, as each covariate has expectation zero under π. But, unless f happens to be a linear combination of these $(l+1)$ functions, one cannot hope to obtain a presistent estimator.

To obtain presistency, one may instead have l grow with the sample size m, say $l(m) = m^\alpha$ where $\alpha < 1$. To make this statement more concrete, assume $\mathcal{X} = [0, 1]$ and $\pi(x) = 1$ is the uniform distribution, so that $u(x) \equiv 0$. Further take $e_k(x) = \sin(2\pi kx)$. Then one may rewrite Eqn. 46 as

$$f(x) = \mu + \sum_{k=1}^{\infty} b_k \cos(2\pi kx) \tag{47}$$

for $b_k = ka_k$, and the a_k's chosen so that $\epsilon_\phi(x) = 0$ in Eqn. 46. This leads to a direct connection with non-linear regression using orthogonal bases (Wasserman, 2006, Chap. 8). Note however the non-standard nature our regression problem: In non-linear regression, it is usually assumed that the f_i's are observed with noise and the x_i's correspond to a fixed design. In our case, the f_i's are observed exactly and the x_i's are random.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

Still, if we essentially follow the same type of approach as in the standard case, we can truncate the sum above to $l(m)$, estimate the $l(m)$ regression coefficients as $\hat{b}_k = m^{-1}\sum_{i=1}^{m} f_i \cos(2\pi k x_i)$ and obtain, from Parseval's identity,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_0}\left[\sigma^2_{\mathcal{D}_0}\right] &= \mathbb{E}_X\mathbb{E}_{\mathcal{D}_0}\left[\left(f(X) - \sum_{k=0}^{l(m)} \hat{b}_k(\mathcal{D}_0)\cos(2\pi k X)\right)^2\right] \\ &= \sum_{k=l(m)+1}^{\infty} b_k^2 + \sum_{k=0}^{l(m)} \mathbb{E}_{\mathcal{D}_0}\left[\left(b_k - \hat{b}_k(\mathcal{D}_0)\right)^2\right] \qquad (48)\end{aligned}$$

where the first term is $O(l(m)^{-2\beta})$, β being related to the regularity of f (i.e. the order of the Sobolev space to which f belongs; see e.g. p.148 and Lemma 8.4 of Wasserman (2006)), and the second term is (again, under appropriate regularity conditions, e.g. f is bounded) $O(l(m)/m) = O(m^{-(1-\alpha)})$. Taking $\alpha = 1/(2\beta+1)$, one obtains that $\mathbb{E}_{\mathcal{D}_0}[\sigma^2_{\mathcal{D}_0}] \to 0$ at rate $O(m^{-\gamma})$ where $\gamma = 2\beta/(2\beta+1)$ is the standard non-parametric rate in dimension one. Thus, in the end, the non-standard (noiseless and based on random x_i) nature of our functional regression problem has no bearing on the actual performance of the estimator.

Of course, our presistency result here is quite specialised, but it does suggest that achieving presistency is not an unattainable task. It also suggests a more general (i.e. not specific to π being uniform) approach in practice. Starting again from the decomposition $\phi(x) = \sum_{k=0}^{\infty} a_k e_k(x)$ on an orthogonal basis, one may truncate the sum in (46) to some $l(m) = m^{\alpha}$ and estimate the a_k's using ordinary least squares. Alternatively one may choose l by minimising some variable selection criterion, as is commonly done in this context. In the main text we concentrate on linear smoothers as these provide an elegant and, in our experience, very effective solution to the functional regression problem.

References

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.

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

Birnbaum, A. (1962) On the foundations of statistical inference.*J. Am. Stat. Assoc.*, **57**, 269-326.

Caflisch, R. E. (1998) Monte Carlo and Quasi-Monte Carlo Methods. *Acta Numerica*, **7**, 1-49.

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

Cartwright, J., Eguiluz, V., Hernandez-Garcia, E. and Piro, O. (1999) Dynamics of elastic excitable media. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.*, **9**, 2197-2202.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

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.

Diaconis, P. (1988) Bayesian numerical analysis. In: *S. Gupta J. Berger, editors, Statistical Decision Theory and Related Topics IV, Volume 1* (pp 163-175). Springer-Verlag, New York.

FitzHugh, R. (1961) Impulses and physiological states in theoretical models of nerve membranes. *Biophysics J.*, **1**, 445-466.

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. (2014) Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods. *CRiSM Working Paper Series, University of Warwick* **14**, 19.

Giles, M. (2008) Improved multilevel Monte Carlo convergence using the Milstein scheme. In: *Monte Carlo and quasi-Monte Carlo methods* (pp. 343-358). Springer Berlin Heidelberg.

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.

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.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Heinrich, S. (2001) Multilevel Monte Carlo methods. In: *Large-scale scientific computing* (pp. 58-67). Springer Berlin Heidelberg.

Kendall, P. C. and Bourne, D. E. (1992) *Vector analysis and Cartesian tensors* (3rd ed.). CRC Press, Florida.

Kendall, W. S., Marin, J.-M. and Robert, C. P. (2007) Confidence bands for Brownian motion and applications to Monte Carlo simulations. *Stat. Comput.*, **17**, 1-10.

Lawrence, N.D., Sanguinetti, G. and Rattray, M. (2007) Modelling transcriptional regulation using Gaussian processes. *Adv. Neur. In.*, **19**, 785-792.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *J. R. Statist. Soc. B*, **72**, 417-473.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

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.

Niederreiter, H. (1978) Quasi-Monte Carlo methods and pseudo-random numbers. *B. Am. Math. Soc.*, **84**, 957-1041.

O'Hagan, A. (1987) Monte Carlo is fundamentally unsound. *Statistician*, **36**, 247-249.

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

O'Hagan, A. (1992) Some Bayesian Numerical Analysis. *Bayesian Statistics*, **4**, 345-363.

Oakley, J. and O'Hagan, A. (2002) Bayesian Inference for the Uncertainty Distribution of Computer Model Outputs. *Biometrika*, **89**, 769-784.

Oates, C. J., Papamarkou, T. and Girolami, M. (2014) The Controlled Thermodynamic Integral for Bayesian Model Comparison. *CRiSM Working Paper Series, University of Warwick*, **14**, 9.

Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C.E., Roberts, S.J. and Ghahramani, Z. (2012) Active learning of model evidence using Bayesian quadrature. *Adv. Neur. In.*, **26**, 46-54.

Owen, A. B. (1997) Scramble net variance for integrals of smooth functions. *Ann. Stat.*, **25**, 1541-1562.

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

Papamarkou, T., Mira, A. and Girolami, M. (2014b) Monte Carlo Methods and Zero Variance Principle. In: *Current Trends in Bayesian Methodology with Applications*, in press.

Pereyra, M. (2013) Proximal Markov chain Monte Carlo algorithms. arXiv:1306.0187.

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

Philippe, A. and Robert, C. (2001) Riemann sums for MCMC estimation and convergence monitoring. *Stat. Comput.*, **11**, 103-115.

Quiñonero-Candela, J., Rasmussen, C. E. and Williams, C. K. (2007) Approximation methods for Gaussian process regression. In: *Large-scale kernel machines* (eds. Bottou, L., Chapelle, O., DeCoste, D. and Weston, J., pp.203-223), MIT Press.

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.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

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.

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

van der Pol, B. (1926) On relaxation-oscillations. *The London, Edinburgh and Dublin Phil. Mag. & J. of Sci.*, **2**, 978-992.

van der Vaart, A. and van Zanten, H. (2011) Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.*, **12**, 2095-2119.

Vyshemirsky, V. and Girolami, M. A. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833-839.

Wasserman, L. (2006) *All of nonparametric statistics*. Springer, New York.

Wasserman, L. (2013) Consistency, Sparsistency and Presistency. Blog post: http://normaldeviate.wordpress.com/2013/09/11/consistency-sparsistency-and-presistency/

Wheeler, M. W., Dunson, D. B., Pandalai, S. P., Baker, B. A. and Herring, A.H. (2014) Mechanistic Hierarchical Gaussian Processes. *J. Am. Stat. Assoc.*, to appear.

Williams, C. K. and Vivarelli, F. (2000) Upper and lower bounds on the learning curve for Gaussian processes. *Mach. Learn.*, **40**, 77-102.

Zhou, Y., Johansen, A. M. and Aston J. A. D. (2013) Towards Automatic Model Comparison an Adaptive Sequential Monte Carlo Approach. *CRiSM Working Paper, University of Warwick*, **13**, 4.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism

Supplement to "Control functionals for Monte Carlo integration"

A simple case study

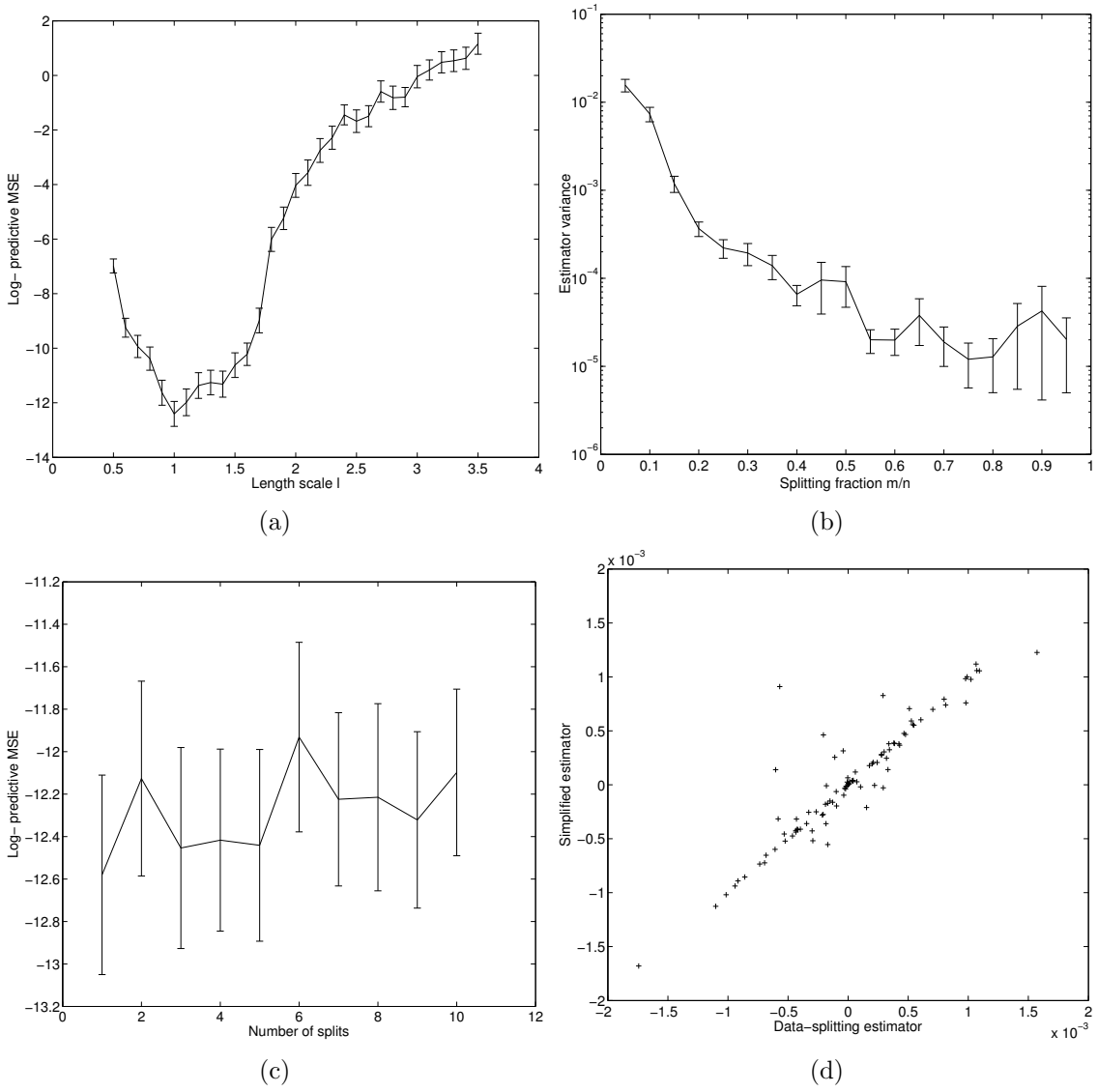
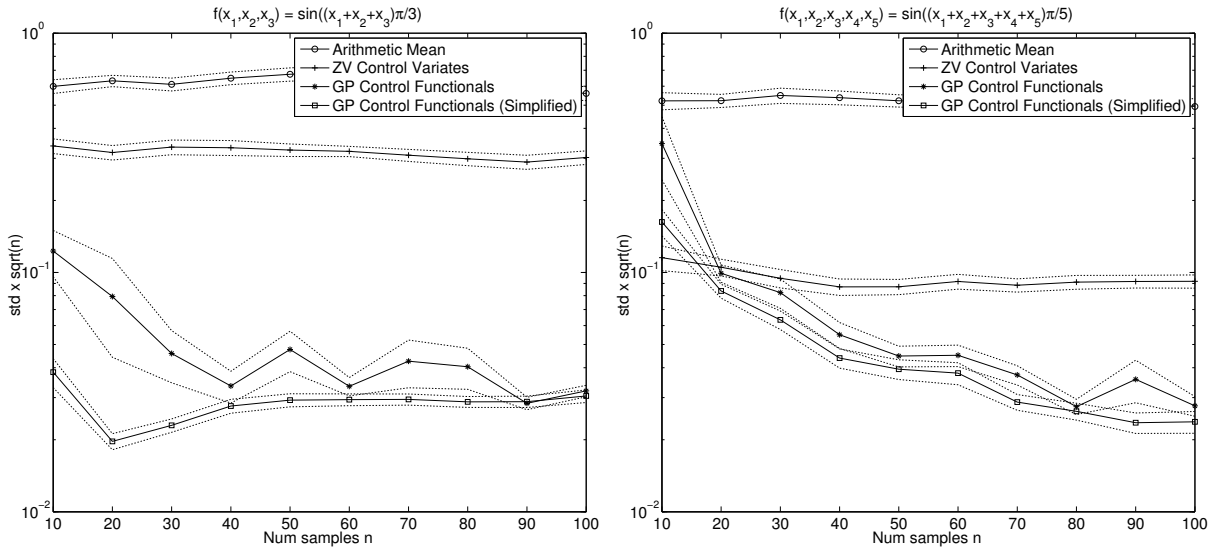


Figure 1: Simple sine/Gaussian case study. (a) Determining the GP covariance length-scale hyper-parameter $\ell \approx 1$ by minimising the predictive mean square error (MSE). (b) Determining an appropriate ratio $m/n \approx 0.8$ of training to test samples, again by consideration of the predictive MSE. (c) Examining the effect of using multiple, randomly selected sample splits. (d) Comparing the sample-splitting estimator with the simplified estimator. [Here the sample-splitting estimator uses just one split of the samples.] In each of (a-d) we considered 100 independent realisations of the sampling process. The number of samples was taken to be $n = 50$ throughout.

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism



(a) Relative variance of estimators ($d = 3$) (b) Relative variance of estimators ($d = 5$)

Figure 2: Simple sine/Gaussian case study (continued). Empirical assessment of asymptotic properties. Here we consider examples with dimension (a) $d = 3$, (b) $d = 5$.

Marginalisation in hierarchical models

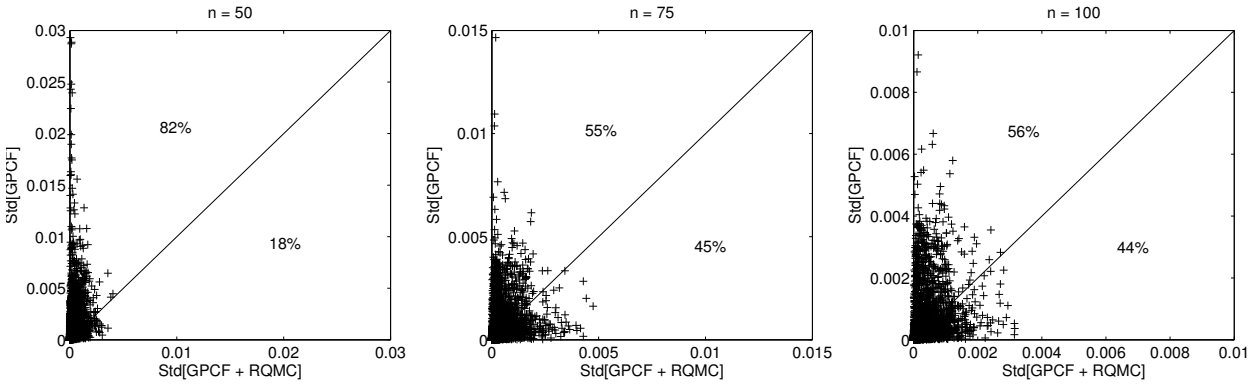


Figure 3: Exploring the use of randomised quasi Monte Carlo (RQMC) with Gaussian process control functionals (GPCF). [Here we display the sampling standard deviation of Monte Carlo estimators for the posterior predictive mean $\mathbb{E}[Y_*|\boldsymbol{y}]$ in the SARCOS robot arm example, computed over 10 independent realisations. Each point, representing one Monte Carlo integration problem, is represented by a cross. Percentages indicate the number of crosses above/below the diagonal.]

CRiSM Paper No. 14-22, www.warwick.ac.uk/go/crism