Control Functionals for Monte Carlo Integration

by Chris J. Oates, Mark Girolami, and Nicolas Chopin

Medha Agarwal

Department of Statistics, University of Washington Seattle, WA

June 10, 2022

Abstract

This report explores the variance reduction technique called control functionals (CF) proposed by Oates et al. [2017]. The problem of high precision Monte Carlo estimation is at the heart of statistics, wherein variance reduction techniques like control variates have provided simple yet effective solutions. However, challenges like slow convergence, incompatibility with unnormalized target distribution, and biasedness arise with almost all existing methods. Oates et al. [2017] propose CF method which can be intuitively understood as a non-parametric development of control variates. While solving the major misspecification issue with control variates, CF also outperforms other contemporary methods by displaying super-$\sqrt{n}$ convergence, unbiasedness, compatibility with unnormalized target, and *post-hoc* implementation. The paper provides theoretical results on convergence rates of CF and presents a comparative study of CF's performance through numerical examples on real data. This report reproduces all theoretical and empirical results. A possible methodological extension is also discussed at the end.

1 Introduction

Estimating the expectation of measurable functions efficiently and accurately is the crucial task that drives research in Monte Carlo algorithms. While getting samples that represent the true target distribution well is one active area of research in Monte Carlo, researchers are increasingly interested in constructing estimators that converge to true mean faster than sample averages.

Suppose we are interested in estimating the expectation $\mu(f) = \int_\Omega f(x)\pi(x)dx$ where f is the test function and π is the target density function of random variables X that take value in $\Omega \subseteq \mathbb{R}^d$.

Typical sample average estimator using n IID random draws $\{X_i\}_{i=1}^n$ from π is of the form

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\,.$$

Using central limit theorem, we know that $\hat{\mu}$ has an error variance of $\sigma(f)^2/n$ where $\sigma^2(f)$ is the variance of f under π. Therefore, $\hat{\mu}$ converges to $\mu(f)$ at the rate of $O_P(\sqrt{n})$. The idea behind variance reduction methods is to construct estimators that have lower σ. This has given rise to volumes of literature on variance reduction techniques like antithetic variables, Riemann sums, stratified sampling, control variates, importance sampling, and a plethora of other sophisticated techniques. See chapter 8 in Owen [2013], chapter 4 in Robert et al. [1999], and chapter 5 in Rubinstein and Kroese [2016] for a detailed discussion.

Oates et al. [2017] suggest the following four desiderata for evaluating any variance reduction method - 1) unbiased estimation, 2) compatibility with unnormalized target density π, 3) super-$\sqrt{n}$ convergence (faster than $\sqrt{n}$-convergence), and 4) *post-hoc* scheme. The last refer to schemes that can be retrospectively applied after the samples have been received. Unfortunately, no existing method satisfies all four properties. The authors summarize all the major variance reduction technique along with where they fall on the four desiderata in Table 1.

Estimation Method	Unbiased	Unnormalized π	Super $\sqrt{n}$	Post hoc
MC (MCMC) + arithmetic mean	✓(×)	×(✓)	×	×
MC + importance sampling	✓(×)	×(✓)	×	×
MC + antithetic variables	✓	×	×	×
Quasi MC	×	×	✓	×
Randomized quasi MC	✓	×	✓	×
MC (/MCMC) + Rao-Blackwellization	✓(×)	×(✓)	×	✓
MC (/MCMC) + control variates	✓(×)	×(✓)	×	✓
MC (/MCMC) + Riemann sums	×	×(✓)	✓	✓
Bayesian MC	×	×	✓	✓
MC (/MCMC) + control functionals	✓(×)	×(✓)	✓	✓

Table 1: Estimator properties based on the four desiderata. The properties change when handling an unnormalized π and have been shown in parenthesis.

Control functionals (CF) is a powerful method proposed by Oates et al. [2017] that can be intuitively understood as a non-parametric extension of control variates (CV) [Mira et al., 2013,

Rubinstein and Marcus, 1985]. It utilizes the gradient information of target density to achieve variance reduction and satisfies all four of the above properties. The authors present theoretical convergence guarantees and also non-asymptotic error bounds. A thorough empirical performance analysis is carried out on real data. This includes Bayesian hierarchical models and estimation of model evidence using thermodynamic integration on non-linear differential equations models. The main competitor for a comparative performance assessment of CF is the zero-variance control variate [Mira et al., 2013] method described in next section. The authors have suggested to make use of a follow-up `R` package `ZVCV` implementing CF and ZV-CV.

2 Literature review

In this section, properties of variance reduction techniques that offer super-$\sqrt{n}$ convergence are briefly discussed. A note on CV is also presented discussing how almost all CV methods suffer from the issue of missepcification. This helps us motivate how it is ameliorated by CF, particularly in a high dimensional setting.

2.1 Methods that guarantee super-$\sqrt{n}$ convergence

Super-$\sqrt{n}$ convergence is generally achieved either using Rao-Blackwellization type argument where first-level convergent approximation of integrand is constructed or using a QMC type argument where we replace the random points by low discrepancy point sets. Existing methods that do satisfy super-$\sqrt{n}$ convergence like (R)QMC [Dick and Pillichshammer, 2014] and Bayesian Monte Carlo (BMC) [Briol et al., 2015, Rasmussen and Ghahramani, 2003] are not amenable to unnormalized target density. This is troublesome, especially in Bayesian paradigm, where posterior expectations are often desirable but posterior density is only known up to a normalizing constant. Although, note that BMC, like importance sampling, can use samples from any unnormalized proposal distribution. The point of deviation is that the target density π should be exactly evaluated at those points. Briol et al. [2015] show that BMC displays super-$\sqrt{n}$ convergence and is only asymptotically unbiased. In later section, we will see that the CF method proposed by the authors is equivalent to BMC plus an additional bias correction term. It is interesting how CF and BMC converge to the same estimator, despite originating from two different paradigms. Philippe [1997] show that Riemann sums, while converging at super-$\sqrt{n}$ rate, give a biased estimator with a $O(n^{-1})$ bias and are computationally prohibitive with increasing dimensions.

2.2 Background and prior work on control variates

In control variates, we search for a function h such that expectation of h with respect to target density is known and h approximates f closely. The surrogate function $\tilde{f}(x) = f(x) - h(x) + \mathbb{E}_\pi[h(X)]$ is constructed and used for Monte Carlo instead of f. Note that $\mathbb{E}[\tilde{f}] = \mu$ and $\text{Var}(\tilde{f}) = \text{Var}(f - h)$. Variance reduction is obtained when there is heavy correlation between f and h. Many methods for approximating f have been suggested in literature, most of which seek to approximate f by a linear combination of some basis statistics $\{s_i\}_{i=1}^m$, i.e. $h = \sum_{i=1}^m a_i s_i$ for $a_i \in \mathbb{R}$ [Andradóttir et al., 1993, Assaraf and Caffarel, 1999, Mira et al., 2013]. This inherently involves solving a regression problem for the coefficients $\{a_i\}_{i=1}^m$.

Assaraf and Caffarel [1999] propose a zero variance principle for Monte Carlo integration of observables that seeks a Hermitian operator H and an integrable function ψ such that $\int H(x,y)\sqrt{\pi(y)}dy = 0$ and the tuple (H, ψ) satisfies the fundamental equation of zero variance

$$\nabla_{\mathbf{x}} \cdot \{\pi(\mathbf{x})\phi(\mathbf{x})\} = \{f(\mathbf{x}) - \mu(f)\}\pi(\mathbf{x})\,. \tag{1}$$

The surrogate is then constructed as $\tilde{f} = f + \int H(x,y)\psi(y)dy/\sqrt{\pi(x)}$. The choice of H and ψ such that it satisfies (at least approximately) (1) determines the variance reduction obtained in this setup. Banking on this general framework, Mira et al. [2013] propose zero-variance control variate (ZV-CV) estimator wherein H is the Hamiltonian operator and ψ is restricted to belong to a low-degree polynomial function class. The coefficients that parametrize ψ are optimized for minimizing $\sigma^2(\tilde{f})$ by solving a regression problem. Oates et al. [2016] extend the ZV-CV approach to estimators of model evidence using thermodynamic integration.

The key issue in control variates is that it often solves a misspecified regression problem because f does not necessarily belong to the linear span of $\{s_i\}_{i=1}^m$. Mijatović and Vogrinc [2018] alleviate this problem by increasing m with n so that $\{s_i\}_{i=1}^m$ is dense with respect to the function space that contains f. However, while being an excellent non-parametric approach for approximating f, their method does not work with unnormalized π. CF is a non-parametric development of CV that heals this misspecification issue by finding the approximation of f inside a reproducing kernel Hilbert space (RKHS) as we will see in the next section.

3 Control Functionals

The main idea in control functionals is to - 1) construct a gradient-based RKHS $\mathcal{H}_+$ such that the expectation of each function belonging to this space is analytically tractable, and 2) show that a close approximation of f belonging to this space is available in closed form. While each element of $\mathcal{H}_+$ is a valid control functional, the authors show that if the constructed space is large enough, the problem is well posed, i.e. $f \in \mathcal{H}_+$. Finding f in $\mathcal{H}_+$ is cast as an optimization problem. Hereon, throughout the text, relevant references to the main paper are mentioned in italicized parenthesis.

3.1 Problem Setup

Let $\mathcal{L}^2(\pi)$ denote the space of measurable functions that have finite L_2 norm with respect to the Lebesgue measure π. In addition, $C^k(\Omega, \mathbb{R}^j)$ denotes the space of measurable functions $g : \Omega \to \mathbb{R}^j$ with continuous partial derivatives up to order k. Recall that $\mu(f) := \int_\Omega f(\mathbf{x})\pi(\mathbf{x})dx$ and $\sigma^2(f) = \int_\Omega (f(\mathbf{x}) - \mu(f))^2\pi(\mathbf{x})d\mathbf{x}$. Another defining assumption of the method is that π is continuously differentiable everywhere on Ω. For n states $\mathcal{D} := \{\mathbf{x}\}_{i=1}^n$, it is assumed that the gradients $\{\nabla_x \log \pi(\mathbf{x}_i)\}$ are calculated and stored.

First, the available states are split into two disjoint sets $\mathcal{D}_0 = \{\mathbf{x}\}_{i=1}^m$ and $\mathcal{D}_1 = \{\mathbf{x}\}_{i=m+1}^n$ where $1 \leq m < n$. $\mathcal{D}_1$ is assumed to have IID samples from π independent from $\mathcal{D}_0$. The surrogate function uses an approximation of f constructed solely from $\mathcal{D}_0$ and denoted by $s_{f,\mathcal{D}_0}$ giving the surrogate function $f_{\mathcal{D}_0}(\mathbf{x}) := f(\mathbf{x}) - s_{f,\mathcal{D}_0}(\mathbf{x}) + \mu(s_{f,\mathcal{D}_0})$. We will show that upon construction of $s_{f,\mathcal{D}_0}$, $f_{\mathcal{D}_0} \in \mathcal{L}^2(\pi)$ and the final estimator of μ is of the form

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

For $m < n$, conditioned on $\mathcal{D}_0$, we have $\mathbb{E}_{\mathcal{D}_1}[\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)] = \mu(f)$ and $\mathrm{Var}_{\mathcal{D}_1}(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)) = (n-m)^{-1}\sigma^2(f - s_{f,\mathcal{D}_0})$. The expectations here are taken with respect to the distribution of $n-m$ samples in $\mathcal{D}_1$.

Note that this looks very familiar to traditional control variates approach except for the split in states. The key point of deviation from CV that also helps us achieve super-$\sqrt{n}$ convergence is assuming $m = O(n^\gamma)$ which ensures that m increases with n as $n \to \infty$. Further, if we assume that the approximation $s_{f,\mathcal{D}_0}$ is good such that $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(m^{-\delta})$ for some $\delta \geq 0$, then $\mathbb{E}_{\mathcal{D}_0}[\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f))^2]] = O(n^{-1-\gamma\delta})$ *(Proposition 1)*. This forms the main result of the

paper claiming super-$\sqrt{n}$ convergence. The rest of the methodology section is devoted to obtaining such function $s_{f,\mathcal{D}_0}$ belonging to a carefully constructed RKHS. Oates et al. [2017] derived the main form of $s_{f,\mathcal{D}_0}$ from application of Stein's operator on a function $\phi \in C^1(\Omega, \mathbb{R}^d)$.

3.2 Control variates based on Stein operator

Recall the zero-variance set-up proposed by Assaraf and Caffarel [1999] that seeks a Hermitian operator H and an integrable function ψ such that they satisfy the fundamental equation of zero-variance principle given by (1). Oates et al. [2017] choose Stein operator for deriving CF because it will help furnish a mean-zero RKHS for ψ. Assuming that the density π belongs to $C^1(\Omega, \mathbb{R})$ *(Assumption 1)*, let $\mathbf{u}(\mathbf{x}) := \nabla_{\mathbf{x}} \log\{\pi(\mathbf{x})\}$. The Stein operator is defined as

$$\begin{aligned} \mathbb{S}_\pi : C^1(\Omega, \mathbb{R}^d) &\to C^0(\Omega, \mathbb{R}) \\ \phi(.) &\mapsto \mathbb{S}_\pi(\phi)(.) := \nabla_{\mathbf{x}} \cdot \phi(.) + \phi(.) \cdot \nabla_{\mathbf{x}} \log \pi(.)\,. \end{aligned}$$

For a choice of ϕ that will be described later, let $\psi := \mathbb{S}_\pi(\phi)$, then the surrogate function is defined as $s_{f,\mathcal{D}_0} = c + \psi(x)$ where c is a constant. The strength of choosing this form of surrogate lies in the fact that under a regularity condition, $\mu(\psi) = 0$. Therefore, ψ is recognized as a control variate. Let $\partial\Omega$ denotes the boundary of the state space Ω. For $\mathbf{x} \in \partial\Omega$, $\mathbf{n}(\mathbf{x})$ denotes the unit normal to $\partial\Omega$. It is assumed that *(Assumption 2)*

$$\oint_{\partial\Omega} \pi(\mathbf{x})\phi(\mathbf{x}) \cdot \mathbf{n}((\mathbf{x})S(d\mathbf{x}) = 0\,,$$

where $\oint_{\partial\Omega}$ denotes the surface integral over $\partial\Omega$ and $S(d\mathbf{x})$ denotes the surface element at $\mathbf{x} \in \partial\Omega$. Under this assumption and that $\pi \in C^1(\Omega, \mathbb{R})$, we can use the divergence theorem to obtain that

$$\mu(\psi) = \int_\Omega \psi(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \int_\Omega \nabla_{\mathbf{x}}\{\phi(\mathbf{x})\pi(\mathbf{x})\}d\mathbf{x} = \oint_{\partial\Omega} \pi(\mathbf{x})\phi(\mathbf{x}) \cdot \mathbf{n}((\mathbf{x})S(d\mathbf{x}) = 0\,,$$

which implies that $\mu(s_{f,\mathcal{D}_0}) = c$ *(Proposition 2)*. The choice of c and ϕ completely determines the control variate. Note that ϕ is chosen as a low-degree polynomial in Assaraf and Caffarel [1999] and Mira et al. [2013]. This leads to linear regression problem wherein the number of covariates increases with increasing dimensions of the target. South et al. [2018] propose a penalized regression technique for high dimensional cases that leads to some performance gains. However, in general, restricting ϕ to a finite dimensional parametric family leads to the misspecification issue. This paper enables a fully non-parametric approximation of f by optimizing c and ϕ over a carefully constructed Hilbert space described in the next subsection.

3.3 Constructing reproducing kernel Hilbert spaces

Recall the approximation $s_{f,\mathcal{D}_0}(\mathbf{x}) = c + \psi = c + \nabla_{\mathbf{x}} \cdot \phi(\mathbf{x}) + \phi(\mathbf{x}) \cdot u(\mathbf{x})$, where $c \in \mathbb{R}$ is a constant. To summarize the key idea, if ϕ is allowed to belong to a RKHS, then we can show that ψ also belongs to a RKHS. Since $\mu(\psi) = 0$, all elements of this RKHS are legitimate control functionals. The exact form of c and ψ is then obtained by taking the projection of f on this RKHS. Let each component ϕ_i of the function ϕ belongs to a RKHS $\mathcal{H}$ with a kernel $k : \Omega \times \Omega \to \mathbb{R}$, i.e.

1. for all $\mathbf{x} \in \Omega$, we have $k(\cdot, \mathbf{x}) \in \mathcal{H}$, and
2. for all $\mathbf{x} \in \Omega$ and $h \in \mathcal{H}$ we have $h(\mathbf{x}) = \langle h, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$.

It can be shown that the Cartesian product space $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$ is also a Hilbert space with inner product $\langle \phi_1, \phi_2 \rangle_{\mathcal{H}^d} = \sum_{i=1}^{d} \langle \phi_{1i}, \phi_{2i} \rangle_{\mathcal{H}}$. The function ϕ now belongs to $\mathcal{H}^d$. It is assumed that $\mathcal{H}$ is constructed only using k that belong to $C^2(\Omega \times \Omega, \mathbb{R})$ *(Assumption 3)* and which satisfies the following two conditions *(Assumption 2')*

$$\oint_{\partial\Omega} k(\mathbf{x}, \mathbf{x}')\pi(\mathbf{x}')\mathbf{n}((\mathbf{x}')S(d\mathbf{x}') = 0 \quad \text{and} \quad \oint_{\partial\Omega} \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')\pi(\mathbf{x}') \cdot \mathbf{n}((\mathbf{x}')S(d\mathbf{x}') = 0\,. \tag{2}$$

Under these assumptions, it can be shown that $\psi = \nabla_{\mathbf{x}} \cdot \phi(\mathbf{x}) + \phi(\mathbf{x}) \cdot u(\mathbf{x})$ belongs to a reproducing kernel Hilbert space $\mathcal{H}_0$ with kernel

$$k_0(\mathbf{x}, \mathbf{x}') := \nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + u(\mathbf{x}) \cdot \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + u(\mathbf{x}') \cdot \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}') + u(\mathbf{x}) \cdot u(\mathbf{x}') k(\mathbf{x}, \mathbf{x}')\,. \tag{3}$$

Using the assumptions that $\pi \in C^1(\Omega, \mathbb{R})$, $k \in C^2(\Omega \times \Omega, \mathbb{R})$, and the surface integral conditions in (2), we can use divergence theorem again to show that $\int_\Omega k(\mathbf{x}, \mathbf{x}')\pi(\mathbf{x}')d\mathbf{x}' = 0$ *(Lemma 1)* for π almost all $\mathbf{x} \in \Omega$. As a consequence, $\mathcal{H}_0$ contains only valid control functionals. In order to even have a possibility of achieving variance reduction by approximating f by elements in $\mathcal{H}_0$, it is necessary that $\mathcal{H}_0 \subset \mathcal{L}^2(\pi)$. This can be done by ensuring that $\int_\Omega k_0(\mathbf{x}, \mathbf{x})\pi(\mathbf{x})d\mathbf{x} < \infty$ *(Assumption 4)*. The reproducing property of RKHS followed by Cauchy-Schwartz inequality gives $|\psi(\mathbf{x})| = |\langle \psi, k_0(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_0}| \leq \|\psi\|_{\mathcal{H}_0} \|k_0(\cdot, \mathbf{x})\|_{\mathcal{H}_0}$, and using the reproducing property again $\|k_0(\cdot, \mathbf{x})\|^2_{\mathcal{H}_0} = k_0(\mathbf{x}, \mathbf{x})$. Now the Assumption 4 in paper allows us to write

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

The authors verified that the assumption holds for their choice of k. Therefore, we have successfully created a RKHS of mean zero functions.

3.4 Consistent approximation

The aim is to consistently approximate f by an optimal $s_{f,\mathcal{D}_0} = c + \psi$ where $c \in \mathbb{R}$ and $\psi \in \mathcal{H}_0$. Let $\mathcal{C}$ denote the reproducing kernel Hilbert space of constant functions with kernel $k_{\mathcal{C}}(\mathbf{x}, \mathbf{x}') = 1$ for all $\mathbf{x}, \mathbf{x}' \in \Omega$ and equipped with norm $\|\cdot\|_{\mathcal{C}}$. Let $\|\cdot\|_{\mathcal{H}_0}$ be the norm associated with $\mathcal{H}_0$. Define $\mathcal{H}_+ := \mathcal{C} + \mathcal{H}_0 = \{c + \psi : c \in \mathcal{C}, \psi \in \mathcal{H}_0\}$. For any $h \in \mathcal{H}_+$, the vector space $\mathcal{H}_+$ is equipped with norm $\|h\|^2_{\mathcal{H}_+} := \|c\|^2_{\mathcal{C}} + \|\psi\|^2_{\mathcal{H}_0}$ which is well-defined because $h = c + \psi$ is the unique representation. Let $h = c + \psi$ and $h' = c' + \psi'$ for $c, c' \in \mathcal{C}$ and $\psi, \psi' \in \mathcal{H}_0$. Then $h + h' = (c + c') + (\psi + \psi')$ which implies that $h + h' \in \mathcal{H}_+$ and for $\lambda \in \mathbb{R}$, $\lambda h = \lambda c + \lambda\psi$ which implies that $\lambda h \in \mathcal{H}_+$. This implies that $\mathcal{H}_+$, equipped with the norm $\|\cdot\|_{\mathcal{H}_+}$, is a RKHS with kernel $k_+(\mathbf{x}, \mathbf{x}') := k_{\mathcal{C}}(\mathbf{x}, \mathbf{x}') + k_{\mathcal{H}_0}(\mathbf{x}, \mathbf{x}')$.

It is assumed that the problem is well-posed, i.e. $f \in \mathcal{H}_+$. This is a strong assumption but it is certainly weaker than other CV methods discussed earlier that restrict f to some finite dimensional parametric class. Following the notation introduced by Assaraf and Caffarel [1999], the task at hand is to find a $\phi \in \mathcal{H}^d$ such that the tuple $(\mathbb{S}_\pi, \phi)$ satisfy the fundamental equation of zero variance in (1). Since this fundamental equation is underdefined, the well-posedness can be instantly satisfied by choosing k such that $\mathcal{H}^d$ is big enough to contain at least one solution.

3.5 Obtaining solution in closed form

The optimization problem of obtaining the $s_{f,\mathcal{D}_0}$ belonging to $\mathcal{H}_+$ solved using the regularized least squares (RLS) functional approximation, i.e.

$$s_{f,\mathcal{D}_0} = \underset{g \in \mathcal{H}_+}{\arg\min} \left[.\frac{1}{m} \sum_{j=1}^{m} f(\mathbf{x}_j) - g(\mathbf{x}_j)^2 + \lambda \|g\|^2_{\mathcal{H}_+} \right], \tag{4}$$

where $\lambda > 0$. Recall that finding the optimal $g \in \mathcal{H}_+$ is equivalent to finding the optimal $c \in \mathcal{C}$ and $\psi \in \mathcal{H}_0$ and the optimization problem becomes

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

Using the representor theorem [Steinwart and Christmann, 2008, theorem 5.5], keeping c constant, we know that there exists a unique solution $\hat{\psi}$ that solves the optimization problem. Additionally there exist $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_m) \in \mathbb{R}^m$ such that $\psi(\mathbf{x}) = \sum_{i=1}^m \beta_i k_0(\mathbf{x}_i, \mathbf{x})$. Denote $\mathbf{f}_0 := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))^T$, $\mathbf{1}_p = (1, \ldots, 1)^T$ for a vector of p ones, $\mathbf{f}_1 = (f(\mathbf{x}_{m+1}), \ldots, f(\mathbf{x}_n))^T$, and

$(\mathbf{K}_0)_{i,j} = k_0(\mathbf{x}_i, \mathbf{x}_j)$. The solution obtained for c and $\boldsymbol{\beta}$ after solving the system is

$$\hat{c} = \frac{\mathbf{1}_m^T \mathbf{K}_0^{-1} \mathbf{f}_0}{1 + \mathbf{1}_m^T \mathbf{K}_0^{-1} \mathbf{1}_m}.$$
$$\hat{\boldsymbol{\beta}} = \mathbf{K}_0^{-1}(\mathbf{f}_0 - \hat{c}\mathbf{1}_m)\,. \tag{5}$$

To incorporate the small effect of λ, replace $\mathbf{K}_0$ by $\mathbf{K}_0 + \lambda\mathbf{I}$ in the above solutions where $\mathbf{I}$ is the $m \times m$ identity matrix. Denoting $(\mathbf{K}_{1,0})_{i,j} := k_0(\mathbf{x}_{m+i}, \mathbf{x}_j)$ for $i \in \{1, \ldots, n-m\}$ and $j \in \{1, \ldots, m\}$, the associated fitted values for $(n-m)$ samples in $\mathcal{D}_1$ are $\hat{\mathbf{f}}_1 := \hat{c}\mathbf{1}_{n-m} + \mathbf{K}_{1,0}\hat{\boldsymbol{\beta}}$. The CF estimator obtained using this solution is *(Lemma 3)*

$$\begin{aligned}
\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \frac{1}{n-m}\sum_{i=m+1}^{n} f_{\mathcal{D}_0}(\mathbf{x}_i) &= \frac{1}{n-m}\sum_{i=m+1}^{n} f(\mathbf{x}_i) - s_{f,\mathcal{D}_0}(\mathbf{x}_i) + \mu(s_{f,\mathcal{D}_0}). \\
&= \frac{1}{n-m}\mathbf{1}_{n-m}^T(\mathbf{f}_1 - \hat{\mathbf{f}}_1) + \hat{c} \\
&= \underbrace{\frac{1}{n-m}\mathbf{1}_{n-m}^T(\mathbf{f}_1 - \hat{\mathbf{f}}_1)}_{(i)} + \underbrace{\frac{\mathbf{1}_m^T(\mathbf{K}_0 + \lambda m\mathbf{I})^{-1}\mathbf{f}_0}{1 + \mathbf{1}_m^T(\mathbf{K}_0 + \lambda m\mathbf{I})^{-1}\mathbf{1}_m}}_{(ii)}\,,
\end{aligned} \tag{6}$$

The authors point out the following interesting remarks on the final CF estimator.

Remark 1. *The estimator $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is a weighted sum of function values $\mathbf{f}_0$ and $\mathbf{f}_1$. These weights are independent of the function f and therefore supplies the major strength of the algorithm. Same set of sample $\{\mathbf{x}_i\}$ can be used to estimate the expectation of multiple test functions. This observation also motivates the possible extension to this work, discussed in Section 6.*

Remark 2. *The samples from $\mathcal{D}_1$ appears only in Term (i) of* (6)*. The authors claim that since this term vanishes in probability as $m \to \infty$, any variability due to $\mathcal{D}_1$ also vanishes, further explaining the super $\sqrt{n}$-convergence of $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$.*

Remark 3. *It is interesting to note that the Term (ii) in* (6) *is algebraically equivalent to BMC, i.e. Term (ii) is equal to the posterior mean of $\mu(f)$ based on a Gaussian prior on $f \sim \mathcal{GP}(0, k_+)$ and data $\mathcal{D}_0$ [Rasmussen and Ghahramani, 2003]. Equation 6 therefore improves the BMC by adding the bias correction term (i) and generalizing BMC to unnormalized target.*

4 Results

4.1 Super $\sqrt{n}$-convergence

By construction, of Hilbert space $\mathcal{H}_+$ and RLS optimal $s_{f,\mathcal{D}_0} \in \mathcal{H}_+$, we know that $\mathbb{E}_{\mathcal{D}_0}[s_{f,\mathcal{D}_0}] = 0$. The super $\sqrt{n}$-convergence of $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is established in two parts - 1) under a boundedness

assumption on kernel k_0, it is first established that the functional approximation produces vanishing errors, i.e. $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O_P(m^{-\delta})$ and 2) proving proposition 1 to show that $\mathbb{E}_{\mathcal{D}_0}[\mathbb{E}_{\mathcal{D}_1}[(\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f))^2]] = O(n^{-1-\gamma\delta})$ when $m = O(n^\gamma)$ for $\gamma \in [0, 1]$. Note that the convergence rate is maximized when $\gamma = 1$.

The approximation error for approximating f by an element in $\mathcal{H}_+$ has been studied extensively using the integration operator by Smale and Zhou [2004, 2005, 2007] defined as

$$Tg(\mathbf{x}) := \int_\Omega k_+(\mathbf{x}, \mathbf{x}')g(\mathbf{x}')\pi(\mathbf{x}')d\mathbf{x}', \qquad \mathbf{x} \in \Omega, g \in \mathcal{L}^2(\pi) \backslash \{f \text{ such that } f = 0 \quad \pi - a.s.\}$$

with the restricted that Ω is compact and k_+ is a Mercer kernel. The approximation error in case of a general metric space and kernel is developed in spirit of Sun and Wu [2009], Theorem 1.1 which says that if

(a) $\sup_{\mathbf{x} \in \Omega} k_+(\mathbf{x}, \mathbf{x}') < \infty$ and

(b) $T^{-1/2} f \in L^2(\pi)$,

then with $\lambda = O(m^{-1/2})$, we have $\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(m^{-1/6})$. Condition (a) can be satisfied by showing that $\sup_{\mathbf{x} \in \Omega} k_+(\mathbf{x}, \mathbf{x}') \leq 1 + \sup_{\mathbf{x} \in \Omega} k_0(\mathbf{x}, \mathbf{x}')$. Boundedness of k_0 is assumed in general and easily verified for the choice of k_0 in experiments. Condition (b) is verified using arguments in proposition 3.3 of Sun and Wu [2009] that state that $\|T^{-1/2}h\|_{L^2(\pi)} = \|h\|_{\mathcal{H}_+}$ for all $h \in \mathcal{H}_+$. Since $f \in \mathcal{H}_+$ by the well-posedness assumption, we have $\|T^{-1/2}f\|_{L^2(\pi)} = \|f\|_{\mathcal{H}_+} < \infty$ as shown in subsection 3.3.

Using the rate on approximation error, the unbiasedness of $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$, and $m = O(n)$, we can show the main result of the paper *(Theorem 2)* that establishes super $\sqrt{n}$-convergence

$$\mathbb{E}_{\mathcal{D}_0}[\mathbb{E}_{\mathcal{D}_1}[\{\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) - \mu(f)\}^2]] = (n - m)^{-1}\mathbb{E}_{\mathcal{D}_0}[\sigma^2(f - s_{f,\mathcal{D}_0})] = O(n^{-7/6})\,. \tag{7}$$

4.2 Convergence rates for estimators based on Stein's method

Explicit convergence rates are derived for CF in the follow-up paper Oates et al. [2019]. Suppose a and b are related to the smoothness levels of π and f respectively such that $\pi \in C^{a+1}(\Omega, \mathbb{R})$ and $k \in C^{b+1}(\Omega \times \Omega, \mathbb{R})$, then the authors show that CF incurs an integration error of $O(n^{-1/2-(a \wedge b)/d + \epsilon})$ where $a \wedge b := \min(a, b)$, d is the problem dimension, and $\epsilon > 0$ is an arbitrary constant.

Remark 4. *Solving the linear system for calculating $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ involves inverting the $m \times m$ matrix $\mathbf{K}_0$. The cost associated with this operation is $O(n^3)$. Therefore, for a cost c, CF achieve*

an error rate of

$$O\left(\left(c^{1/3}\right)^{-1/2-(a\wedge b)/d+\epsilon)}\right) = O(c^{-1/6-(a\wedge b)/3d+\epsilon}) \,. \tag{8}$$

On the other hand, for cost c, sample means gives an error rate of $O(c^{-1/2})$*. Therefore, it is only advisable to use CF of* $a \wedge b > d$*. Otherwise, Monte Carlo is more efficient.*

From (8), it is evident that CF too suffers from curse of dimension and perform worse than Monte Carlo when d is high. In most Bayesian applications, the desired quantity is a posterior expectation. However, independent sampling from posterior distribution is often not possible and various MCMC methods are instead employed.

Remark 5. *Oates et al. [2019] prove that the convergence rates remain the same even when the samples in* $\mathcal{D}$ *are generated using an MCMC method that supplies uniformly ergodic Markov chains.*

Remark 6. *The data split between* $\mathcal{D}_0$ *and* $\mathcal{D}_1$ *can be optimized by minimizing the convergence rate over* $\nu := m/n$*. This gives optimal split* $\nu^* = 2(a \wedge b)/d$*. Therefore, if* π *and* f *are very smooth, i.e.* $a \wedge b \gg d$*, then all the samples should be assigned to* $\mathcal{D}_0$ *and the estimator becomes equivalent to a numerical quadrature, whereas if* $a \wedge b \ll d$*, it is advisable to do plain Monte Carlo.*

5 Experiments

The authors use three different examples to evaluate the performance of control functionals. An additional illustrative example is presented with multimodal target to show how CF does not offer performance gains even under smooth multimodal targets. The kernel used throughout the examples is $k(\mathbf{x}, \mathbf{x}') = (1 + \alpha_1\|\mathbf{x}\|_2^2)^{-1}(1 + \alpha_1\|\mathbf{x}'\|_2^2)^{-1} \exp\{(-2\alpha_2)^{-1}\|\mathbf{x} - \mathbf{x}'\|_2^2\}$. The kernel parameters $\alpha = (\alpha_1, \alpha_2)$ for each example are chosen via cross-validation on $\mathcal{D}_0$. Assuming that samples in $\mathcal{D}_0$ are iid, $\mathcal{D}_0$ is randomly split into m' training samples $\mathcal{D}_{0,0}$ and $m - m'$ test samples $\mathcal{D}_{0,1}$. The parameters α such that they minimize $\|\hat{\mathbf{f}}_{(0,1)} - \mathbf{f}_{(0,1)}\|$ where $\mathbf{f}_{(0,1)}$ are the functional evaluations for samples in $\mathcal{D}_{0,1}$ and $\hat{\mathbf{f}}_{(0,1)}$ are the corresponding fitted values. Note that this cross-validation does not introduce bias as it is done only on the samples in $\mathcal{D}_0$. The cross-validated value of α reported by the authors is chosen for each example.

The authors also propose using simplified CF wherever unbiased estimation is not essential and variance reduction offered by simplified CF is more valuable. In all cases, λ is chosen such that the kernel matrix $\mathbf{K}_0 + \lambda\mathbf{I}$ has condition number lower than 10^{10}.

5.1 Illustrative Example

The goal is to estimate the expectation of $f(X) = \sin(\pi/d)\sum_{i=1}^{d} X$ where X is a d-dimensional standard Gaussian random variable. A simple scalar case, i.e. $d = 1$ is considered with n going up to 100. The authors report that CF performs best when $m/n = 0.5$, i.e $\mathcal{D}_0$ and $\mathcal{D}_1$ each get half of the samples. The cross-validated hyperparameters for the kernel are $\alpha_1 = 0.1$ and $\alpha_2 = 1$. Figure 1 displays the sampling distribution of estimators - 1) arithmetic means, 2) zero-variance control variates, 3) control functionals, and 4) simplified CF as n is increased from 10 to 100. Both CF and simplified CF show dramatic variance reduction for $d = 1$ case compared to vanilla MC and CV. The bias introduced by simplified CF seems to vanish asymptotically.

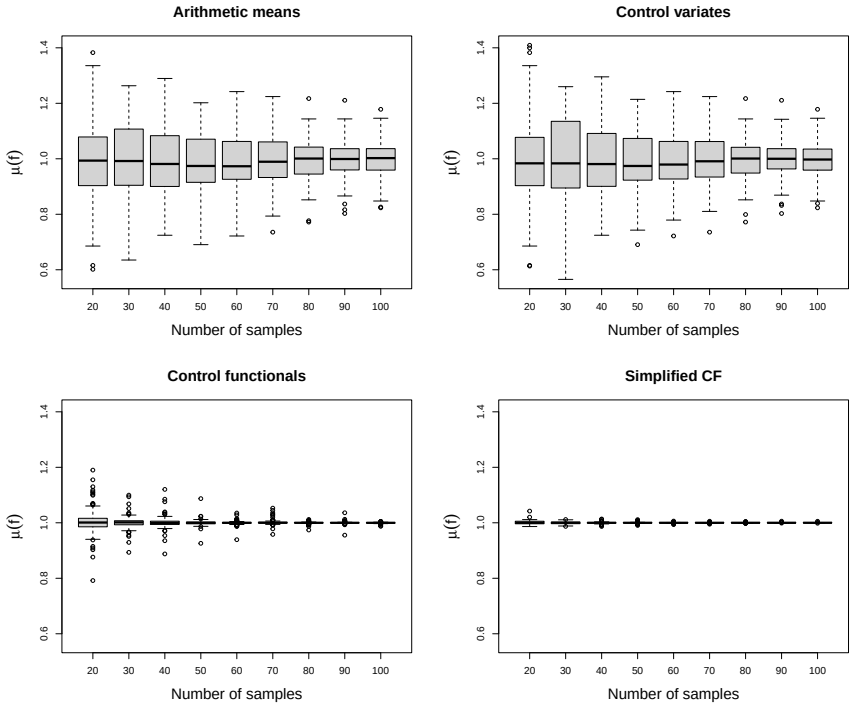

Figure 1: Sampling distribution of MC estimators based on 100 replications for $d = 1$.

To visualize the super $\sqrt{n}$-convergence, we have plotted the MSE times n versus the sample size n in Figure 2. Note that for sample means this should be approximately a horizontal line. Notice that in this synthetic example both CF and simplified CF achieve incredible super $\sqrt{n}$-convergence.

However, if we further investigate CF on this simple example for high dimensions, we see a performance worse than arithmetic means. Figure 3 shows the sampling distribution of all four estimators when $d = 10$ and n is increased from 40 to 200. Note that in this synthetic example the additional computational costs incurred by CF are not worth it. However, in following applications

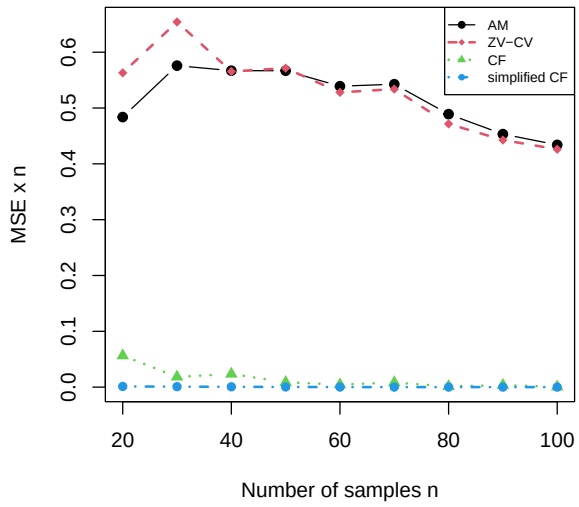

Figure 2: Estimator mean square error times n for Monte Carlo estimators.

we will see that CF offers valuable variance reduction when model complexity is borne out of difficulty in sampling from π or evaluating f.

5.2 Marginalization in hierarchical models

Marginalizing over hyper-parameters in hierarchical models is often desired in Bayesian paradigm and achieved using Monte Carlo. In this example, the authors consider a Gaussian process model and make predictions on new data by marginalizing over hyper-parameters. Consider a p-dimensional Gaussian process (GP) model for data $\{(y_i, \mathbf{z}_i)\}_{i=1}^N$ where y_i is the response variable and $\mathbf{z}_i$ is the vector of covariates. Let $\mathbf{y}$ denote the vector of all responses, i.e. $(y_1, \ldots, y_N)$. The assumed model is

$$y_i = g(\mathbf{z}_i) + \epsilon_i \qquad \text{where } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)\,,$$

and a Gaussian prior is placed on $g \sim \mathcal{GP}(0, c(\mathbf{z}, \mathbf{z}'; \theta))$ where $c(\mathbf{z}, \mathbf{z}'; \theta) = \theta_1 \exp[-\{1/2\theta_2\}^2 \|\mathbf{z}-\mathbf{z}'\|_2^2]$. Here $\theta = (\theta_1, \theta_2)$ are hyperparameters with assigned hyperpriors as $\theta_1 \sim \Gamma(\alpha, \beta)$ and $\theta_2 \sim \Gamma(\gamma, \delta)$. Let the joint distribution of θ be denoted by $\pi(\theta)$.

Suppose we wish to predict the response y_* for an unseen state $\mathbf{z}_*$. The estimator employed is a marginalized Bayesian posterior mean given by the following expression

$$\hat{y_*} = \mathbb{E}[Y_*|\mathbf{y}] = \int \mathbb{E}[Y_*|\mathbf{y}, \theta]\pi(\theta)d\theta\,.$$

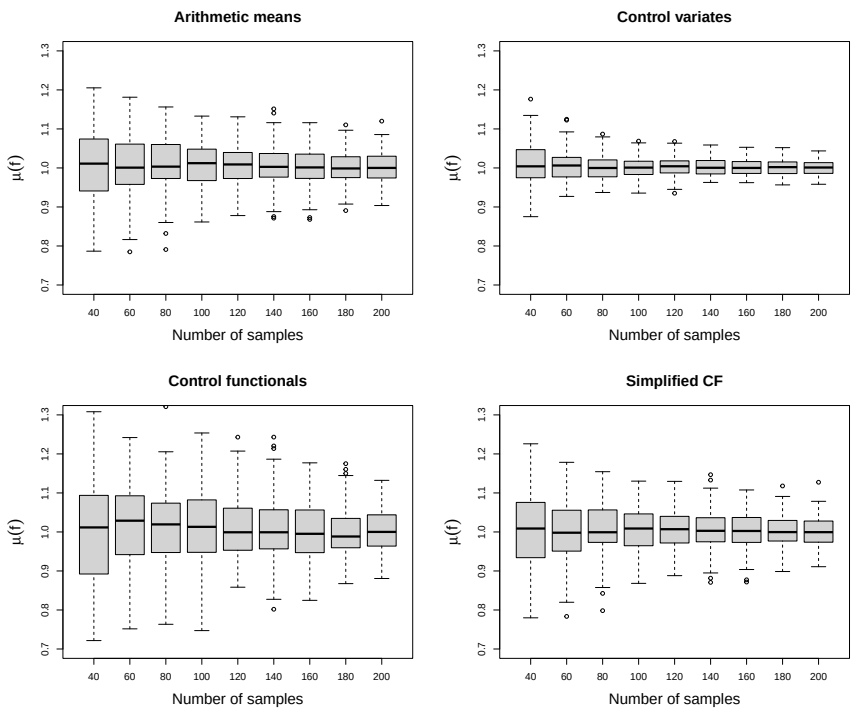

Figure 3: Sampling distribution of MC estimators based on 100 replications for $d = 10$.

Here the Monte Carlo estimator is obtained by sampling from $\pi(\theta)$. In terms of notations of the paper the integrand is

$$f(\theta) = \mathbb{E}[y_*|\mathbf{y}, \theta] = C_{*,N}(C_N + \sigma^2 I_{N\times N})^{-1}\mathbf{y}$$

where $(C_N)_{i,j} = c(\mathbf{z}_i, \mathbf{z}_j; \theta)$ and $(C_{*,N})_{1,j} = c(\mathbf{z}_*, \mathbf{z}_j; \theta)$. Here the main model complexity lies in evaluating f because evaluating f at each test point incurs a computational cost of $O(N^3)$ which can be very prohibitive considering that N is very high in real data. The authors suggest using a subset of regressors approach where we use a subset of full data, i.e. for $N' < N$

$$f(\theta) \approx C_{*,N'}(C_{N',N}C_{N,N'} + \sigma^2 C_{N'})^{-1}C_{N',N}\mathbf{y}\,. \tag{9}$$

5.2.1 Sarcos robot arm

Now we apply the above method to a Sarcos robot arm dataset which consists of 21 dimensional covariates (seven positions, seven velocities, and seven accelerations) and 7 dimensional response (seven joint torques). The dataset consists of 44484 training points and 4449 test points. This implies that we need to solve 4449 Monte Carlo estimation problems. First a random subset of size $N = 1000$ is sampled from the training data and then $N' = 100$ is used for approximation of f

using (9). Here $\alpha = \gamma = 25$ and $\beta = \delta = 0.04$ so the both hyperparameters have prior mean 1 and standard deviation 0.2.

For each test point, the sampling procedure is repeated 10 times and predictions are made using sample means, ZV-CV, and simplified CF. Figure 4 plots the sampling standard deviation of arithmetic means vs CF (top) and ZV-CV vs CF (bottom) for three different sample sizes n calculated over 10 replications. Each point in the plots represent a test point. Notice that CF offers lower estimation variance than other methods in all settings.

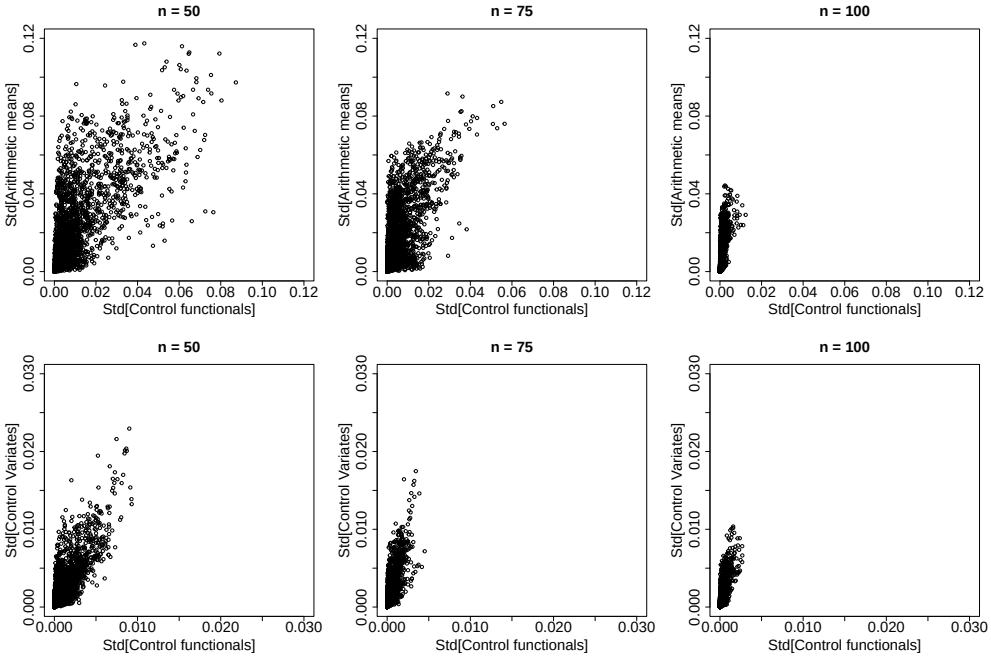

Figure 4: Sampling standard deviation of AM vs CF (top) and ZV-CV vs CF (bottom) calculated over 10 replications.

5.3 Normalizing constant for non-linear ordinary differential equation models

Suppose we are presented with data $\mathbf{y}$ that is assumed to have arisen from one of two (or more) models m_i characterised by a probability density $p(\mathbf{y}|m_i)$. Given a priori model probabilities $p(m_i)$, the data $\mathbf{y}$ induces posterior probabilities $p(m_i|\mathbf{y})$ that are the basis for Bayesian model comparison. Bayes factor is an important model comparison technique. A natural approach for computing Bayes factors is to directly compute the evidence $p(\mathbf{y}|m_i) = \int p(\mathbf{y}|m_i, \theta)p(\theta|m_i)d\theta$ provided by data $\mathbf{y}$ in favour of model m_i, where θ are parameters associated with model m_i. Yet for almost all models of interest, the evidence is unavailable in closed form and must be approximated/estimated.

5.3.1 Thermodynamic integration

Empirical investigations by Friel and Wyse [2012] have revealed that thermodynamic integration (TI) is among the most promising approach for estimation of model evidence. TI targets the model evidence directly; in what follows we therefore implicitly condition upon a model m and aim to compute the evidence $p(\mathbf{y}) := p(\mathbf{y}|m)$ provided by data $\mathbf{y}$ in favour of model m. The power posterior is defined as $p(\theta|\mathbf{y}, t) = p(\mathbf{y}|\theta)^t p(\theta)/Z_t(\mathbf{y})$ where the normalizing constant is given by $Z_t(\mathbf{y}) = \int p(\mathbf{y}|\theta)^t p(\theta) d\theta$. Here t is known as an inverse temperature parameter. The normalising constant $Z_0(\mathbf{y})$ is equal to one and $Z_1(\mathbf{y})$ is equal to $p(\mathbf{y})$, the model evidence that we aim to estimate. The standard thermodynamic identity is

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

where the expectation in the integrand is with respect to the power posterior whose density is given above. In TI, this one-dimensional integral is evaluated numerically using a 4 quadrature approximation over a discrete temperature ladder $0 = t_0 < t_1 < \ldots, < t_m = 1$. The second-order quadrature is

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

where $\hat{\mu}_i$ and $\hat{\nu}_i$ are Monte Carlo estimates of the posterior mean and variance respectively of $\log p(\mathbf{y}|\theta)$ when θ arises from $p(\theta|\mathbf{y}, t_i)$.

5.3.2 Non-linear ODE models - van der Pol oscillator

The non-linear system of van der Pol oscillator is described in detail in the appendix. In this experiment, TI is performed for computing model evidence using the temperature schedule $t_i = (i/30)^5$ as recommended by Calderhead and Girolami [2009]. Let n be the number of samples obtained and each sample is indexed by its observation time. In this example, $n = 11$ samples are obtained at times $s_i = i$ where $i = 0, 1, \ldots, 10$.

Oates et al. [2016] studied the ZV-CV method for thermodynamic integration and showed that this controlled thermodynamic integration (CTI) method provides significant variance reduction when the integrand f can be approximated by a low-degree polynomial and the target density π resembles a Gaussian distribution. The performance of simple CF on van der Pol oscillator system is compared to CTI and sample means in Figure 5. Both CTI and simplified CF display some bias but that is okay because TI using numerical quadrature itself produces bias.

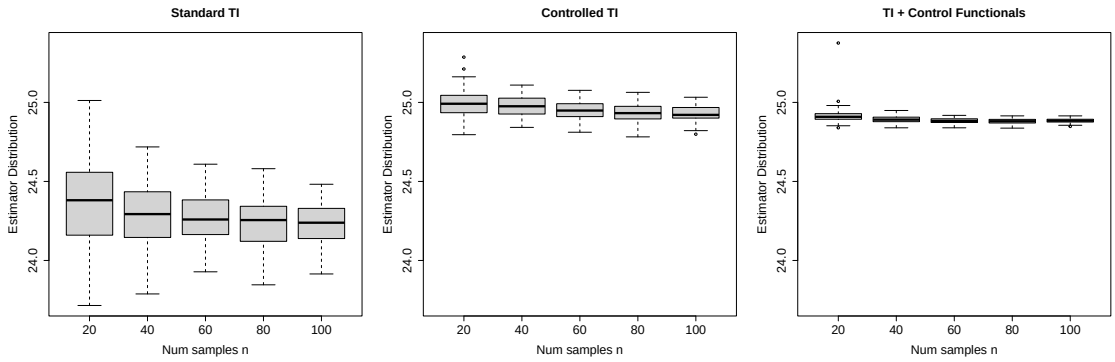

Figure 5: Sampling distribution of different Monte Carlo estimators over 100 replications.

In a previous paper introducing CTI, Oates et al. [2016] point out that CTI is ineffective for non-linear ODE models, especially when f cannot be approximated by a low-degree polynomial. They considered the Goodwill oscillator system which has a rough multimodal log-likelihood surface rendering a multimodal power posterior. It is interesting to note that no experimental results for CF have been provided for the Goodwill oscillator system wherein we know that CTI fails due to the multimodal nature of both - the test function and target density. A preliminary analysis of CF on smooth bimodal target densities reveals that CF actually performs worse than ZV-CV and even arithmetic means. This gives reason to believe that CF will not perform well on Goodwill oscillator as well wherein there are more than two well-separated modes.

5.4 Gaussian mixtures

Consider the scalar Gaussian mixture density

$$\pi(x) = p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)$$

where p is the mixing proportion, μ_1, μ_2 are the means and σ_1^2, σ_2^2 are the variance of two mixing Gaussian distributions. We are interested in estimating two kinds of integrands. The first is a polynomial example $f(x) = x^2$ wherein we can expect ZV-CV to perform well, infact provide essentially exact estimates. The second is an extremely smooth test function $f(x) = e^x$. The true expectation is known for both test functions. Here again we only compare to simplified CF because the samples are not independent and obtained using random-walk Metropolis-Hastings. The sample size is $n = 200$ with a 500 points burn-in.

Table 2 presents the average estimate along the mean squared error calculated over 100 replications for arithmetic means, ZV-CV, and simplified CF. CF gives dramatic variance reduction for

cases (a) and (b) where either π is unimodal or the two modes are not separated by more than 1 standard deviation. However, as soon as the modes are well-separated, the variance of CF estimator blows up for both test functions.

Case	Problem		Estimator		
	$f(x)$	π	Arithmetic Means	ZV-CV	CF (simplified)
(a)	$\exp(x)$	$\mathcal{N}(0,1)$	1.6486 $\pm$0.0944	1.6277 $\pm$0.0049	1.6447 $\pm 10^{-4}$
(b)	x^2	$0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(1,1)$	1.957 $\pm$0.1448	2.0082 $\pm$0.0022	1.9875 $\pm 10^{-4}$
(c)	x^2	$0.5\mathcal{N}(-2,1)+0.5\mathcal{N}(2,1)$	4.9121 $\pm$0.4016	4.8835 $\pm$0.0305	2.6159 $\pm$6.9807
(d)	$\exp(x)$	$0.5\mathcal{N}(-2,1)+0.5\mathcal{N}(2,1)$	6.1447 $\pm$6.2988	5.9863 $\pm$2.4247	4.829 $\pm$20.6664

Table 2: Average estimate and mean squared error of Monte Carlo estimators over 100 replications for different combinations of test function f and target density π.

6 Extension

In many real-world scenarios, deriving samples directly from the target distribution is either not feasible or desirable. Consider two applications - (a) problems arising in high energy physics, finance, insurance, etc that often rely on rare event sampling, and (b) problems where sampling from a proxy distribution is computationally easier and offers variance reduction.

In both situations importance sampling is employed to significant benefit. A valuable extension for future research is to adapt the CF method to settings where samples are acquired from a suitable proposal distribution instead of target itself. For convenience of notation, let us call it importance control functionals (ICF).

We have seen how CF performs poorly for multimodal target densities, particularly when Markov chains tend to get stuck in one mode for a long time. Even when the assumption of uniform ergodicity is satisfied, the number of steps until the Markov chain traverses all high probability areas of the state space can be quite large. Owing to its $O(n^3)$ complexity, this sample size can be computationally prohibitive for CF. A foreseeable benefit of ICF is in ameliorating CF in this situation as a random walk through the target can be replaced by IID samples from the suitable proposal.

To write the expression for ICF, first note that $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \frac{1}{n-m}\mathbf{1}_{n-m}^T(\mathbf{f}_1 - \mathbf{K}_{1,0}\hat{\boldsymbol{\beta}} - \hat{c}\mathbf{1}_{n-m}) + \hat{c} = \frac{1}{n-m}\left(\mathbf{1}_{n-m}^T\mathbf{f}_1 - \mathbf{1}_{n-m}^T\mathbf{K}_{1,0}\hat{\boldsymbol{\beta}}\right).$$

Plugging in the solution for $\hat{c}$ and $\hat{\boldsymbol{\beta}}$ from (5), it can be seen that $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$ is a weighted sum of $[\mathbf{f}_0^T, \mathbf{f}_1^T]^T$

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f) = \frac{1}{n-m}\left(\mathbf{1}_{n-m}^T\mathbf{f}_1 + \underbrace{\left[\left(\frac{\mathbf{1}_{n-m}^T\mathbf{K}_{1,0}\mathbf{K}_0^{-1}\mathbf{1}_m}{1+\mathbf{1}_m^T\mathbf{K}_0^{-1}\mathbf{1}_m}\right)\mathbf{K}_0^{-1}\mathbf{1}_m - \mathbf{K}_0^{-1}\mathbf{K}_{1,0}^T\mathbf{1}_{n-m}\right]^T}_{=:A}\mathbf{f}_0\right).$$

The authors claim that the weights add up to 1. However, this claim is only true as $n \to \infty$ since the weights in the above expression add up to

$$1 - \frac{\mathbf{1}_{n-m}^T\mathbf{K}_{1,0}\mathbf{K}_0^{-1}\mathbf{1}_m}{(n-m)(1+\mathbf{1}_m^T\mathbf{K}_0^{-1}\mathbf{1}_m)}. \tag{10}$$

Therefore, it can be seen that the sum of weights of $\mathbf{f}_0$ goes to zero explaining the variance reduction obtained with $\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)$. Now, suppose the samples $\{\mathbf{x}_i\}_{i=1}^n$ are generated from a proposal density q and the corresponding weights are $w(\mathbf{x}_i) := \pi(\mathbf{x}_i)/q(\mathbf{x}_i)$. Since the method is amenable to unnormalized π, we will set up the method using normalized weights where $\bar{w}_1(\mathbf{x}_j) := w(\mathbf{x}_j)/\sum_{i=m+1}^n w(\mathbf{x}_i)$ for $j \in \{m+1, \ldots, n\}$ and $\bar{w}_0(\mathbf{x}_j) := w(\mathbf{x}_j)/\sum_{i=1}^m w(\mathbf{x}_i)$ for $j \in \{1, \ldots, m\}$. Let $\mathbf{W}_0$ denote the diagonal matrix of $\{\bar{w}_0(\mathbf{x}_i)\}_{i=1}^m$ and $\mathbf{W}_1$ denote the diagonal matrix with entries $\{\bar{w}_1 too(\mathbf{x}_i)\}_{i=m+1}^n$, then we can propose ICF estimator as

$$\hat{\mu}(\mathcal{D}_0, \mathcal{D}_1; f)_{IS} := \mathbf{1}_{n-m}^T\mathbf{W}_1\mathbf{f}_1 + A^T\mathbf{W}_0\mathbf{f}_0. \tag{11}$$

As long as the proposal is chosen carefully such that the samples are representative of the true target density, $f_{\mathcal{D}_0}$ should be able to approximate f well regardless of the fact that the samples in $\mathcal{D}_0$ used for approximation are obtained from q instead. Of course, such an extension would require significant mathematical work to establish its convergence properties but can be established in spirit of Oates et al. [2019].

7 Discussion

Oates et al. [2017] have developed a strong non-parametric approach for estimating expectations with super-$\sqrt{n}$ convergence. In this paper, the authors work with a bounded state space Ω and establish an $O_P(n^{-7/6})$ error variance of CF estimator assuming that both π and f are continuously

differentiable. Convergence results for unbounded Ω are more challenging and remain a component of active research.

Additionally, a follow-up paper by the same authors [Oates et al., 2019] reveals that curse of dimension is intrinsic to this method, much like all functional approximation methods. The authors prove that the CF estimator incurs an $O_P(n^{-1-2(a\wedge b)/d+\epsilon})$ where a and b are related to the smoothness of π and f respectively. Note that, as expected, the convergence is faster for smoother functions and lower dimensions.

It is natural to expect that a quasi Monte Carlo (QMC) setup would yield faster convergence instead of IID Monte Carlo. To strengthen this intuition, Oates and Girolami [2016] study the performance of CF estimator in quasi Monte Carlo (QMC) setting. QMC algorithms can theoretically achieve a $O_P(n^{-\alpha/d})$ rate for integrands of dimension d and derivatives of order α. However, this rate-optimal α-QMC is often not employed in practice because either α is unknown or an explicit QMC algorithm for functions of smoothness α is unavailable. Therefore, sub-optimal QMC algorithms for smoothness α_L are used when $\alpha_L < \alpha$. In this short, yet theoretically rich study the authors show how CF leads to faster convergence in a variety of deterministic and random sampling cases. A particularly interesting result states that under quasi-uniform sampling of points used to construct the surrogate function, CF+QMC accelerates α_L-QMC by a factor of $O_P(n^{-(\alpha-\alpha_L)/d})$.

Acknowledgement

I would like to thank my mentor Vincent Roulet for advice and encouragement to think about theoretical implications of this method and conduct thorough simulation studies. I would also like to thank the instructors for the course STAT 572, Marina Meila and Yen-Chi Chen, for their constant support thoughout the quarter. Their help has been crucial to structure this project. In addition, I would like to acknowledge the help of my colleague and friend Lars van der Laan with an application study of the method. Finally, I would like to thank the authors of the paper for writing such an insightful and thought-provoking paper.

References

Sigrún Andradóttir, Daniel P Heyman, and Teunis J Ott. Variance reduction through smoothing and control variates for markov chain simulations. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 3(3):167–189, 1993.

Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms. *Physical review letters*, 83(23):4682, 1999.

François-Xavier Briol, Chris J Oates, Mark Girolami, Michael A Osborne, and Dino Sejdinovic. Probabilistic integration: A role for statisticians in numerical analysis. *arXiv preprint arXiv:1512.00933*, 1293, 2015.

Ben Calderhead and Mark Girolami. Estimating bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

Josef Dick and Friedrich Pillichshammer. Discrepancy theory and quasi-Monte Carlo integration. In *A panorama of discrepancy theory*, pages 539–619. Springer, 2014.

Nial Friel and Jason Wyse. Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Aleksandar Mijatović and Jure Vogrinc. On the Poisson equation for Metropolis–Hastings chains. *Bernoulli*, 24(3):2401–2428, 2018.

Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance markov chain Monte Carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.

Chris Oates and Mark Girolami. Control functionals for quasi-Monte Carlo integration. In *Artificial Intelligence and Statistics*, pages 56–65. PMLR, 2016.

Chris J Oates, Theodore Papamarkou, and Mark Girolami. The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111 (514):634–645, 2016.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

Chris J Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141–1159, 2019.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Anne Philippe. Processing simulation output by Riemann sums. *Journal of Statistical Computation and Simulation*, 59(4):295–314, 1997.

Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian Monte Carlo. *Advances in neural information processing systems*, pages 505–512, 2003.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016.

Reuven Y Rubinstein and Ruth Marcus. Efficiency of multivariate control variates in Monte Carlo simulation. *Operations Research*, 33(3):661–677, 1985.

Steve Smale and Ding-Xuan Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41(3):279–305, 2004.

Steve Smale and Ding-Xuan Zhou. Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Leah F South, Chris J Oates, Antonietta Mira, and Christopher Drovandi. Regularised zero-variance control variates for high-dimensional variance reduction. *arXiv preprint arXiv:1811.05073*, 2018.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Hongwei Sun and Qiang Wu. Application of integral operator for regularized least-square regression. *Mathematical and Computer Modelling*, 49(1-2):276–285, 2009.

A Thermodynamic integration

We consider nonlinear dynamical systems of the form

$$\frac{d\mathbf{x}}{ds} = f(\mathbf{x}, s, \theta), \qquad \mathbf{x}(0) = \mathbf{x}_0 \,. \tag{12}$$

We assume only a subset of the variables are observed under noise, so that when $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_b]$, only $\mathbf{x}_a$ is observed and $\mathbf{y}$ is a $d \times n$ matrix of observations of the variable $\mathbf{x}_a$. Write $s_1 < s_2 < \cdots < s_n$ for the times at which observations are obtained, such that $\mathbf{y}(s_j) = \mathbf{y}\cdot, j$, where $\mathbf{y}\cdot, j$ is the jth column of data matrix $\mathbf{y}$. We consider a Gaussian observation process with likelihood

$$p(\mathbf{y}|\theta, \mathbf{x}_0, \sigma) = \prod_{i=1}^{n} N(\mathbf{y}(s_j)|\mathbf{x}_a(s_j; \theta, \mathbf{x}_0), \sigma^2 I) \,,$$

where $\mathbf{x}_a(s_j; \theta, \mathbf{x}_0)$ is the solution of the system in (12) and $\sigma > 0$ is assumed known. The authors use an augmented ODE method to solve for the gradients of log power posterior. The resulting gradient function can be computed using a set of differential equations that can be computed with negligible costs when differential geometric sampling schemes are employed. In their case, they use manifold Metropolis-adjusted Langevin algorithm (m-MALA) developed by Girolami and Calderhead [2011].

However, in my experimentation, I noticed that even simple finite differences method provides good gradient calculations and solving the augmented ODE for gradient calculations is quite unnecessary. Similarly, the power posterior is unimodal Gaussian-like distribution. Therefore, a regular random-walk Metropolis Hastings algorithm can be employed to get samples that represent the target distribution well. This saves a lot of computational effort.

In this example, the authors illustrate their method on the van der Pol oscillator, a non-conservative oscillator with non-linear damping. Here a position $x(s) \in \mathbb{R}$ evolves in time s according to the second order differential equation

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

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 in (12)

$$f(\mathbf{x}, s, \theta) = \theta(1 - x_1^2)x_2 - x_1 \,,$$

where only the first component x_1 is observed. This system was solved numerically using $\theta = 1, x_0 = [0, 2]$. Observations were made once every time unit, up to 10 units, and Gaussian measurement

noise of standard deviation $\sigma = 0.1$ was added. A log-normal prior was placed on θ such that $\log(\theta) \sim N(0, 0.25)$. For this experiment, the temperature schedule is $t_i = (i/30)^5$ as recommended by Calderhead and Girolami [2009]. Here n is the number of samples obtained and each sample is indexed by its observation time. In this example, $n = 11$ samples are obtained at times $s_i = i$ where $i = 0, 1, \ldots, 10$.