Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

ON THE STABILITY OF KERNELIZED CONTROL FUNCTIONALS ON PARTIAL AND BIASED STOCHASTIC INPUTS

Henry Lam
Haofeng Zhang

Department of Industrial Engineering and Operations Research
Columbia University
500 W. 120th Street
New York, NY 10027

ABSTRACT

We investigate some theoretical properties of kernelized control functionals (CFs), a recent technique for variance reduction, regarding its stability when applied to subsets of input distributions or biased generating distributions. This technique can be viewed as a highly efficient control variate obtained by carefully choosing a function of the input variates, where the function lies in a reproducing kernel Hilbert space with known mean thus ensuring unbiasedness. In large-scale simulation analysis, one often faces many input distributions for which some are amenable to CFs and some may not due to technical difficulties. We show that CFs retain good theoretical properties and lead to variance reduction in these situations. We also show that, even if the input variates are biasedly generated, CFs can correct for the bias but with a price on estimation efficiency. We compare these properties with importance sampling, in particular a version using a similar kernelized approach.

1 INTRODUCTION

We study some theoretical properties of kernelized control functions (CFs) regarding its stability when applied in stochastic simulation analysis. CF is a recent technique first proposed by Oates et al. (2017), which can be viewed as a highly efficient control variate by suitably choosing a function on the input variates. For instance, suppose we want to estimate $\mathbb{E}_\pi[f(X)]$ by Monte Carlo simulation, where $\mathbb{E}_\pi[\cdot]$ denotes the expectation for X under distribution π. CF seeks to find a function $s_m(\cdot)$ applied on X such that $\mathbb{E}_\pi[s_m(X)]$ is known, and that $f(X)-s_m(X)$ has a very low variance so that the sample average of

$$f(X)-s_m(X)+\mathbb{E}_\pi[s_m(X)]$$

for randomly simulated X is a highly efficient (possibly super-efficient, i.e., exceeding the canonical square-root convergence) estimator for $\mathbb{E}_\pi[f(X)]$. This function $s_m(\cdot)$ is constructed as a functional approximation for $f(\cdot)$ by utilizing a beginning set of samples. It lies in a reproducing kernel Hilbert space (RKHS) with the special property that any element can be disintegrated into a constant and a function that has mean zero under π. The latter property is a consequence of applying a Stein operator, with respect to π, to a "primary" RKHS in order to obtain the approximation basis for $s_m(\cdot)$.

Our interest in this technique is to apply this to potentially large-scale simulation analysis. In this situation, typically the modeler faces several, possibly many input models or distributions. To apply CF, one needs to know the parametric forms of these input distributions (up to normalizing constants, for constructing the Stein operators) and also ensures that these distributions satisfy a set of technical conditions. It could well be the case that some of these input distributions are amenable to this technique while some are not. Thus, whereas Oates et al. (2017) (and subsequent related works Oates et al. 2019; Liu et al. 2016; Liu

and Lee 2017; Chwialkowski et al. 2016) assume all input distributions are known, we focus on whether CF behaves stably when applied to only a subset of these distributions.

We show that CF retains super-efficiency in terms of the error rate associated with the subset of applicable input distributions. For those distributions where CF does not apply on, the final output gives the canonical square-root error rate. Thus, if the amenable input distributions contribute to most of the variance, then the CF outputs are effectively super-efficient. We prove this result on partially applied CF via a direct use of the so-called regularized least-square (RLS) functional approximation that looks for a closest function in an RKHS from data (Cucker and Smale 2002). Moreover, we contrast this conclusion with a closely related method referred to as black-box importance sampling (Liu and Lee 2017; Chwialkowski et al. 2016; Liu et al. 2016), which relies on assigning weights over the samples, where the weights are optimized from a kernel induced by the same Stein operator as CF. However, since this method does not approximate the function f but rather uses the weights, it is plausible that without additional adjustment it could have subpar performance when the kernel is induced by only part of all input distributions.

Our second interest is to investigate the convergence of CF when the input variates are generated from a "wrong" or approximating distribution than the underlying model. We show that, if the likelihood ratio between the approximating and the underlying distributions are controllable, then CF can automatically correct for the bias, though with a price on estimation efficiency (i.e., sub-canonical convergence). This property can be attributed to the fact that s_m are able to well approximate f, in terms of variance under π, even if the utilized data have a different distribution. In this sense, CF can act as a bias corrector that is similar to importance sampling. It also can be viewed as a more powerful version of the weighted Monte Carlo studied in Glasserman and Yu (2005). However, its loss of efficiency (at least as a consequence of our analysis) indicates that importance sampling, or a combination with it, is more efficient in bias reduction than CF.

The remainder of this paper is as follows. Section 2 first describes our setup and notations. Section 3 develops some machinery from RLS that we need to utilize in our analysis. Section 4 presents our result on partially applied CF. Section 5 studies the bias reduction property of CF under a deviating generating distribution.

2 SETUP AND THE CONTROL FUNCTIONAL FRAMEWORK

Consider a random vector (X,Y) where X takes values in an open set $\Omega \subset \mathbb{R}^d$ and Y takes values in an open set $\Theta \subset \mathbb{R}^p$. We assume X admits a positive (marginal) density $\pi_x(x) > 0$ on Ω with respect to d-dimensional Lebesgue measure, which has a parametric form that is known up to a normalizing constant. Similarly, we also denote $\pi_{y|x}(y|x)$ as the distribution of Y given X, and π as the joint distribution of (X,Y) (these can be viewed as densities without ambiguity, but the assumption of having a density is not necessary for these distributions).

Our goal is to estimate the expectation of $f(X,Y)$, which we write as $\mu := \mathbb{E}_\pi[f(X,Y)]$. Our premise is that we can run simulation and have access to a collection of i.i.d. samples $D = \{(x_i,y_i)\}_{i=1}^n$ where (x_i,y_i) are sampled from π (or some other distributions as discussed in Section 5). The vector X is assumed to be the "dominating" factor in the simulation, contributing to the most output variance, whereas Y contributes a small variance (which will be rigorized later). Moreover, we assume that $\nabla_x \log \pi_x(x)$ is well-defined and is computable for given x_i's, so that we can apply CF on X as we will discuss. Throughout this paper, we also suppose $f : \Omega \times \Theta \to \mathbb{R}$ satisfies $\mathbb{E}_\pi[f(X,Y)^2] < \infty$.

For convenience, for any measurable function $g : \Omega \times \Theta \to \mathbb{R}$, we write $\mu(g) = \mathbb{E}_\pi[g(X,Y)]$; for any measurable function $g : \Omega \to \mathbb{R}$, we write $\mu_x(g) = \mathbb{E}_{\pi_x}[g(X)]$. Let $L^2(\pi_x)$ denote the space of measurable functions $g : \Omega \to \mathbb{R}$ for which $\mu_x(g^2)$ is finite, with the norm written as $\|\cdot\|_{L^2(\pi_x)}$. Let $C^k(\Omega, \mathbb{R}^j)$ denote the space of (measurable) functions from Ω to $\mathbb{R}^j$ with continuous partial derivatives up to order k. The region Ω can be bounded or unbounded; in the former case, the boundary $\partial\Omega$ is assumed to be piecewise smooth (i.e., infinitely differentiable).

Following the framework in Oates et al. (2017), we divide the data D into two disjoint subsets as $D_0 = \{(x_i, y_i)\}_{i=1}^m$ and $D_1 = \{(x_i, y_i)\}_{i=m+1}^n$, where $1 \le m \le n$. We use D_0 to construct a CF $s_m(\cdot) \in L^2(\pi_x)$ that is a partial approximation to f (that only depends on x), and consider the function

$$f_m(x,y) = f(x,y) - s_m(x) + \mu_x(s_m).$$

The final estimator is then given by a sample average of $f_m(\cdot,\cdot)$ on D_1, i.e.,

$$\hat{\mu} := \frac{1}{n-m} \sum_{j=m+1}^{n} f_m(x_j, y_j).$$

It is clear that we have unbiasedness, since $\mathbb{E}_\pi[\hat{\mu}|D_0] = \mu(f_m) = \mu$ for any given D_0 and hence $\mathbb{E}_\pi[\hat{\mu}] = \mu$, where here $\mathbb{E}_\pi[\cdot|D_0]$ and $\mathbb{E}_\pi[\cdot]$ are with respect to the data distribution.

Denote the "score function" of the density π_x by $\mathbf{u}(x) := \nabla_x \log \pi_x(x)$. The CF s_m is in the form

$$s_m(x) := c + \psi(x)$$

$$\psi(x) := \nabla_x \cdot \phi(x) + \phi(\mathbf{x}) \cdot \mathbf{u}(x)$$

where $c \in \mathbb{R}$ is a constant and $\phi \in C^1(\Omega, \mathbb{R}^d)$. Note that, under suitable conditions (that we describe below), $\mu_x(\psi) = 0$ via integration by parts, which constitutes the Stein operator applied on the function ϕ.

Next we specify our choice of each component of $\phi(x)$. Suppose $\phi_i : \Omega \to \mathbb{R}$ is in a Hilbert space $\mathscr{H} \subset L^2(\pi) \cap C^1(\Omega, \mathbb{R})$ with inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$. Moreover, we require that $\mathscr{H}$ is an RKHS. This implies that there exists a symmetric positive definite function $k : \Omega \times \Omega \to \mathbb{R}$ such that for all $x \in \Omega$, we have $k(\cdot, x) \in \mathscr{H}$ and for all $x \in \Omega$ and $h \in \mathscr{H}$, we have $h(x) = \langle h(\cdot), k(\cdot, x) \rangle$.

The vector-valued function $\phi(x) : \Omega \to \mathbb{R}^d$ is defined in the Cartesian product space $\mathscr{H}^d := \mathscr{H} \times \cdots \times \mathscr{H}$, which is an RKHS with the inner product $\langle \phi, \phi' \rangle_{\mathscr{H}^d} = \sum_{i=1}^d \langle \phi_i, \phi_i' \rangle_{\mathscr{H}}$. We will see next that under some reasonable assumptions ψ also belongs to an RKHS $\mathscr{H}_0$ with a kernel denoted k_0.

Following Oates et al. (2017), we make the following assumptions and conclusions:

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

Assumption 2 Let $\mathbf{n}(x)$ be the unit normal to the boundary $\partial\Omega$ of the state space Ω. For π_x-almost all $x \in \Omega$ the kernel k satisfies

$$\oint_{\partial\Omega} k(x,x') \pi_x(x') \mathbf{n}(x') S(dx') = \mathbf{0}$$

and

$$\oint_{\partial\Omega} \nabla_x k(x,x') \pi_x(x') \cdot \mathbf{n}(x') S(dx') = 0.$$

The notation $\oint_{\partial\Omega}$ denotes a surface integral over $\partial\Omega$ and $S(dx')$ denotes the surface element at $x' \in \partial\Omega$.

Assumption 3 The kernel k belongs to $C^2(\Omega \times \Omega, \mathbb{R})$.

Assumption 4 The gradient-based kernel k_0 satisfies

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

This implies that

$$\int_\Omega k_0(x,x) \pi_x(x) dx < \infty.$$

These assumptions conclude the following:

1. Assumption 1 allows $\mathbf{u}(x)$ to be well-defined.

2. With Assumptions 1 and 2, $\mu_x(\psi) = 0$ and so $\mu_x(s_m) = c$.
3. With Assumptions 1 and 3, ψ belongs to $\mathscr{H}_0$, the RKHS with kernel

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

4. With Assumptions 1, 2 and 3, the gradient-based kernel k_0 satisfies

$$\int_\Omega k_0(x,x')\pi_x(x')dx' = 0$$

for π_x-almost all $x \in \Omega$.
5. With Assumptions 1, 2, 3 and 4, we have $\mathscr{H}_0 \subset L^2(\pi_x)$.

Next, let $\mathscr{C}$ denote the RKHS of constant functions with kernel $k_{\mathscr{C}}(x,x') = 1$ for all $x,x' \in \Omega$. The norms associated to $\mathscr{C}$ and $\mathscr{H}_0$ is denoted by $\|\cdot\|_{\mathscr{C}}$ and $\|\cdot\|_{\mathscr{H}_0}$ respectively. $\mathscr{H}_+ = \mathscr{C} + \mathscr{H}_0$ denotes the set $\{c + \psi : c \in \mathscr{C}, \psi \in \mathscr{H}_0\}$. Equip $\mathscr{H}_+$ with the structure of a vector space, with addition operator $(c+\psi) + (c'+\psi') = (c+c') + (\psi+\psi')$ and multiplication operator $\lambda(c+\psi) = (\lambda c) + (\lambda\psi)$, each well-defined due to uniqueness of the representation $f = c + \psi, f' = c + \psi$ with $c, c' \in \mathscr{C}$ and $\psi, \psi' \in \mathscr{H}_0$. It is known that $\mathscr{H}_+$ can be constructed as an RKHS with kernel $k_+(x,x') := k_{\mathscr{C}}(x,x') + k_0(x,x')$ and with norm $\|f\|^2_{\mathscr{H}_+} := \|c\|^2_{\mathscr{C}} + \|\psi\|^2_{\mathscr{H}_0}$.

We will use crucially the decomposition

$$f(X,Y) = \bar{f}(X) + \varepsilon(X,Y)$$

where $\bar{f}(X) = \mathbb{E}[f(X,Y)|X]$ can be viewed as the contribution of the fluctuation on f from X, and $\varepsilon(X,Y) = f(X,Y) - \bar{f}(X)$ is the residual. Note that, by definition, $\bar{f}(X)$ and $\varepsilon(X,Y)$ are uncorrelated.

To state our next assumption, we denote $(\mathscr{H}_+)^\pi_0 := \{f \in \mathscr{H}_+ : f = 0 \text{ a.e. with respect to } \pi_x\}$ and $(\mathscr{H}_+)^\pi_1 := (\mathscr{H}_+)^{\pi\perp}_0$ the orthogonal complement $(\mathscr{H}_+)^\pi_0$ in $\mathscr{H}_+$. $\overline{(\mathscr{H}_+)_1}^\pi$ is the closure of $(\mathscr{H}_+)^\pi_1$ in $L^2(\pi_x)$, which is equal to $\overline{\mathscr{H}_+}^\pi$. See the definitions before Lemma 3 for reference. We assume a basic well-posedness condition:

Assumption 5 $\bar{f} \in \overline{\mathscr{H}_+}^\pi$.

We note that usually $\bar{f}$ is unknown in practice so it is not easy to check Assumption 5. However, Oates et al. (2019) showed that under some reasonable conditions, $\overline{\mathscr{H}_+}^\pi = L^2(\pi_x)$, and in this case, Assumption 5 appears mild. We also list the following stronger assumption that is used in Oates et al. (2017) and Oates et al. (2019):

Assumption 6 $\bar{f} \in (\mathscr{H}_+)^\pi_1$.

To make precise that X is the "dominating" factor in contributing to the simulation noise, we make the following assumption:

Assumption 7 $M_0 := \mathbb{E}_\pi[\varepsilon(X,Y)^2] < \infty$ where M_0 is a small constant.

We have seen that when the decomposition of s_m into c and ψ is known, then finding its mean $\mu_x(s_m)$ is straightforward and equal c. The effectiveness of the discussed approach lies on the approximation quality of s_m for $\bar{f}$. For the choice of s_m, we will use the so-called regularized least-squares (RLS) functional approximation in the RKHS $\mathscr{H}_+$. The next section presents our RLS analysis, which obtains s_m in a different path from Oates et al. (2017) who does not consider the extra component Y and uses a more simplified machinery.

3 REGULARIZED LEAST SQUARE FUNCTIONAL APPROXIMATION

This section develops some theoretical results about RLS. Let $z = (f(x_1,y_1), \cdots, f(x_m,y_m))^T$. The samples we need are $\{(x_j, z_j = f(x_j,y_j))\}_{j=1,\cdots,m}$. Suppose π is the underlying sampling distribution. For this

section, we do not assume any information about π. We call f_π a regression function defined by

$$f_\pi(x) = \int z d\pi(z|x) = \mathbb{E}_\pi[f(X,Y)|X = x]$$

which is $\bar{f}$ in the setting in Section 2.

The aim here is to learn the regression function $f_\pi(x)$ by constructing a good approximating function s_m from the data. Let $\mathscr{H}$ be a generic RKHS associated with the kernel $k(x,y)$. Let $\|\cdot\|_{\mathscr{H}}$ denote the norm on $\mathscr{H}$. Note that $k_x = k(x,\cdot)$ is a function in $\mathscr{H}$.

For this section, we only impose the following assumption:

Assumption 8 $\kappa := \sup_{x\in\Omega}\sqrt{k(x,x)} < \infty$, $M_0 := \mathbb{E}_\pi[(z - f_\pi(x))^2] < \infty$. For any $g \in \mathscr{H}$, $g(x)$ is π-measurable. $f_\pi \in L^2(\pi)$.

It follows that for any $g \in \mathscr{H}$,

$$\sup_{t\in\Omega}|g(t)|^2 = \sup_{t\in\Omega}|\langle g, k_t\rangle|^2 \le \sup_{t\in\Omega}\|g\|_{\mathscr{H}}^2\|k_t\|_{\mathscr{H}}^2 \le \kappa^2\|g\|_{\mathscr{H}}^2.$$

So under Assumption 8, any $g \in \mathscr{H}$ is a bounded function. We point out the following inequality that we will use frequently:

$$\|g\|_{L^p(\pi)} \le \kappa\|g\|_{\mathscr{H}}, \forall 1 \le p \le \infty.$$

The RLS functional approximation is given by

$$s_m(x) := \arg\min_{g\in\mathscr{H}}\left\{\frac{1}{m}\sum_{j=1}^m (f(x_j,y_j) - g(x_j)) + \lambda\|g\|_{\mathscr{H}}\right\}$$

where $\lambda > 0$ is a regularization parameter. Note that although we have y_j available, the needed data in this approximation are only the numbers $\{z_j = f(x_j,y_j)\}_{j=1,\cdots,m}$ and $\{x_j\}_{j=1,\ldots,m}$. A nice property of RLS is that there is an explicit formula for its solution, as stated below.

Lemma 1 Let $K = (k(x_i,x_j))_{m\times m}$, $\hat{k}(x) = (k(x_1,x),\cdots,k(x_m,x))^T$. Then the RLS solution is given as $s_m(x) = \beta^T\hat{k}(x)$ where $\beta = (K + \lambda m I)^{-1}z$. □

Lemma 1 shows us a way to calculate s_m in practice. On the other hand, to derive the property of s_m, we need to use some tools from functional analysis. To begin, we give an equivalent form of s_m in terms of linear operators. Define the sampling operator $S_x : \mathscr{H} \to \mathbb{R}^m$ associated with a discrete subset $\{x_i\}_{i=1}^m$ of X by

$$S_X(g) = (g(x_i))_{i=1}^m,\ f \in \mathscr{H}.$$

The adjoint of the sampling operator, $S_x^T : \mathbb{R}^m \to \mathscr{H}$, is given by

$$S_X^T(c) = \sum_{i=1}^m c_i k_{x_i}, c \in \mathbb{R}^m.$$

Note that the compound mapping $S_x^T S_x$ is a positive self-adjoint operator on $\mathscr{H}$. Let I denote the identity mapping on $\mathscr{H}$. We have:

Lemma 2 The RLS solution can be written as follows:

$$s_m = \left(\frac{1}{m}S_x^T S_x + \lambda I\right)^{-1}\frac{1}{m}S_x^T(z).$$

□

A proof can be found in Smale and Zhou (2005). It is also easy to derive this result directly from Lemma 1.

Next, we use the following established theorem. This result can be found in Theorem 2.4 and Proposition 2.10 in Sołtan (2018).

Theorem 1 [Continuous Functional Calculus] Let A be a bounded self-adjoint linear operator. Let $C(\sigma(A))$ be the set of real-valued continuous functions defined on the spectrum of A. Then for any $f \in C(\sigma(A))$, $f(A)$ is self-adjoint and $\|f(A)\| = \sup_{x\in\sigma(A)} |f(x)|$. □

Define $L : L^2(\pi) \to L^2(\pi)$ as the integral operator

$$(Lg)(x) := \int_\Omega k(x,x')g(x')\pi(x')dx',\ x \in \Omega,\ g \in L^2(\pi).$$

This operator can be viewed as a linear operator on $L^2(\pi)$ or on $\mathscr{H}$. Unless specified otherwise, we always assume the domain of L is $L^2(\pi)$. Sun and Wu (2009) shows that L is a compact and positive self-adjoint operator on $L^2(\pi)$. Denote

$$\mathscr{H}_0 := \{f \in \mathscr{H} : f = 0 \ \text{ a.e. with respect to } \pi\} \text{ and}$$

$$\mathscr{H}_1 := \mathscr{H}_0^\perp, \text{ the orthogonal complement } \mathscr{H}_0 \text{ in } \mathscr{H}.$$

Note that both $\mathscr{H}_0$ and $\mathscr{H}_1$ are closed subspaces in $\mathscr{H}$ with respect to the norm $\|\cdot\|_{\mathscr{H}}$. It is well-known that $\mathscr{H}/\mathscr{H}_0$ is isometrically isomorphic to $\mathscr{H}_1$. So $\mathscr{H}_1$ is essentially the quotient space of $\mathscr{H}$ induced by the equivalence relation a.e. with respect to π, the same equivalence relation in $L^2(\pi)$. In practice, we may treat $\mathscr{H}_1$ as $\mathscr{H}$.

Let $\overline{\mathscr{H}_1}^\pi$ be the closure of $\mathscr{H}_1$ in $L^2(\pi)$. The following theorem indicates a useful property of the integral operator L (a proof can be found in Sun and Wu (2009)).

Lemma 3 $L^{1/2}$ is an isometric isomorphism from $(\overline{\mathscr{H}_1}^\pi, \|\cdot\|_{L^2(\pi)})$ onto $(\mathscr{H}_1, \|\cdot\|_{\mathscr{H}})$. □

Denote Range(L^r) the range of L^r where L is regarded as a positive self-adjoint operator on $L^2(\pi)$. When we write $L^{-r}g \in L^2(\pi)$, it should be understood as $g \in \text{Range}(L^r)$ and $L^{-r}g$ is a preimage of g.

Next, consider an oracle or a data-free limit of s_m as

$$f_\lambda := \operatorname*{arg\,min}_{g\in\mathscr{H}} \left\{\mu_x((g-f_\pi)^2) + \lambda\|g\|_{\mathscr{H}}^2\right\}.$$

We have the following explicit expression (a proof can be found in Cucker and Smale (2002)):

Lemma 4 The solution of f_λ is given as $f_\lambda = (L+\lambda I)^{-1}Lf_\pi$. □

To show that $s_m - f_\pi$ is small, we split it into two parts

$$s_m - f_\pi = (s_m - f_\lambda) + (f_\lambda - f_\pi). \tag{1}$$

The first part in (1) comes from the statistical noise in the RLS regression, whereas the second part can be viewed as the bias of the functional approximation. We study the asymptotic error of each part in the next set of results. A proof of the following proposition can be found in Sun and Wu (2009).

Proposition 1 Suppose that $L^{-r}f_\pi \in L^2(\pi)$ where $0 \le r \le 1$. Then $\|f_\lambda - f_\pi\|_{L^2(\pi)} \le \lambda^r \|L^{-r}f_\pi\|_{L^2(\pi)}$. □

If we want to obtain a better bound for $f_\lambda - f_\pi$ by using this proposition, we may want r to be as large as possible, but meanwhile $L^{-r}f_\pi \in L^2(\pi)$ becomes a more restrictive constraint. However, we have the following proposition that can bypass this tradeoff.

Proposition 2 The range of L satisfies

$$\overline{\text{Range}(L)}^\pi = \overline{\mathscr{H}}^\pi.$$

□

Proof. Take any $f_1 \in \overline{\mathscr{H}}^{\pi} = \overline{\mathscr{H}_1}^{\pi}$. For any $\varepsilon > 0$, there exists $f_2 \in \mathscr{H}_1$ such that $\|f_1 - f_2\|_{L^2(\pi)} \leq \varepsilon$. It follows from Lemma 3 that there exists $g_1 \in \overline{\mathscr{H}_1}^{\pi}$ such that $L^{1/2} g_1 = f_2$. There exists $g_2 \in \mathscr{H}_1$ such that $\|g_1 - g_2\|_{L^2(\pi)} \leq \frac{\varepsilon}{\kappa}$. Again, it follows from Lemma 3 that there exists $h_1 \in \overline{\mathscr{H}_1}^{\pi}$ such that $L^{1/2} h_1 = g_2$. Then we have

$$\|Lh_1 - f_2\|_{L^2(\pi)} \leq \kappa \|Lh_1 - f_2\|_{\mathscr{H}} = \kappa \|L^{1/2} g_2 - L^{1/2} g_1\|_{\mathscr{H}} = \kappa \|g_2 - g_1\|_{L^2(\pi)} \leq \varepsilon$$

and

$$\|Lh_1 - f_1\|_{L^2(\pi)} \leq \|Lh_1 - f_2\|_{L^2(\pi)} + \|f_2 - f_1\|_{L^2(\pi)} \leq 2\varepsilon.$$

This implies that $f_1 \in \overline{\text{Range}(L)}^{\pi}$ so

$$\overline{\text{Range}(L)}^{\pi} \supset \overline{\mathscr{H}}^{\pi}.$$

On the other hand, since

$$\text{Range}(L) \subset \text{Range}(L^{1/2}) \subset \mathscr{H}$$

we have

$$\overline{\text{Range}(L)}^{\pi} \subset \overline{\mathscr{H}}^{\pi}.$$

□

The following two propositions can be obtained via direct computation. Proofs can be found in Sun and Wu (2009).

Proposition 3 We have

$$\|s_m - f_\lambda\|_{\mathscr{H}} \leq \frac{1}{\lambda} \|\Delta\|_{\mathscr{H}}$$

where

$$\Delta := \frac{1}{m} \sum_{i=1}^{m} (z_i - f_\lambda(x_i)) k_{x_i} - L(f_\pi - f_\lambda).$$

□

Proposition 4 We have

$$\mathbb{E}_\pi[\|\Delta\|_{\mathscr{H}}^2] \leq \frac{1}{m} \kappa^2 (\mu_x((f_\pi - f_\lambda)^2) + M_0).$$

With this, we have the following estimate:

Corollary 2 We have

$$\mathbb{E}_\pi[\|s_m - f_\lambda\|_{\mathscr{H}}^2] \leq \frac{\kappa^2 (\mu_x((f_\pi - f_\lambda)^2) + M_0)}{\lambda^2 m}.$$

Proof. Combining Proposition 3 and Proposition 4, we obtain

$$\mathbb{E}_\pi[\|s_m - f_\lambda\|_{\mathscr{H}}^2] \leq \frac{\kappa^2 (\mu_x((f_\pi - f_\lambda)^2) + M_0)}{\lambda^2 m}.$$

Also note that

$$\mu_x((s_m - f_\lambda)^2) \leq \kappa^2 \|s_m - f_\lambda\|_{\mathscr{H}}^2.$$

□

Finally, putting everything together we have:

Corollary 3 Suppose that $L^{-r} f_\pi \in L^2(\pi)$ where $0 \le r \le 1$. Then

$$\mathbb{E}_\pi[\mu_x((f_\pi - s_m)^2)] \le \left(\frac{2\kappa^4}{\lambda^{2-2r} m} + 2\lambda^{2r} \right) \mu_x((L^{-r} f_\pi)^2) + \frac{2\kappa^4 M_0}{\lambda^2 m}.$$

In particular, taking $\lambda = m^{-1/2}$, we have

$$\mathbb{E}_\pi[\mu_x((f_\pi - s_m)^2)] \le C_\kappa m^{-r} \mu_x((L^{-r} f_\pi)^2) + 2\kappa^4 M_0$$

where $C_\kappa = 2\kappa^4 + 2$ only depends on κ.

Proof. Proposition 1 shows that if $L^{-r} f_\pi \in L^2(\pi)$, then

$$\mu_x((f_\lambda - f_\pi)^2) \le \lambda^{2r} \mu_x((L^{-r} f_\pi)^2).$$

We note that

$$\mu_x((f_\pi - s_m)^2) \le 2(\mu_x((f_\pi - f_\lambda)^2) + \mu_x((f_\lambda - s_m)^2)).$$

So taking expectation, we have

$$\mathbb{E}_\pi[\mu_x((f_\pi - s_m)^2)] \le \left(\frac{2\kappa^4}{\lambda^2 m} + 2 \right) \mu_x((f_\pi - f_\lambda)^2) + \frac{2\kappa^4 M_0}{\lambda^2 m} \le \left(\frac{2\kappa^4}{\lambda^{2-2r} m} + 2\lambda^{2r} \right) \mu_x((L^{-r} f_\pi)^2) + \frac{2\kappa^4 M_0}{\lambda^2 m}.$$

□

Corollary 3 shows that s_m computed through RLS approximates f_π closely, measured by a mean square error under π of order m^{-r}. It appears that Corollary 3 is more refined than the theory used by Oates et al. (2017). Accordingly, we will prove a better result in the next section.

4 CONTROL FUNCTIONALS ON PARTIAL INPUTS

This section presents the properties of $\hat{\mu}$ defined in Section 2. Note that $\mathscr{H}$ in Section 3 coincides with $\mathscr{H}_+$ in Section 2. First, the following expression of s_m is a direct consequence of Lemma 1.

Lemma 5 Let

$$z = (f(x_1, y_1), \cdots, f(x_m, y_m))^T$$
$$K_+ = (k_+(x_i, x_j))_{m \times m}$$
$$\hat{k}_+(x) = (k_+(x_1, x), \cdots, k_+(x_m, x))^T$$

and

$$\hat{k}_0(x) = (k_0(x_1, x), \cdots, k_0(x_m, x))^T.$$

Then the RLS solution is given as $s_m(x) = \beta^T \hat{k}_+(x)$ where $\beta = (K_+ + \lambda m I)^{-1} z$. □

We remark that $s_m(x)$ is a linear combination of z. Moreover, these coefficients only depend on the RKHS $\mathscr{H}_0$, free of the function of interest f. The following computes the mean of s_m (the proof is straightforward and thus skipped):

Lemma 6 Let $s_m(x) = \beta^T \hat{k}_+(x)$ as given in Lemma 1. We have $\mu_x(s_m) = \beta^T \mathbf{1}$. □

Combining Lemmas 5 and 6 gives us an explicit form of the estimator $\hat{\mu}$. To describe the error of $\hat{\mu}$, we first state the following observation of Oates et al. (2017) that translates the error of s_m into the error of the two-phase estimator $\hat{\mu}$.

Proposition 5 Assume

$$\mathbb{E}_\pi[\mu((f-s_m)^2)] = I_1.$$

Then the mean square error of $\hat{\mu}$ is given by

$$\mathbb{E}_\pi[(\hat{\mu}-\mu)^2] = \mathbb{E}_\pi[\mathbb{E}_\pi[(\hat{\mu}-\mu)^2|D_0]] \leq \frac{I_1}{n-m}.$$

□

Proof. For $i = m+1,\ldots,n$, we have $\mathbb{E}_\pi[f_m(x_i,y_i)-\mu|D_0] = 0$. By the independence of D,

$$\mathbb{E}_\pi[(\hat{\mu}-\mu)^2|D_0] = \left(\frac{1}{n-m}\right)^2 \sum_{i=m+1}^{n} \mathbb{E}_\pi[(f_m(x_i,y_i)-\mu)^2|D_0].$$

It is well-known that $\mathbb{E}[(X-a)^2]$ is minimized when $a = E(X)$. This implies that

$$\mathbb{E}_\pi[(f_m(x_i,y_i)-\mu)^2|D_0] \leq \mathbb{E}_\pi[(f_m(x_i,y_i)-\mu(s_m))^2|D_0].$$

The right-hand side is exactly $\mathbb{E}_\pi[(f-s_m)^2|D_0]$. Therefore

$$\mathbb{E}_\pi[(\hat{\mu}-\mu)^2|D_0] \leq \left(\frac{1}{n-m}\right)^2 \sum_{i=m+1}^{n} \mu((f-s_m)^2) = \frac{1}{n-m}\mu((f-s_m)^2).$$

□

The main theorem in this section is:

Theorem 4 Suppose Assumptions 1-5, 7 hold and take an RLS estimate with $\lambda = m^{-\frac{1}{2}}$ and take $m = O(n)$. Then the estimator $\hat{\mu}$ is an unbiased estimator of μ with

$$\mathbb{E}_\pi[(\hat{\mu}-\mu)^2] = O(C_\kappa(C_f n^{-2} + M_0 n^{-1}))$$

where C_f is a constant free of m (and n), $C_\kappa = 2\kappa^4 + 2$ and the outside O only depends on the ratio m/n. □

Proof of Theorem 4. We apply the results from Section 3. We first check that the setting here accords with the conditions in Section 3. Recall that the set of samples in Section 3 corresponds to $\{(x_j, z_j = f(x_j,y_j))\}_{j=1,\cdots,m}$, and the f_π there corresponds to $\bar{f}(x)$ here. It follows from Assumption 4 and $k_+(x,x') = 1 + k_0(x,x')$ that

$$\kappa = \sup_{x\in\Omega} \sqrt{k_+(x,x)} < \infty$$

so Assumption 8 is satisfied. Besides, M_0 there is exactly M_0 here since by definition, we have

$$M_0 = \mathbb{E}_\pi[(z-f_\pi(x))^2] = \mathbb{E}_\pi[(f(x,y)-f(x))^2] = \mathbb{E}_\pi[\varepsilon(x,y)^2] < \infty.$$

Assumption 5 assumes that $\bar{f} \in \overline{\mathscr{H}_+}^\pi$ which is a weaker condition than $L^{-r}\bar{f} \in L^2(\pi_x)$ so we cannot apply Corollary 3 directly. However, we have shown in Proposition 2 that $\overline{\mathscr{H}_+}^\pi = \overline{\text{Range}(L)}^\pi$. Consider the following approximation approach: Fix $\varepsilon = M_0$. There exists a $g \in \text{Range}(L)$ such that $\|\bar{f}-g\|^2_{L^2(\pi_x)} \leq \varepsilon$. Let $h = f-g$ so $\bar{h} = \bar{f}-g$. Let s_m^h, s_m^g be the RLS functional approximation of h, g respectively. As we point out after Lemma 5, s_m^h is a linear functional of h, so we write

$$h - s_m^h = (f-s_m) - (g-s_m^g).$$

Next we apply Corollary 3 (with $r = 1$) to the samples $\{(x_i, g(x_i))\}$: Since $g \in \text{Range}(L)$ and g is a function of x only, $M_0^g = 0$ and

$$\mathbb{E}_\pi[\mu_x((g - s_m^g)^2)] \leq C_\kappa m^{-1} \mu_x((L^{-1}g)^2).$$

Again, we apply Corollary 3 (with $r = 0$) to the samples $\{(x_i, h(x_i, y_i))\}$: We note that $h \in L^2(\pi)$ and $\bar{g} = g$ so

$$M_0^h := \mathbb{E}_\pi[(h(x,y) - \bar{h}(x))^2] = \mathbb{E}_\pi[(f(x,y) - \bar{f}(x))^2]$$

which is the same as M_0 and thus

$$\mathbb{E}_\pi[\mu_x((\bar{h} - s_m^h)^2)] \leq C_\kappa \mu_x(\bar{h}^2) + (C_\kappa - 2)M_0 \leq (2C_\kappa - 2)M_0$$

where $C_\kappa = 2\kappa^4 + 2$.

Combining the above, we obtain

$$\mathbb{E}_\pi[\mu_x((\bar{f} - s_m)^2)] \leq 2((2C_\kappa - 2)M_0 + C_\kappa m^{-1} \mu_x((L^{-1}g)^2)).$$

Finally it remains to show a bound for

$$\mathbb{E}_\pi[\mu((f - s_m)^2)].$$

To this end, we split $f - s_m$ into two parts as follows:

$$\mathbb{E}_\pi[\mu((f - s_m)^2)] = \mathbb{E}_\pi[\mu((\bar{f} + \varepsilon - s_m)^2)] \leq 2(\mathbb{E}_\pi[\mu_x((\bar{f} - s_m)^2)] + \mathbb{E}_\pi[\varepsilon(X,Y)^2]) = 2(\mathbb{E}_\pi[\mu_x((\bar{f} - s_m)^2)] + M_0).$$

Hence we obtain

$$\mathbb{E}_\pi[\mu((f - s_m)^2)] \leq 8C_\kappa M_0 + 4C_\kappa m^{-1} \mu_x(L^{-1}g)^2).$$

Using Proposition 5 and noting that $m = O(n)$, we can write

$$\mathbb{E}_\pi[(\hat{\mu} - \mu)^2] = O(C_\kappa(C_f n^{-2} + M_0 n^{-1})).$$

□

Consider a case where M_0 is a relatively small number compared with C_f. Then the bound in Theorem 4 essentially becomes

$$\mathbb{E}_\pi[(\hat{\mu} - \mu)^2] = O(n^{-2}).$$

Therefore even in the case that we know very little about Y, the CF method applied on X only still improves the Monte Carlo rate. Note that Theorem 4 provides a different rate from Oates et al. (2017), the reason being that in our proof we employ a more refined inequality developed in Section 3 together with a different approximation approach.

5 BIASED GENERATING DISTRIBUTION

In this section, we digress our investigation and consider the case where we could only generate X from a distribution q_x different from π_x. We have explicit closed-form formula for π_x as described in Section 2, but we may not have that for q_x (though Monte Carlo samples are available). For this section, suppose the auxiliary variable Y does not show up (so we write $q = q_x$, $\pi = \pi_x$) and the regression function is f itself, leaving the discussion with the presence of auxiliary variables to future work.

We use the same CF method on X: We construct the RKHS $\mathscr{H}$ and $\mathscr{H}_+$ based on π, then construct the estimator s_m in exactly the same way (note that s_m only depends on $\mathscr{H}_+$ and the data, free of the underlying distribution), and we obtain the formula for $\mu_x(s_m)$ as we did in Lemma 6. Our goal is still to estimate μ. In this case, we do not have unbiasedness anymore since $\mathbb{E}_q[f_m(X) - \mu | D_0]$ is not necessarily equal to 0. However, we will see that we can still construct reasonable estimators for μ under q under the CF framework.

We introduce further notations. For any measurable function $g : \Omega \to \mathbb{R}$, we write $\nu_x(g) = \int_\Omega g(x)q_x(x)dx$. Let $L^2(q_x)$ denote the space of measurable functions $g : \Omega \to \mathbb{R}$ for which $\nu_x(g^2)$ is finite, with the norm written as $\|\cdot\|_{L^2(q_x)}$. Let $L_q : L^2(q_x) \to \mathscr{H}_+$ denote the integral operator

$$(L_q g)(x) := \int_\Omega k(x,x')g(x')q_x(x')dx',\ x \in \Omega,\ g \in L^2(q_x).$$

While s_m is constructed from $\mathscr{H}_+$ (induced by the original distribution π), the results on RLS that we developed in Section 3 can still be applied. In fact, π in Section 3 (independent of the choice of RKHS) stands for the underlying distribution of the samples which is exactly q in this section. In particular, Corollary 3 (with distribution q) is still valid in the current case.

We introduce the following assumptions that will be used in this section:

Assumption 9 $f \in (\mathscr{H}_+)_1^q$.

Assumption 10 $\mathbb{E}_\pi[\pi_x/q_x] < \infty$.

The following theorem reveals that the estimator $\mu_x(s_m)$ (computed in Lemma 6) is a rough approximation to μ under q. (For this estimator, we set $n = m$ and use the entire data set D to construct s_m.)

Theorem 5 Suppose Assumptions 1, 2, 3, 4, 9 and 10 hold and take a RLS estimate with $\lambda = m^{-1/2}$. Then the estimator $\mu_x(s_m)$ is an estimator of μ with

$$\mathbb{E}_q[(\mu_x(s_m) - \mu))^2] = O(m^{-\frac{1}{2}}).$$

□

Proof. First we note that $f_q = f$, $M_0 = 0$ and

$$f \in (\mathscr{H}_+)_1^q \subset \text{Range}(L_q^{\frac{1}{2}}).$$

It follows from Corollary 3 that

$$\mathbb{E}_q[\nu_x((f - s_m)^2)] \le C_\kappa m^{-\frac{1}{2}} \nu_x((L_q^{-1/2} f)^2) = C_\kappa m^{-\frac{1}{2}} \|f\|^2_{\mathscr{H}_+}$$

where $C_\kappa = 2\kappa^4 + 2$. Next, we have

$$|\mu_x(s_m) - \mu| \le \mathbb{E}_\pi[|s_m - f|] = \int_\Omega |s_m(t) - f(t)|\pi_x(t)dt.$$

It follows from Cauchy-Schwarz inequality that

$$\begin{aligned}\left(\int_\Omega |s_m(t) - f(t)|\pi_x(t)dt\right)^2 &\le \left(\int_\Omega |s_m(t) - f(t)|^2 q_x(t)dt\right)\left(\int_\Omega \frac{(\pi_x(t))^2}{q_x(t)}dt\right) \\ &= \nu_x((f - s_m)^2)\mathbb{E}_\pi[\pi_x/q_x].\end{aligned}$$

Therefore we obtain

$$\mathbb{E}_q[(\mu_x(s_m) - \mu))^2] \le \mathbb{E}_q[\nu_x((f - s_m)^2)](\mathbb{E}_\pi[\pi_x/q_x])^2.$$

Then the result follows from Assumption 10. □

Theorem 5 implies that the CF estimator still retains consistency regardless of the generating distribution of X, as long as this distribution is close to the target distribution in the sense of a controllable likelihood ratio. However, the convergence rate is subcanonical. This theorem can be compared with the corresponding results in importance sampling (e.g., Liu and Lee 2017) that achieves a better convergence $\mathbb{E}_q[(\hat{\mu} - \mu))^2] = O(m^{-1})$. In our future work, we will refine our analysis to improve our convergence rate as well as expanding the analyses to more general settings.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1653339/1834710 and IIS-1849280.

REFERENCES

Chwialkowski, K., H. Strathmann, and A. Gretton. 2016. "A Kernel Test of Goodness of Fit". In *Proceedings of the 33rd International Conference on Machine Learning*, edited by M. F. Balcan and K. Q. Weinberger, Volume 48, 2606–2615. New York, NY, USA: JMLR.org.

Cucker, F., and S. Smale. 2002. "Best Choices for Regularization Parameters in Learning Theory: On the BiasVariance Problem". *Foundations of Computational Mathematics* 2(4):413–428.

Glasserman, P., and B. Yu. 2005. "Large Sample Properties of Weighted Monte Carlo Estimators". *Operations Research* 53(2):298–312.

Liu, Q., and J. Lee. 2017. "Black-box Importance Sampling". In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, edited by A. Singh and J. Zhu, Volume 54, 952–961. Fort Lauderdale, FL, USA: PMLR.

Liu, Q., J. D. Lee, and M. Jordan. 2016. "A Kernelized Stein Discrepancy for Goodness-of-fit Tests". In *Proceedings of the 33rd International Conference on Machine Learning*, edited by M. F. Balcan and K. Q. Weinberger, Volume 48, 276–284. New York, NY, USA: JMLR.org.

Oates, C. J., J. Cockayne, F.-X. Briol, and M. Girolami. 2019. "Convergence Rates for a Class of Estimators Based on Stein's Method". *Bernoulli* 25(2):1141–1159.

Oates, C. J., M. Girolami, and N. Chopin. 2017. "Control Functionals for Monte Carlo Integration". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79(3):695–718.

Smale, S., and D.-X. Zhou. 2005. "Shannon Sampling II: Connections to Learning Theory". *Applied and Computational Harmonic Analysis* 19(3):285–302.

Sołtan, P. 2018. *A Primer on Hilbert Space Operators*. Cham, Switzerland: Birkhäuser Basel.

Sun, H., and Q. Wu. 2009. "Application of Integral Operator for Regularized Least-square Regression". *Mathematical and Computer Modelling* 49(1-2):276–285.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is khl2114@columbia.edu.

HAOFENG ZHANG is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. His obtained his bachelor's degree in mathematics from the University of Science and Technology of China. His email address is hz2553@columbia.edu.