Some New Perspectives on the Method of Control Variates

Peter W. Glynn and Roberto Szechtman
Email: glynn@stanford.edu, rober@stanford.edu

Department of Management Science and Engineering, Stanford University,
Stanford CA 94305, USA

Abstract. The method of control variates is one of the most widely used variance reduction techniques associated with Monte Carlo simulation. This paper studies the method of control variates from several different viewpoints, and establishes new connections between the method of control variates and: conditional Monte Carlo, antithetics, rotation sampling, stratification, and nonparametric maximum likelihood. We also develop limit theory for the method of control variates under weak assumptions on the estimator of the optimal control coefficient.

1 Introduction

Suppose that a simulation analyst wishes to compute a quantity α that can be expressed as the expectation of a real-valued random variable (rv) X, so that $\alpha = EX$. The conventional sampling-based algorithm for computing α involves simulating n independent and identically distributed (iid) copies of the rv X, denoted $X_1, \cdots, X_n$. The corresponding estimator for α is then just the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

However, in many situations, the analyst can take advantage of existing problem structure so as to create a more efficient means of computing α. One powerful approach to exploiting problem structure is the method of control variates. Specifically, suppose that there exists a random variable Y, jointly distributed with X, for which EY is known. Then, the control variate $C = Y - EY$ is guaranteed to be a "mean zero" random variable, so that $X(\lambda) = X - \lambda C$ is an estimator for α. Consequently, this creates the possibility of computing α by generating iid copies $(X_1, C_1), \cdots, (X_n, C_n)$ of the pair (X, C) and estimating α via $\bar{X}_n(\lambda) = \bar{X}_n - \lambda \bar{C}_n$, where $\bar{C}_n = n^{-1} \sum_{i=1}^{n} C_i$ and $\lambda \in \mathbb{R}$ is an arbitrary scalar. By choosing the "control coefficient" λ judiciously, one can therefore obtain a variance reduction relative to the conventional estimator $\bar{X}_n$.

In particular, it is natural to choose λ so as to minimize the variance of $X(\lambda)$. The variance-minimizing choice is just

$$\lambda^* = \operatorname{cov}(X, C) / \operatorname{var} C.$$

Since λ^* involves moment quantities that are generally unknown to the analyst, it must be estimated somehow. Fortunately, one of the great strengths of Monte Carlo sampling-based methodology is its ability to internally estimate such problem-dependent parameters by using the sample moments of the (X_i, C_i)'s to compute the population moments of (X, C). In particular, λ^* can be estimated via

$$\lambda_1(n) = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(C_i - \bar{C}_n)}{\sum_{i=1}^n (C_i - \bar{C}_n)^2},$$

thereby suggesting the estimator $\alpha_1(n) = \bar{X}_n - \lambda_1(n)\bar{C}_n$. This estimator for α can easily be implemented in many practical problem settings. In particular, relative to the conventional estimator $\bar{X}_n$, the estimator $\alpha_1(n)$ requires only that the simulation code collect the control variate outcomes $C_1, \cdots, C_n$, compute $\lambda_1(n)$, and form $\alpha_1(n)$.

Note that this variance-reduction method is "non-invasive", in the sense that it requires only that additional statistics be collected during the simulation run, and it does not require that the simulation analyst modify the code that is used to generate the X_i's themselves. Thus, control variates can be implemented (for example) in the background, while some visualization involving the X_i's proceeds in the foreground.

Because of the relative ease with which control variates can be implemented, it is perhaps the most widely applied of all variance reduction techniques. It is therefore of some interest and importance to understand this method in greater depth. This paper is intended to provide new insights into the foundations of the method of control variates, and to describe certain connections between the method of control variates and some other widely used variance reduction techniques. The following results are among the major contributions of this paper:

1. a complete discussion, in Section 2, of the method of control variates from the Hilbert space viewpoint (associated with the L^2 space of square-integrable random variables);
2. limit theory (see Theorem 1) that establishes that any consistent estimator of $\boldsymbol{\lambda}^*$ yields "first order" asymptotic optimality;
3. a discussion, in Section 3, of how to construct control variates in the setting of terminating simulations that involve a random number of input random variables;
4. an argument, in Section 5, that proves that control variates reduces variance "cooperatively" with the method of conditional Monte Carlo;
5. a discussion, in Sections 4, 6, and 7, of the relationships between control variates and numerical integration, antithetics, stratification, and rotation sampling;

6. an introduction to "constrained Monte Carlo", as well as a proof that constrained Monte Carlo coincides asymptotically with control variates in the equality-constrained context.

For additional discussion of the basic ideas underlying the method of control variates (as well as a range of different applications in which the idea is relevant), see, for example, [12], [4], [11], or [6].

2 Multiple Control Variates

The method of control variates easily generalizes to the setting in which there is a collection of random variables $Y_1, \cdots, Y_d$ (jointly distributed with X) for which EY_i is known for $1 \leq i \leq d$. Set

$$\boldsymbol{C} = (Y_1 - EY_1, \cdots, Y_d - EY_d)^T.$$

Let $(X, \boldsymbol{C}_1), \cdots, (X_n, \boldsymbol{C}_n)$ be n iid replicates of the random variable $(X, \boldsymbol{C})$. For $\boldsymbol{\lambda} \in \mathbb{R}^d$ (with $\boldsymbol{\lambda}$ encoded as a column vector), let $X(\boldsymbol{\lambda}) = X - \boldsymbol{\lambda}^T \boldsymbol{C}$ and put

$$\bar{X}_n(\boldsymbol{\lambda}) = \bar{X}_n - \boldsymbol{\lambda}^T \bar{\boldsymbol{C}}_n.$$

As in the scalar setting, the goal is to now choose $\boldsymbol{\lambda}$ so as to minimize the variance of $X(\boldsymbol{\lambda})$. Assume that $E(X^2 + \boldsymbol{C}^T \boldsymbol{C}) < \infty$. Then,

$$\mathrm{var}\, X(\boldsymbol{\lambda}) = \mathrm{var}\, X - 2\boldsymbol{\lambda}^T EX\boldsymbol{C} + \boldsymbol{\lambda}^T E\boldsymbol{C}\boldsymbol{C}^T \boldsymbol{\lambda}.$$

To compute the minimizing $\boldsymbol{\lambda}^*$, there are several equivalent approaches one can follow, each of which offers insight into the general principles underlying the method of control variates.

Approach 1. Use calculus to minimize $\mathrm{var}\, X(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda} \in \mathbb{R}^d$. Assuming the covariance matrix $E\boldsymbol{C}\boldsymbol{C}^T$ is non-singular, the "first-order optimality" conditions yield

$$\boldsymbol{\lambda}^* = (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} EX\boldsymbol{C}. \tag{1}$$

It is easily shown, via the "second-order optimality conditions", that $\boldsymbol{\lambda}^*$ is indeed the variance-minimizing choice of $\boldsymbol{\lambda}$.

Approach 2. Let L^2 be the space of all square-integrable random variables that are defined on the same probability space as X, equipped with inner product $\langle Z_1, Z_2 \rangle = EZ_1 Z_2$ and norm $\|Z\| = (EZ^2)^{1/2}$ for $Z, Z_1, Z_2 \in L^2$. Also, let $\mathcal{G}$ be the linear subspace of all random variables $W \in L^2$ such that $W = \boldsymbol{\lambda}^T \boldsymbol{C}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^d$. It is easily seen that $\mathcal{G}$ is a closed linear subspace of L^2. Observe that

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^d} \mathrm{var}\, X(\boldsymbol{\lambda}) = \min_{W \in \mathcal{G}} \|X - \alpha - W\|,$$

so that minimizing variance is equivalent to finding the closest point $W \in \mathcal{G}$ to $X - \alpha$. The Hilbert space projection theorem ([9]) asserts that the minimizing $W^* = \boldsymbol{\lambda}^{*T} \boldsymbol{C} \in \mathcal{G}$ is characterized via the relation

$$\langle X - W^*, W \rangle = 0$$

for $W \in \mathcal{G}$. In other words,

$$EWW^* = E(X - \alpha)W = EXW$$

for $W \in \mathcal{G}$. Hence,

$$\boldsymbol{\lambda}^T E \boldsymbol{C} \boldsymbol{C}^T \boldsymbol{\lambda}^* = \boldsymbol{\lambda}^T E X \boldsymbol{C}$$

for all $\boldsymbol{\lambda} \in \mathbb{R}^d$. This implies that $\boldsymbol{\lambda}^*$ must satisfy the equations

$$E \boldsymbol{C} \boldsymbol{C}^T \boldsymbol{\lambda}^* = E X \boldsymbol{C}.$$

We will exploit this Hilbert space viewpoint in Section 5's discussion of the relationship of control variates with conditional Monte Carlo.

Approach 3. By applying Gram-Schmidt orthogonalization ([9]), we may assume that the d components of the control variate vector are orthogonal in our L^2 inner product. Equivalently, since $E \boldsymbol{C} \boldsymbol{C}^T$ is symmetric and non-negative definite, one can express $E \boldsymbol{C} \boldsymbol{C}^T$ as RDR^T, where D is a diagonal matrix with non-negative diagonal entries and $RR^T = I$. Set $\boldsymbol{B} = R \boldsymbol{C}$. Then $E \boldsymbol{B} \boldsymbol{B}^T = D$, so that the components of $\boldsymbol{B}$ are orthogonal in our inner product, and the variance of the i'th component of $\boldsymbol{B}$ is D_{ii} (for $1 \leq i \leq n$).

We may now choose to view $\boldsymbol{B}$ as our vector of controls (rather that $\boldsymbol{C}$). This suggests finding $\boldsymbol{\beta}^* \in \mathbb{R}^d$ so as to minimize $\text{var}(X - \boldsymbol{\beta}^T \boldsymbol{B})$ over $\boldsymbol{\beta} \in \mathbb{R}^d$. Our discussion above establishes that

$$\beta_i^* = \frac{\langle X, B_i \rangle}{\langle B_i, B_i \rangle}$$

(provided $\langle B_i, B_i \rangle \neq 0$), so that

$$\boldsymbol{\beta}^* \boldsymbol{B} = \sum_{i=1}^d \frac{\langle X, B_i \rangle}{\langle B_i, B_i \rangle} B_i.$$

The right hand side of this expression is precisely what is obtained when one attempts to express X via an "orthogonal expansion" in terms of the rv's $B_1, \cdots, B_d$. Thus, the optimal linear combination of control variates can be viewed in terms of an orthogonal expansion. If X is in the "span" of the "basis vectors" $B_1, \cdots, B_d$, the variance of $X(\boldsymbol{\lambda}^*)$ is reduced to zero. In general, $X(\boldsymbol{\lambda}^*)$ has variability in proportion to that component of X that can not be explained in terms of the basis $B_1, \cdots, B_d$ (i.e. lies in the "orthogonal

complement" to $B_1, \cdots, B_d$).

The above analysis provides several different theoretical justifications for the formula (1) for $\boldsymbol{\lambda}^*$. As in the case of a scalar control, $\boldsymbol{\lambda}^*$ must typically be estimated from the sample data $(X_1, \boldsymbol{C}_1), \cdots, (X_n, \boldsymbol{C}_n)$. The vector analog of $\lambda_1(n)$ is, of course,

$$\boldsymbol{\lambda}_1(n) = (\sum_{i=1}^{n}(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n)(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n)^T)^{-1} \sum_{i=1}^{n}(X_i - \bar{X}_n)(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n).$$

It should be noted, however, that because it is known that $E\boldsymbol{C} = 0$, we can alternatively estimate $\boldsymbol{\lambda}^*$ via any of the following estimators:

$$\boldsymbol{\lambda}_2(n) = (\sum_{i=1}^{n}\boldsymbol{C}_i\boldsymbol{C}_i^T)^{-1} \sum_{i=1}^{n}(X_i - \bar{X}_n)(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n),$$

$$\boldsymbol{\lambda}_3(n) = (\sum_{i=1}^{n}(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n)(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n)^T)^{-1} \sum_{i=1}^{n}X_i\boldsymbol{C}_i,$$

$$\boldsymbol{\lambda}_4(n) = (\sum_{i=1}^{n}\boldsymbol{C}_i\boldsymbol{C}_i^T)^{-1} \sum_{i=1}^{n}X_i\boldsymbol{C}_i.$$

In addition, one can apply the above idea component-wise, choosing to center some components of $\boldsymbol{C}_i$ by their corresponding sample means and leaving the other components uncentered.

But other estimators for $\boldsymbol{\lambda}^*$ may also be available. In many applications settings, the entire joint distribution of $\boldsymbol{C}$ may be known in closed form, so that the simulation analyst may know $E\boldsymbol{C}\boldsymbol{C}^T$. This suggests the estimators

$$\boldsymbol{\lambda}_5(n) = (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \sum_{i=1}^{n}(X_i - \bar{X}_n)(\boldsymbol{C}_i - \bar{\boldsymbol{C}}_n),$$

$$\boldsymbol{\lambda}_6(n) = (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \sum_{i=1}^{n}X_i\boldsymbol{C}_i$$

for $\boldsymbol{\lambda}^*$. Furthermore, it may be that only part of the covariance structure of $\boldsymbol{C}$ is known to the analyst. Among the possibilities that exist are:

1. the joint distribution of only a subset of the control vector components is known, so that the elements of a principal submatrix of $E\boldsymbol{C}\boldsymbol{C}^T$ are known, with the other elements unknown. (By re-labeling the components of $\boldsymbol{C}$ if necessary, this is equivalent to assume that some "northwest" square sub-block of $E\boldsymbol{C}\boldsymbol{C}^T$ is known, with the rest of $E\boldsymbol{C}\boldsymbol{C}^T$ unknown.)
2. Some of the control variate components are known to be orthogonal with respect to one another, so that the off-diagonal entries corresponding to

pairs of such orthogonal random variables are guaranteed to be zero. This can arise if some subset of controls forms a collection of mutually independent random variables. It may also occur in some settings in which X arises from the output of a stochastic process which is correlated with an associated martingale; the martingale differences can then be used as individual control variates, and such martingale differences are guaranteed to be orthogonal with respect to one another.

In both of the above circumstances, one may choose to substitute the known covariances for the corresponding covariance estimator, with or without "centering by the sample mean". Thus, in the context of a vector-valued control $\boldsymbol{C}$, an enormous number of different estimators for $\boldsymbol{\lambda}^*$ suggest themselves. The following result shows that regardless of the choice of estimator for $\boldsymbol{\lambda}^*$, the resulting estimator for α achieves the maximal degree of variance reduction possible (subject only to the stipulation that the estimator for $\boldsymbol{\lambda}^*$ be consistent).

Let $\Rightarrow$ denote weak convergence and suppose that $N(\mu, \sigma^2)$ is a normally distributed random variable with mean μ and variance σ^2.

Theorem 1. Assume that $E\boldsymbol{C}\boldsymbol{C}^T$ is non-singular and that $E(X^2 + \boldsymbol{C}^T\boldsymbol{C}) < \infty$. Suppose that $E\boldsymbol{C} = 0$ and that $\boldsymbol{\lambda}(n) \Rightarrow \boldsymbol{\lambda}^*$ as $n \to \infty$. Set $\alpha_c(n) = \bar{X}_n - \boldsymbol{\lambda}(n)^T \bar{C}_n$. Then,

$$n^{1/2}(\bar{X}_n(\boldsymbol{\lambda}^*) - \alpha_c(n)) \Rightarrow 0$$

as $n \to \infty$, so that

$$n^{1/2}(\alpha_c(n) - \alpha) \Rightarrow \sqrt{\operatorname{var} X(\boldsymbol{\lambda}^*)} N(0, 1)$$

as $n \to \infty$. In addition,

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \boldsymbol{\lambda}(n)^T \boldsymbol{C}_i)^2 \Rightarrow EX(\boldsymbol{\lambda}^*)^2$$

as $n \to \infty$.

Proof. Note that $\bar{X}_n(\boldsymbol{\lambda}^*) - \alpha_c(n) = (\boldsymbol{\lambda}(n) - \boldsymbol{\lambda}^*)^T \bar{C}_n$. Furthermore, the multivariate central limit theorem (CLT) ensures that

$$n^{-1/2}\bar{C}_n \Rightarrow N(\boldsymbol{0}, E\boldsymbol{C}\boldsymbol{C}^T)$$

as $n \to \infty$, where $N(\boldsymbol{0}, E\boldsymbol{C}\boldsymbol{C}^T)$ is a d-dimensional multivariate normal random vector with mean vector $\boldsymbol{0}$ and covariance matrix $E\boldsymbol{C}\boldsymbol{C}^T$. Since $\boldsymbol{\lambda}(n) \Rightarrow \boldsymbol{\lambda}^*$, it follows from a "converging together" argument (see [2], p. 27) that

$$n^{1/2}(\bar{X}_n(\boldsymbol{\lambda}^*) - \alpha_c(n)) = (\boldsymbol{\lambda}(n) - \boldsymbol{\lambda}^*)^T \cdot n^{1/2}\bar{C}_n \Rightarrow \boldsymbol{0} \cdot N(\boldsymbol{0}, E\boldsymbol{C}\boldsymbol{C}^T) = 0$$

as $n \to \infty$, proving the first assertion. For the second result, observe that

$$(X_i - \boldsymbol{\lambda}^{*T}\boldsymbol{C}_i)^2 - (X_i - \boldsymbol{\lambda}(n)^T\boldsymbol{C}_i)^2 = (2X_i - (\boldsymbol{\lambda}(n) + \boldsymbol{\lambda}^*)^T\boldsymbol{C}_i)(\boldsymbol{C}_i^T(\boldsymbol{\lambda}(n) - \boldsymbol{\lambda}^*)),$$

so

$$\frac{1}{n-1}\sum_{i=1}^{n} X_i(\boldsymbol{\lambda}^*)^2 - \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \boldsymbol{\lambda}(n)^T\boldsymbol{C}_i)^2$$

$$= \frac{2}{n-1}\sum_{i=1}^{n} X_i\boldsymbol{C}_i^T(\boldsymbol{\lambda}(n) - \boldsymbol{\lambda}^*) - (\boldsymbol{\lambda}(n) + \boldsymbol{\lambda}^*)^T \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{C}_i\boldsymbol{C}_i^T(\boldsymbol{\lambda}(n) - \boldsymbol{\lambda}^*)$$

$$\Rightarrow 2E X\boldsymbol{C}^T \cdot \boldsymbol{0} - 2\boldsymbol{\lambda}^{*T} \cdot E\boldsymbol{C}\boldsymbol{C}^T \cdot \boldsymbol{0} = \boldsymbol{0}$$

as $n \to \infty$. Since $n^{-1}\sum_{i=1}^{n} X_i(\boldsymbol{\lambda}^*)^2 \to E X(\boldsymbol{\lambda}^*)^2$ a.s. as $n \to \infty$, this establishes the second assertion ∎

Theorem 1 asserts that all estimators of α in which $\boldsymbol{\lambda}(n)$ is chosen to be consistent for $\boldsymbol{\lambda}^*$ are within $o_p(n^{-1/2})$ of one another (where $o_p(a_n)$ denotes a sequence of rv's $(\xi_n : n \geq 1)$ such that $a_n^{-1}\xi_n \Rightarrow 0$ as $n \to \infty$). Thus, at the level of "first-order" central limit theory, the choice of $\boldsymbol{\lambda}(n)$ is irrelevant. The specific choice of the estimator $\boldsymbol{\lambda}(n)$ can therefore only make a difference at the "second-order" level.

Of course, it should be noted that knowledge of all or part of the covariance structure of $\boldsymbol{C}$ permits one to potentially use the corresponding sample covariances (centered by their known population values) as control variates. By adding such additional control variates, one can obtain further (asymptotic) variance reduction that would reflect itself at a "first order" level by reducing the magnitude of the variance associated with $X(\boldsymbol{\lambda}^*)$.

The second assertion of Theorem 1 is important in that it offers theoretical support for the variance estimation needed to construct confidence intervals for α. Specifically, let $v(n) = (n^{-1}\sum_{i=1}^{n}(X_i - \boldsymbol{\lambda}(n)^T\boldsymbol{C}_i))^2 - \alpha_c(n)^2$. Under the conditions of Theorem 1 (and assuming that var $X(\boldsymbol{\lambda}^*) > 0$), Theorem 1 permits us to assert that

$$P(\alpha \in [L_n, R_n]) \to 1 - \delta$$

as $n \to \infty$, where

$$L_n = \alpha_c(n) - z\sqrt{v(n)/n},$$

$$R_n = \alpha_c(n) + z\sqrt{v(n)/n},$$

and z is selected so that $P(-z \leq N(0,1) \leq z) = 1 - \delta$. Hence the interval $[L_n, R_n]$ is guaranteed to be an asymptotic $100(1 - \delta)\%$ confidence interval for α (with asymptotic half-width $n^{-1/2} \cdot z\sqrt{\text{var } X(\boldsymbol{\lambda}^*)}$).

3 Construction of Control Variates for Terminating Simulations

Suppose that generation of the rv X involves simulating an associated stochastic process over some finite time horizon. In many applications, the stochastic process is itself driven by an associated iid "noise" sequence of random input vectors. The number τ of such input vectors needed to simulate the process over the given time horizon is generally random. For example, in simulating a continuous-time Markov chain to a deterministic time t, the number of exponential random variables required is random.

To make our idea more concrete, suppose that $\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_\tau$ is the sequence of iid random (column) vectors required to generate X. Typically $E\boldsymbol{V}_j$ is known in closed form. It is therefore tempting to use

$$\frac{1}{\tau}\sum_{i=1}^{\tau} \boldsymbol{V}_i - E\boldsymbol{V}_i$$

as a control variate. While this centered sample mean is indeed a control (with mean zero) when τ is deterministic, it generally is not a control when τ is random, for in that case

$$E(\tau^{-1}\sum_{i=1}^{\tau}\boldsymbol{V}_i) \neq E\boldsymbol{V}_i$$

usually holds. (However, if τ is large, the difference between the left-hand side and right-hand side is generally small.)

Fortunately, there is typically an easy way to remedy the situation. In general, the rv τ will be a stopping time that is adapted to a filtration to which the $\boldsymbol{V}_j$'s are also adapted. If $E\tau < \infty$ and $E(\sum_{i=1}^{\tau}|\boldsymbol{V}_j|) < \infty$, Wald's first moment identity asserts ([3]) that

$$E\left(\sum_{i=1}^{\tau}\boldsymbol{V}_i\right) = E\boldsymbol{V}_i \cdot E\tau,$$

so that $\boldsymbol{C} = \sum_{i=1}^{\tau}(\boldsymbol{V}_i - E\boldsymbol{V}_i)$ is a control variate satisfying $E\boldsymbol{C} = \boldsymbol{0}$. Note that $\boldsymbol{C}$ is a legitimate mean-zero control even if the $\boldsymbol{V}_i$'s are non-identically distributed in i (provided that they are independent in i); this follows from the generalized Wald's first moment identity.

Furthermore, if the $\boldsymbol{V}_i$'s are iid with finite second moments and $E\tau^2 < \infty$, then Wald's second moment identity ensures that

$$E\boldsymbol{C}\boldsymbol{C}^T = E\boldsymbol{V}_1\boldsymbol{V}_1^T \cdot E\tau.$$

Consequently, if the covariance structure of $\boldsymbol{V}_1$ is known, the covariance structure of $\boldsymbol{C}$ is known up to the scalar $E\tau$. This offers the opportunity, as discussed in Section 2, to develop estimators of the optimal control coefficient $\boldsymbol{\lambda}^*$ that can take advantage of the known covariance structure of the $\boldsymbol{V}_i$'s.

For additional discussion of the subtleties that can arise in constructing control variates in the terminating simulating setting, see [10].

4 Connections to Numerical Integration

The method of control variates has a counterpart that arises in the theory of numerical integration. Consider, for example, the problem of numerically computing the integral

$$\alpha = \int_0^1 f(x)dx,$$

where $f(x) \sim x^{-1/2}$ as $x \downarrow 0$. A standard "trick" in numerical integration is to "subtract out the singularity" (namely, the singularity in f at $x = 0$), so that we write

$$\alpha = \int_0^1 [f(x) - x^{-1/2}]dx + \int_0^1 x^{-1/2} = \int_0^1 g(x)dx + 2,$$

where $g(x) = f(x) - x^{-1/2}$ is then free of the singularity at zero (and consequently easier to numerically integrate via, for example, a quadratic rule). The key to the idea is to take advantage of the fact that a closely related integrand can be analytically integrated; see [5], p. 470, for further details on this technique. Clearly, the philosophy underlying this approach is essentially identical to that associated with the method of control variates.

However, control variates and conventional numerical integration can interact in other ways, as well. For example, in certain contexts, it may be that the distribution of the rv Y (jointly distributed with X) is known, and yet no analytic closed-form for EY may be available. Typically, computing EY will then involve a one-dimensional integration with respect to the probability density of Y. This integration can be implemented numerically via an integration scheme that is not sampling-based, such as a quadrature scheme (or even an approach base on quasi-random sequences). Suppose that $\mu(n)$ is the resulting numerical approximation to EY obtained after expending a computational effort of n floating-point operations. Typically, such numerical integration schemes enjoy a rate of convergence of the form

$$\mu(n) = EY + O(n^{-1/2-p})$$

as $n \to \infty$, for some $p > 0$ (where $O(a_n)$ represents a deterministic sequence having an absolute value that can be bounded by some finite multiple of a_n).

Given a computer budget of c floating point operations, suppose that we then allocate $\lfloor c^\delta \rfloor$ operations to the numerical computation of EY and the remainder of the budget allocated to sampling of the pair (X, Y), with $\delta = (1+p)/(1+2p)$. We estimate $\alpha = EX$ via a sample mean of the rv $X - \lambda(Y - \mu(\lfloor c^\delta \rfloor))$. But $c^{1/2}(\mu(\lfloor c^\delta \rfloor) - EY) \to 0$, as $c \to \infty$ and the fraction of the budget allocated to computing $\mu(\lfloor c^\delta \rfloor)$ is converging to zero as $c \to \infty$. It follows easily that the sample mean has precisely the same "first-order" central limit behavior as that obtained by replicating the rv $X - \lambda(Y - EY)$. In other words, at the level of such a "first-order" analysis, our need to compute EY numerically via such an integration scheme has no impact whatsoever on the asymptotic convergence rate of our control variates methodology.

5 Connections to Conditional Monte Carlo

One powerful variance reduction technique that is often available in the presence of suitable problem structure is "conditional Monte Carlo". Suppose that the probability space supporting the rv X is the triplet $(\Omega, \mathcal{F}, P)$. It is often possible to explicitly compute $Z = E(X|\mathcal{H})$ for some sub-σ-field $\mathcal{H} \subseteq \mathcal{F}$, at the same time as one generates X. This provides us with two unbiased estimators of α, namely X and Z. As a consequence, $C = X - Z$ can be used as a control variate.

In this special setting, λ^* can be explicitly computed without a need to resort to estimation via sample covariances and sample variances.

Theorem 2. Suppose that $EX^2 < \infty$ and set $X(\lambda) = X - \lambda(X - E(X|\mathcal{H}))$. Then var $X(\lambda)$ is minimized over λ at $\lambda^* = 1$.

Proof. Suppose that $W \in L^2$ is $\mathcal{H}$-measurable. Then, the Cauchy-Schwarz inequality implies that XW and ZW are integrable rv's. Furthermore, the conditional expectation Z satisfies

$$EXW = EZW$$

for all such W. In particular,

$$EXZ = EZ^2,$$

so that $EZ(X - Z) = 0$. It follows that $EXC = EX(X - Z) = E(X - Z + Z)(X - Z) = E(X - Z)^2 + EZ(X - Z) = EC^2$, so that $\lambda^* = EXC/EC^2 = 1$. ∎

Thus, the best possible linear combination of X and Z is just $Z = E(X|\mathcal{H})$ itself. In other words, the conditional Monte Carlo estimator for α, based on a sample mean of iid copies of Z, is also optimal when viewed from a control

variates perspective.

Recall that for $X \in L^2$, the variance decomposition formula

$$\text{var } X = E \text{ var}(X|\mathcal{H}) + \text{var } E(X|\mathcal{H})$$

holds; see, for example, [4]. Use of conditional Monte Carlo reduces variance from var X (the variance associated with the conventional Monte Carlo estimator for α, based on replicating X itself) to var $E(X|\mathcal{H})$. Thus, conditional Monte Carlo eliminates the contribution to var X from $E \text{ var}(X|\mathcal{H})$.

We will now discuss how the method of control variates can frequently be used to reduce the contribution to var X that is contributed by var $E(X|\mathcal{H})$. In particular, suppose that $\mathcal{H}$ is the σ-algebra that is generated by a scalar rv Γ, and that the distribution of Γ is known. More precisely, suppose that the simulation analyst knows the entire moment sequence $(E\Gamma^k : k \geq 1)$ for the rv Γ. As a consequence, $Y_i = \Gamma^i$ has known mean for $i \geq 1$, so that for each $d \geq 1$,

$$\boldsymbol{C}(d) = (\Gamma - E\Gamma, \Gamma^2 - E\Gamma^2, \cdots, \Gamma^d - E\Gamma^d)^T$$

is a d-dimensional control variate with mean $\boldsymbol{0}$. We will now analyze the behavior of $\boldsymbol{C}(d)$ as $d \rightarrow \infty$.

Let $W^*(d) = \boldsymbol{\lambda}^*(d)^T \boldsymbol{C}(d)$ be the associated linear combination of the components of $\boldsymbol{C}(d)$ that minimizes $\text{var}(X - \boldsymbol{\lambda}^T \boldsymbol{C}(d))$ over $\boldsymbol{\lambda} \in \mathbb{R}^d$.

Theorem 3. Suppose that $E|\Gamma|^i < \infty$ for $i \geq 1$, and that $EX^2 < \infty$. Then,

$$\|W^*(d) - E(X - \alpha|\Gamma)\| \rightarrow 0$$

as $d \rightarrow \infty$, and $\text{var}(X - W^*(d)) \rightarrow E \text{ var}(X|\Gamma)$ as $d \rightarrow \infty$.

Proof. Let $\mathcal{G}_d$ be the linear subspace of all random variables $W \in L^2$ such that $W = \boldsymbol{\lambda}^T \boldsymbol{C}(d)$ for some $\boldsymbol{\lambda} \in \mathbb{R}^d$. Then, as discussed in Section 2,

$$\langle X - W^*(d), W \rangle = 0 \tag{2}$$

for $W \in \mathcal{G}_d$. Note that for $d_1 \leq d_2$, it follows that

$$EXW^*(d_1) = EW^*(d_1)W^*(d_2). \tag{3}$$

In addition,

$$\begin{aligned} EX^2 &= E(X - W^*(d) + W^*(d))^2 \\ &= E(X - W^*(d))^2 + 2EW^*(d)(X - W^*(d)) + EW^*(d)^2 \qquad (4) \\ &= E(X - W^*(d))^2 + EW^*(d)^2. \end{aligned}$$

Hence

$$
\begin{aligned}
\|W^*(d_2) &- W^*(d_1)\|^2 \\
&= EW^*(d_2)^2 + EW^*(d_1)^2 - 2EW^*(d_1)W^*(d_2) \\
&= EW^*(d_2)^2 + EW^*(d_1)^2 - 2EXW^*(d_1) \text{ (using (3))} \\
&= EW^*(d_2)^2 + EW^*(d_1)^2 - 2EXW^*(d_1) + EX^2 - EX^2 \\
&= E(W^*(d_1) - X)^2 + EW^*(d_2)^2 - EX^2 \\
&= EX^2 - EW^*(d_1)^2 + EW^*(d_2)^2 - EX^2 \text{ (using (4))} \\
&= \|W^*(d_2)\|^2 - \|W^*(d_1)\|^2,
\end{aligned}
$$

so $(\|W^*(d)\|^2 : d \geq 1)$ is a non-decreasing sequence. Furthermore, it is bounded above by $\|X\|^2$.

Clearly, for $\epsilon > 0$, there exists $d = d(\epsilon)$ such that $\|W^*(d_2) - W^*(d_1)\|^2 < \epsilon$ for $d_1, d_2 \geq d$. In other words, $(W^*(d) : d \geq 1)$ is a Cauchy sequence (in L^2) of $\mathcal{H}$-measurable random variables. It follows that there exists a $\mathcal{H}$-measurable rv $W^*(\infty) \in L^2$ such that

$$
\|W^*(d) - W^*(\infty)\| \to 0 \tag{5}
$$

as $d \to \infty$.

Note that (2) implies that

$$
EX\Gamma^i = EW^*(d)\Gamma^i
$$

for $d \geq i$. But the Cauchy-Schwarz inequality and (5) together establish that

$$
E(W^*(\infty) - W^*(d))\Gamma^i \to 0
$$

as $d \to \infty$. We conclude that

$$
EX\Gamma^i = EW^*(\infty)\Gamma^i
$$

for $i \geq 1$. Consequently, the collection of all $\mathcal{H}$-measurable rv's $W \in L^2$ for which

$$
E(\alpha + W^*(\infty))W = EXW \tag{6}
$$

contains the collection $\{\Gamma^i : i \geq 0\}$. Evidently, (6) therefore holds for all $\mathcal{H}$-measurable $W \in L^2$ ([3]). Since $W^*(\infty)$ is itself $\mathcal{H}$-measurable, we conclude that $W^*(\infty) = E(X|\Gamma) - \alpha$, proving the first assertion of the theorem.

For the second assertion,

$$
\begin{aligned}
\|X - W^*(\infty)\| &- \|W^*(\infty) - W^*(d)\| \\
&\leq \|X - W^*(d)\| \leq \|X - W^*(\infty)\| + \|W^*(\infty) - W^*(d)\|,
\end{aligned}
$$

so $\|X - W^*(d)\|^2 \to \|X - W^*(\infty)\|^2$ as $d \to \infty$. Hence, $\text{var}(X - W^*(d)) \to \text{var}(X - W^*(\infty))$ as $d \to \infty$. But $\text{var}(X - W^*(\infty)) = E(X - Z)^2$, where $Z = E(X|\Gamma)$. Since $\langle X - Z, Z \rangle = 0$, evidently

$$\begin{aligned} EX^2 &= E(X - Z + Z)^2 \\ &= E(X - Z)^2 + EZ^2, \end{aligned}$$

so $E(X - Z)^2 = EX^2 - EZ^2 = E(E(X^2|\Gamma)) - (E(X|\Gamma))^2 = E\,\text{var}(X|\Gamma)$, proving the second assertion. ∎

Theorem 3 proves that control variates eliminates that variability in X that is "orthogonal" to the variability eliminated by conditional Monte Carlo, in that control variates (based on powers of Γ) and conditional Monte Carlo (based on conditioning on Γ) are guaranteed to work cooperatively with one another in reducing variance. We note, in passing, that a weaker version of Theorem 3, with stronger hypotheses, appears in [10].

The principles that Theorem 3 illustrates hold much more generally. If conditional Monte Carlo is based on conditioning on a sub-σ-field $\mathcal{H}$, then $\mathcal{H}$-measurable control variates always reduce variance cooperatively. Furthermore, if a sequence of control variates is selected that ultimately spans (in L^2) the space of $\mathcal{H}$-measurable random variables, then the combination of conditional Monte Carlo and control variates can asymptotically reduce the variance to zero.

6   Connections to Antithetics

The method of antithetic variates is a sampling-based means of computing $\alpha = EX$ that attempts to induce an advantageous correlation between the copies of the rv X that are simulated. In particular, the idea is to generate a pair of copies of the rv X, say $X(0)$ and $X(1)$, that are correlated more advantageously than would be the case if the $X(i)$'s had been generated via iid sampling. Given that

$$\text{var}(\frac{1}{2}(X(0) + X(1))) = \frac{1}{2}(\text{var}\,X + \text{cov}(X(0), X(1))),$$

it follows that a necessary and sufficient condition for such a variance reduction (relative to iid sampling) is that $\text{cov}(X(0), X(1)) \leq 0$. See, for example, [12] for one means of inducing such a correlation. In the presence of such negative correlation, an estimator based on n iid samples of $(X(0) + X(1))/2$ has lower variance than one based on $2n$ iid samples of X.

However, a "control variates" possibility also presents itself. In particular, note that $C = X(0) - X(1)$ is a control variate having mean zero. Consequently,

$$X(\lambda) = X(0) - \lambda(X(0) - X(1))$$

is a rv, having mean α, that can be replicated to produce an estimator for $\alpha = EX$. The method of antithetics is recovered when $\lambda = 1/2$. We now prove that $\lambda = 1/2$ is indeed the universally optimal choice of the control coefficient.

Theorem 4. Suppose that $X(0) \overset{\mathcal{D}}{=} X(1) \overset{\mathcal{D}}{=} X$ (where $\overset{\mathcal{D}}{=}$ denotes "equality in distribution"). If $EX^2 < \infty$, then $\mathrm{var}\, X(\lambda)$ is minimized over λ at $\lambda^* = 1/2$.

Proof. Note that

$$\mathrm{var}\, X(\lambda) = [(1 - \lambda)^2 + \lambda^2 + 2\lambda(1 - \lambda)\rho]\,\mathrm{var}\, X,$$

where ρ is the coefficient of correlation between $X(0)$ and $X(1)$. At a minimizer λ^*, the derivative of $\mathrm{var}\, X(\lambda)$ must vanish, so that

$$2(\lambda^* - 1) + 2\lambda^* + (2 - 4\lambda^*)\rho = 0,$$

which has solution $\lambda^* = 1/2$. Since the second derivative is positive at λ^* if $\rho < 1$ (and non-negative if $\rho = 1$), it follows that $\lambda^* = 1/2$ is the minimizer. ∎

Thus, regardless of the covariance (or the sign of the covariance), $\lambda^* = 1/2$ is the optimal control coefficient (so that "equal weighting" of $X(0)$ and $X(1)$ minimizes variance).

A related concept is that of "rotation sampling"; see [7] and [8]. Here, the idea is to generate $d + 1$ correlated copies of X at a time, call them $X(0), \cdots, X(d)$, and to then replicate iid copies of the random vector $(X(0), \cdots, X(d))^T$, from which an estimator of α can then be constructed. The term "rotation sampling" arises from the fact that the $X(i)$'s are generated, via inversion, by adding $i(d + 1)^{-1}$ to each of the uniform rv's associated with $X(0)$, where addition is performed in "modulo one" arithmetic. Thus, the uniforms for the $X(i)$'s can be viewed as the uniforms for $X(0)$, suitably "rotated" around a circle of unit circumference. Note that we can embed $(X(i) : 0 \le i \le d)$ in a doubly infinite stationary sequence $\boldsymbol{X} = (X(i) : -\infty < i < \infty)$, in which $X(i) = X(i + d + 1)$ for all i.

As in the antithetic setting, a "control variates" possibility presents itself. In particular, let $C_i = X(0) - X(i)$ for $1 \le i \le d$, and put $\boldsymbol{C} = (C_1, \cdots, C_d)$. Then

$$X(\boldsymbol{\lambda}) = X(0) - \boldsymbol{\lambda}^T \boldsymbol{C}$$

is an unbiased estimator for $\alpha = EX$. An important issue is the choice of $\boldsymbol{\lambda}$ that minimizes the variance of $X(\boldsymbol{\lambda})$.

Theorem 5. Suppose that $\boldsymbol{X} = (X(i) : -\infty < i < \infty)$ is a stationary process for which $X(i) = X(i + d + 1)$ for all i. If $EX^2(0) < \infty$, then

var $X(\boldsymbol{\lambda})$ is minimized over $\boldsymbol{\lambda}$ by $\lambda_i = 1/(d+1)$ for $1 \le i \le d$.

Proof. Let U be a rv, independent of $\boldsymbol{X}$, that is uniformly distributed on the integers $0, 1, \cdots, d$. Note that by stationarity of $\boldsymbol{X}$,

$$X(0) - \boldsymbol{\lambda}^T \boldsymbol{C} = (1 - \sum_{i=1}^{d} \lambda_i) X(0) + \sum_{i=1}^{d} \lambda_i X(i) \overset{\mathcal{D}}{=} (1 - \sum_{i=1}^{d} \lambda_i) X(U) + \sum_{i=1}^{d} \lambda_i X(U+i).$$

Hence, by the Cauchy-Schwarz inequality,

$$
\begin{aligned}
EX(\boldsymbol{\lambda})^2 &= E\left(E\left[((1 - \sum_{i=1}^{d} \lambda_i) X(U) + \sum_{i=1}^{d} \lambda_i X(U+i))^2 | \boldsymbol{X} \right] \right) \\
&\ge E\left(E\left[(1 - \sum_{i=1}^{d} \lambda_i) X(U) + \sum_{i=1}^{d} \lambda_i X(U+i) | \boldsymbol{X} \right]^2 \right).
\end{aligned}
$$

But

$$E[X(U+i)|\boldsymbol{X}] = \frac{1}{d+1} \sum_{j=0}^{d} X(j).$$

So,

$$E\left[(1 - \sum_{i=1}^{d} \lambda_i) X(U) + \sum_{i=1}^{d} \lambda_i X(U+i) | \boldsymbol{X} \right] = \frac{1}{d+1} \sum_{j=0}^{d} X(j),$$

and consequently

$$
\begin{aligned}
EX(\boldsymbol{\lambda})^2 &\ge E\left(\frac{1}{d+1} \sum_{i=0}^{d} X(j) \right)^2 \\
&= EX(\boldsymbol{\lambda}^*)^2,
\end{aligned}
$$

proving the result. ∎

This theorem proves that the choice of "uniform weighting" on the $X(i)$'s is the variance-minimizing choice for the control coefficients.

7   Connections to Stratification

Suppose that $\alpha = EX$, where X is jointly distributed with some rv Γ. Assume that the range of Γ can be partitioned into a collection of sets $A_1, A_2, \cdots, A_d$ for which $p_i = P(\Gamma \in A_i)$ is known for $1 \le i \le d$ and for which it is possible to generate variates from each of the d conditional distributions

$$P_i(\cdot) = P(X \in \cdot | \Gamma \in A_i).$$

For $1 \leq i \leq d$, let $\bar{X}_n(i)$ be an independently generated sample mean based on n iid replications of the rv $X(i)$. For a given total sample size n, we can then estimate α via

$$\sum_{i=1}^{d} p_i \bar{X}_{\lfloor c_i n \rfloor}(i),$$

where $(c_i : 1 \leq i \leq d)$ is a probability mass function that distributes the total sampling budget among the d "strata". Proportional stratification is the special case where $c_i = p_i$ for $1 \leq i \leq d$. If $EX^2 < \infty$, then it is well known that

$$n^{1/2} \left(\sum_{i=1}^{d} p_i \bar{X}_{\lfloor p_i n \rfloor}(i) - \alpha \right) \Rightarrow \eta N(0,1)$$

as $n \to \infty$, where $\eta^2 \leq \operatorname{var} X$; see, for example, [12]. Thus, proportional stratification is always more efficient that iid sampling. Of course, one (major) disadvantage of stratification is its need to explicitly generate random variates from each of the distributions $\boldsymbol{p}_i(\cdot)$, $1 \leq i \leq d$.

A more easily implemented alternative is post-stratification. Let (X_1, Γ_1), $\cdots, (X_n, \Gamma_n)$ be n iid replications of the random vector (X, Γ) and put

$$\bar{X}(n,i) = \sum_{j=1}^{n} X_j I(\Gamma_j \in A_i) / \sum_{j=1}^{n} I(\Gamma_j \in A_i)$$

for $1 \leq i \leq d$. Note that conditional on $\sum_{j=1}^{n} I(\Gamma_j \in A_i) = k$, $\bar{X}(n,i) \overset{\mathcal{D}}{=} \bar{X}_k(i)$. This suggests the estimator

$$\sum_{i=1}^{d} p_i \bar{X}(n,i).$$

If $EX^2 < \infty$, the post-stratified estimator satisfies the CLT

$$n^{1/2} \left(\sum_{i=1}^{d} p_i \bar{X}(n,i) - \alpha \right) \Rightarrow \eta N(0,1)$$

as $n \to \infty$, so that the "first order" central limit theory allows us to conclude that the post-stratified estimator is as efficient as a stratified estimator with proportional stratification.

However, an obvious control variates possibility also presents itself. In particular, the assumption that the p_i's are known opens the possibility of using

$$\boldsymbol{C} = (I(\Gamma \in A_1) - p_1, \cdots, I(\Gamma \in A_d) - p_d)$$

as a d-dimensional control variate (having mean $\mathbf{0}$). Consequently,

$$X(\boldsymbol{\lambda}) = X - \boldsymbol{\lambda}^T \boldsymbol{C}$$

is a control variates estimand from which α can be estimated. A natural question here is the relationship between the variance minimizing $X(\boldsymbol{\lambda}^*)$ and the estimator obtained through post-stratification.

Theorem 6. Suppose that $EX^2 < \infty$. Then, $\text{var}(\boldsymbol{\lambda})$ is minimized over $\boldsymbol{\lambda}$ by the choice $\boldsymbol{\lambda}^*$, where $\lambda_i^* = E(X|\Gamma \in A_i)$ for $1 \le i \le d$. Furthermore,

$$n^{1/2} (\sum_{i=1}^{d} p_i \bar{X}(n, i) - \bar{X}_n(\boldsymbol{\lambda}^*)) \Rightarrow 0$$

as $n \to \infty$, where $\bar{X}_n(\boldsymbol{\lambda}^*)$ is the sample mean based on $X_1 - \boldsymbol{\lambda}^{*T} \boldsymbol{C}_1, \cdots, X_n - \boldsymbol{\lambda}^{*T} \boldsymbol{C}_n$.

Proof. Put $\mu_i = E(X|\Gamma \in A_i)$ for $1 \le i \le d$. It is easily seen that the i'th diagonal entry of $E\boldsymbol{C}\boldsymbol{C}^T$ is $p_i(1-p_i)$, whereas the (i,j)'th entry (for $i \ne j$) is $-p_i p_j$. Furthermore, the i'th entry of $EX\boldsymbol{C}$ is $(\mu_i - \alpha)p_i$. A simple computation establishes that the vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_d)$ satisfies $E\boldsymbol{C}\boldsymbol{C}^T \boldsymbol{\mu} = EX\boldsymbol{C}$. Consequently, our discussion of Section 2 proves that $\boldsymbol{\lambda}^* = \boldsymbol{\mu}$.

For the second assertion, note that if $\hat{p}_i(n) = n^{-1} \sum_{j=1}^{n} I(\Gamma_j \in A_i)$, then

$$\sum_{i=1}^{d} p_i \bar{X}(n, i) = \sum_{i=1}^{d} \hat{p}_i(n) \bar{X}(n, i) - \sum_{i=1}^{d} (\hat{p}_i(n) - p_i) \bar{X}(n, i)$$

$$= \bar{X}_n - \sum_{i=1}^{d} (\hat{p}_i(n) - p_i) \bar{X}(n, i)$$

$$= \bar{X}_n - \sum_{i=1}^{d} (\hat{p}_i(n) - p_i) \mu_i + o_p(n^{-1/2})$$

$$= \bar{X}_n(\boldsymbol{\lambda}^*) + o_p(n^{-1/2}),$$

as required. ∎

Thus, we may conclude that the post-stratified estimator is actually optimal from a control variates standpoint.

It is interesting to note that the control variates "correction" to $\bar{X}_n$ is just

$$\sum_{i=1}^{d} (\hat{p}_i(n) - p_i) \mu_i = \frac{1}{n} \sum_{j=1}^{n} (g^*(R_j) - \alpha),$$

where $R_j = \sum_{i=1}^{d} iI(\Gamma_j \in A_i)$ and $g^*(R_j) = E(X_j|R_j)$. So, the optimal control variates estimator is a sample mean of iid replicates of $X_1 - (E(X_1|R_1) - \alpha)$. Hence, post-stratification effectively involves using the control variate $E(X_1|R_1) - \alpha$, leading to a variance (per replication) of $\mathrm{var}(X_1 - E(X_1|R_1))$. Thus, post-stratification eliminates exactly the same variance component as does the control variate scheme described in Section 5.

8   Connection to Nonparametric Maximum Likelihood Estimation

We conclude this paper with a discussion of the connection between the method of control variates and the method of nonparametric maximum likelihood estimation (MLE). Nonparametric maximum likelihood arises naturally in the setting of "constrained Monte Carlo".

Monte Carlo computation, in the presence of constraints, has recently been investigated by [13]; see also [14]. The general statement of the problem involves the computation of $\alpha = EX$, where X is the first component of a simulatable random vector $\boldsymbol{X} = (X(0), X(1), \cdots, X(d))$. In other words, we wish to compute $\alpha = EX$, where $X \stackrel{\mathcal{D}}{=} X(0)$. Constrained Monte Carlo deals with the situation in which there exists a given set $B \subseteq \mathbb{R}^{d+1}$ for which it is known that $E\boldsymbol{X} \in B$. Computation, in the presence of such constraints, arises in many different problem settings; see [13] for details.

Pure equality constraints arise when it is known that $\boldsymbol{Y} \triangleq (X(1), \cdots, X(2))$ satisfies the constraint $E\boldsymbol{Y} = \boldsymbol{y}$ for some known $\boldsymbol{y}$. Thus, if we set $\boldsymbol{C} = \boldsymbol{Y} - \boldsymbol{y}$, we can view $\boldsymbol{C}$ as a d-dimensional vector of control variates having mean $\boldsymbol{0}$.

One approach to the problem of constrained Monte Carlo is nonparametric MLE. Specifically, suppose that we have generated n iid copies $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ of the random vector $\boldsymbol{X}$. Nonparametric MLE involves finding a probability distribution of the form

$$\sum_{i=1}^{n} p_i \delta_{\boldsymbol{X}_i}(\cdot)$$

that explains the simulated data set best, subject to the constraint that the probability distribution must satisfy the known constraints on $E\boldsymbol{X}$. (Here, $\delta_{\boldsymbol{X}_i}(\cdot)$ is a unit point mass on $\boldsymbol{X}_i$ and the collection $(p_i : 1 \le i \le n)$ is a probability mass function on $\{1, \cdots, n\}$.) Set $\boldsymbol{p} = (p_1, \cdots, p_n)$. The "highest likelihood" solution for the weights $\boldsymbol{p}$ involves maximizing the likelihood

$$L(\boldsymbol{p}) = \prod_{i=1}^{n} p_i \tag{7}$$

subject to $p_i \geq 0$, $1 \leq i \leq n$, and

$$\sum_{i=1}^{n} p_i = 1 \tag{8}$$

$$\sum_{i=1}^{n} p_i \boldsymbol{X}_i \in B. \tag{9}$$

Assuming that $\boldsymbol{p}_n^* = (p_{1n}^*, \cdots, p_{nn}^*)$ is the maximizer of the above optimization problem, the estimator for α is then

$$\begin{aligned}
\tilde{\alpha}(n) &= \int_{\mathbb{R}^{d+1}} x_0 \sum_{i=1}^{n} p_{in}^* \delta_{\boldsymbol{X}_i}(d\boldsymbol{x}) \\
&= \sum_{i=1}^{n} p_{in}^* X_i(0),
\end{aligned} \tag{10}$$

where $X_i(0)$ is the zero'th component of $\boldsymbol{X}_i$. As indicated above, there is a clear connection between constrained Monte Carlo and the method of control variates when the constraint takes the form $E\boldsymbol{Y} = \boldsymbol{y}$. The remainder of this section is devoted to exploring this connection in greater detail.

In the setting of such an equality constraint, the simulated data can be represented as n iid copies $(X_1, \boldsymbol{C}_1), \cdots, (X_n, \boldsymbol{C}_n)$ of a random vector $(X, \boldsymbol{C})$, in which it is known that $E\boldsymbol{C} = \boldsymbol{0}$. The optimization problem (7)-(10) then becomes:

$$\text{Maximize } L(\boldsymbol{p}) = \prod_{i=1}^{n} p_i \tag{11}$$

subject to $p_i \geq 0$, $1 \leq i \leq n$, and

$$\sum_{i=1}^{n} p_i = 1 \tag{12}$$

$$\sum_{i=1}^{n} p_i \boldsymbol{C}_i = \boldsymbol{0}. \tag{13}$$

If $\boldsymbol{p}^*$ is the maximizer of (11)-(13), the "constrained Monte Carlo" estimator for α then takes the form

$$\tilde{\alpha}(n) = \sum_{i=1}^{n} p_{in}^* X_i. \tag{14}$$

Of course, maximizing (11) is equivalent to maximizing the log-likelihood, namely

$$\mathcal{L}(\boldsymbol{p}) = \sum_{i=1}^{n} \log p_i.$$

Note that the feasible region described by (12), (13), and the non-negativity constraints is non-empty for n sufficiently large, since $\mathbf{0}$ must lie in the convex hull of the $\boldsymbol{C}_i$'s for n large enough (on account of the fact that $E\boldsymbol{C} = \mathbf{0}$). Furthermore, the feasible region is clearly convex. Because $\mathcal{L}$ is strictly concave, it follows that the above optimization problem can have at most one global maximizer, namely $\boldsymbol{p}_n^*$.

The standard means of solving such an optimization problem it to temporarily ignore the non-negativity constraints and to introduce Kuhn-Tucker multipliers $\gamma_n \in \mathbb{R}$ for (12) and $\boldsymbol{\nu}_n \in \mathbb{R}^d$ for (13). As usual, we choose to encode $\boldsymbol{p}_n^*$ and $\boldsymbol{\nu}_n$ as column vectors. At the maximizer $\boldsymbol{p}_n^*$, the multipliers should satisfy

$$(1/p_{in}^*)\boldsymbol{C}_i - \boldsymbol{\nu}_n^T \boldsymbol{C}_i - \gamma_n = 0 \tag{15}$$

for $1 \le i \le n$. Multiplying through (15) by p_{in}^* and summing over i, we get

$$n - \boldsymbol{\nu}_n^T \sum_{i=1}^n p_{in}^* \boldsymbol{C}_i - \gamma_n \sum_{i=1}^n p_{in}^* = 0.$$

Hence, $\gamma_n = n$ and (15) implies that

$$p_{in}^* = (n + \boldsymbol{\nu}_n^T \boldsymbol{C}_i)^{-1} \tag{16}$$

for $1 \le i \le n$. Multiplying through (16) by $\boldsymbol{C}_i$ and summing over i, we arrive at the equation

$$\sum_{i=1}^n \boldsymbol{C}_i (n + \boldsymbol{\nu}_n^T \boldsymbol{C}_i)^{-1} = \mathbf{0}.$$

In other words, the multiplier $\boldsymbol{\nu}_n$ that characterizes p_n^* via relation (16) is a root of the equation $f_n(\boldsymbol{\nu}_n) = \mathbf{0}$, where $f_n : \mathbb{R}^d \to \mathbb{R}^d$ is given by

$$f_n(\boldsymbol{\nu}) = \sum_{i=1}^n \boldsymbol{C}_i (n + \boldsymbol{\nu}_n^T \boldsymbol{C}_i)^{-1}.$$

Given the root $\boldsymbol{\nu}_n \in \mathbb{R}^d$, $\hat{\alpha}_n$ is then defined by

$$\hat{\alpha}(n) = \sum_{i=1}^n X_i (n + \boldsymbol{\nu}_n^T \boldsymbol{C}_i)^{-1}.$$

If the root $\boldsymbol{\nu}_n$ yields p_{in}^*'s that are all positive, then the relaxation obtained by ignoring the non-negativity constraints yields a solution to the fully constrained problem. The proof of Theorem 7 will confirm that the above approach is asymptotically valid (so that the p_{in}^*'s defining $\hat{\alpha}(n)$ can be obtained from the root $\boldsymbol{\nu}_n$ of $f_n(\boldsymbol{\nu}) = \mathbf{0}$). More importantly, Theorem 7 shows that the constrained Monte Carlo estimator $\hat{\alpha}(n)$ is asymptotically identical to the optimal control variates estimator $\bar{X}_n(\boldsymbol{\lambda}^*) = \bar{X}_n - \boldsymbol{\lambda}^{*T} \bar{\boldsymbol{C}}_n$. This result

extends, to $d \geq 2$, an identical conclusion reached by [13] in the setting of scalar control variates.

For a vector $\boldsymbol{x} \in \mathbb{R}^d$, let $\|\boldsymbol{x}\|_\infty \triangleq \max\{|x_i| : 1 \leq i \leq d\}$, where the x_i's are the components of $\boldsymbol{x}$.

Theorem 7. Suppose that $\boldsymbol{ECC}^T$ is non-singular and that $E|X|^2 < \infty$ and $E\|\boldsymbol{C}\|_\infty^4 < \infty$. Then,

$$\hat{\alpha}(n) = \bar{X}_n(\boldsymbol{\lambda}^*) + o_p(n^{-1/2})$$

as $n \to \infty$.

Proof. We will start by establishing that for n sufficiently large, the equation $f_n(\boldsymbol{\nu}) = \boldsymbol{0}$ has a root $\boldsymbol{\nu}_n$ for which all the quantities p_{in}^*, $1 \leq i \leq n$, defined by (16) are positive. It will then follow by convexity and from Theorem 4.38 of [1] that $\boldsymbol{p}_n^* = (p_{1n}^*, \cdots, p_{nn}^*)^T$ is the global solution of the fully constrained problem.

Note that $f_n(\boldsymbol{\nu}) = \boldsymbol{0}$ if and only if $T_n(\boldsymbol{\nu}) = \boldsymbol{\nu}$, where $T_n(\boldsymbol{\nu}) = \boldsymbol{\nu} + n(\boldsymbol{ECC}^T)^{-1} f_n(\boldsymbol{\nu})$. We shall shortly see that $\tilde{\boldsymbol{\nu}}_n = (\boldsymbol{ECC}^T)^{-1} \sum_{i=1}^n \boldsymbol{C}_i$ is an approximate "fixed point" of T_n, so that $\boldsymbol{\nu}_n$ is close to $\tilde{\boldsymbol{\nu}}_n$. In fact, for $0 < \delta < 1/2$, we will show that $\boldsymbol{\nu}_n$ is, for n large, in the set $B_n \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \tilde{\boldsymbol{\nu}}_n\|_\infty \leq n^\delta\}$. (Given that the Law of the Iterated Logarithm ensures that $\tilde{\boldsymbol{\nu}}_n = O((n \log \log n)^{1/2})$ a.s. as $n \to \infty$ and $\delta < 1/2$, $\boldsymbol{\nu}_n \in B_n$ is therefore indeed a statement of "closeness".)

Observe that for $|x| < 1/2$, $|(1+x)^{-1} - (1-x)| \leq 2x^2$. Also, the moment condition $E\|\boldsymbol{C}\|_\infty^4 < \infty$, in conjunction with the Borel-Cantelli lemma, implies that $\|\boldsymbol{C}_n\|_\infty = O(n^{1/4})$ a.s., from which it follows that $\sup\{|\boldsymbol{\nu}^T \boldsymbol{C}_i|/n : \boldsymbol{\nu} \in B_n\} \to 0$ a.s. as $n \to \infty$. Thus, for n sufficiently large,

$$\sup_{\boldsymbol{\nu} \in B_n} |(1 + \boldsymbol{\nu}\boldsymbol{C}_i^T/n)^{-1} - (1 - \boldsymbol{\nu}\boldsymbol{C}_i^T/n)| \leq 3.(\|\tilde{\boldsymbol{\nu}}_n\|_\infty/n)^2 \|\boldsymbol{C}_i\|_\infty^2. \qquad (17)$$

So,

$$T_n(\boldsymbol{\nu}) = \boldsymbol{\nu} + (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \sum_{i=1}^{n} \boldsymbol{C}_i (1 + \boldsymbol{\nu}^T \boldsymbol{C}_i/n)^{-1}$$

$$= \boldsymbol{\nu} + (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \left[\sum_{i=1}^{n} \boldsymbol{C}_i - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_i \boldsymbol{C}_i^T \boldsymbol{\nu} \right]$$

$$+ O\left(\left(\frac{\|\tilde{\boldsymbol{\nu}}_n\|_\infty}{n} \right)^2 \sum_{i=1}^{n} \frac{\|\boldsymbol{C}_i\|_\infty^3}{n} \right) \qquad (18)$$

$$= \boldsymbol{\nu} - \tilde{\boldsymbol{\nu}}_n - (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_i \boldsymbol{C}_i^T \boldsymbol{\nu} + O\left(\frac{\log\log n}{n} \right) \text{ a.s.}$$

$$= \tilde{\boldsymbol{\nu}}_n + O(\log\log n) \text{ a.s.}$$

uniformly in $\boldsymbol{\nu} \in B_n$. Consequently, T_n maps B_n into itself for n sufficiently large.

Furthermore, for $\boldsymbol{x}, \boldsymbol{y} \in B_n$,

$$\|T_n(\boldsymbol{x}) - T_n(\boldsymbol{y})\|_\infty$$

$$= \|(\boldsymbol{x} - \boldsymbol{y}) + (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_i [(1 + \boldsymbol{x}\boldsymbol{C}_i/n)^{-1} - (1 + \boldsymbol{y}\boldsymbol{C}_i/n)^{-1}]\|_\infty$$

$$\leq \|I - (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_i \boldsymbol{C}_i^T (1 + \boldsymbol{x}\boldsymbol{C}_i/n)^{-1} (1 + \boldsymbol{y}\boldsymbol{C}_i/n)\|_\infty \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$$

$$= \|I - (E\boldsymbol{C}\boldsymbol{C}^T)^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_i \boldsymbol{C}_i^T (1 + O(n^{-1/8}))^2\|_\infty \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$$

$$= O\left(\left(\frac{\log\log n}{n} \right)^{1/2} \right) \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$$

uniformly in $\boldsymbol{x}, \boldsymbol{y} \in B_n$. It follows that for n sufficiently large, T_n is a strict contraction on the closed ball B_n. The Banach Fixed Point Theorem (see, for example, [9]) therefore guarantees the existence of a unique fixed point $\boldsymbol{\nu}_n \in B_n$ for the mapping T_n. Furthermore, because $\boldsymbol{\nu}_n \in B_n$, it is easily seen that the resulting p_{in}^*'s are uniformly positive for n large. As argued earlier, we may therefore conclude that $\boldsymbol{\nu}_n$ is the multiplier associated with the desired global maximizer to the fully constrained problem.

Substituting $\boldsymbol{\nu} = \boldsymbol{\nu}_n$ in (18), we find that

$$\boldsymbol{\nu}_n = \tilde{\boldsymbol{\nu}}_n + O(\log\log n) \text{ a.s.}$$

So, (17) shows that

$$\tilde{\boldsymbol{\nu}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i (1 + \boldsymbol{\nu}_n^T \boldsymbol{C}_i / n)^{-1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i (1 - \boldsymbol{\nu}_n^T \boldsymbol{C}_i / n) + O\left(\left(\frac{\|\tilde{\boldsymbol{\nu}}_n\|_\infty}{n}\right)^2 \frac{1}{n}\sum_{i=1}^{n} |X_i| \|\boldsymbol{C}_i\|^2 \infty\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i (1 - \tilde{\boldsymbol{\nu}}_n^T \boldsymbol{C}_i / n) + O\left(\frac{\log\log n}{n}\right) \ \text{a.s.}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i(\boldsymbol{\lambda}^*) + O\left(\frac{\log\log n}{n}\right) \ \text{a.s.},$$

proving the theorem. ∎

References

1. Avriel, M. (1976) Nonlinear Programming: Analysis and Methods. Prentice-Hall, Englewood Cliffs.
2. Billingsley, P. (1968) Convergence of Probability Measures. John Wiley, New York.
3. Billingsley, P. (1995) Probability and Measure, 3rd edn. John Wiley, New York.
4. Bratley, P., Fox, B.L., and L.E. Schrage (1987) A Guide to Simulation. Springer-Verlag, New York.
5. Davis, P.J., and P. Rabinowitz (1984) Methods of Numerical Integration. Academic Press, New York.
6. Evans, M., and T. Swartz (2000) Approximating Integrals Via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford.
7. Fishman, G.S., and B.D. Huang (1983) Antithetic Variates Revisited. Comms. of the ACM 26:11:964-971.
8. Glynn, P.W. , and W. Whitt (1992) The Asymptotic Efficiency of Simulation Estimators . Op. Res. 40:505-520.
9. Kreyszig, E. (1978) Introductory Functional Analysis with Applications. John Wiley, New York.
10. Loh, W.W. (1995) On the Method of Control Variates. PhD Thesis, Stanford University, Stanford.
11. Ross, S.M. (1990) A Course in Simulation. Macmillan Publishing Company, New York.
12. Rubinstein, R.Y. (1981) Simulation and the Monte Carlo Method. John Wiley, New York.
13. Szechtman, R., and P.W. Glynn (2001) Constrained Monte Carlo and the Method of Control Variates. Submitted to Win. Simulation Conf.
14. Szechtman, R. (2001) Efficient Monte Carlo Simulation in the Presence of Constraints. PhD Thesis, Stanford University, Stanford.