Chapter 10: A Hilbert Space Approach To Variance Reduction

Roberto Szechtman

Department of Operations Research, Naval Postgraduate School

Abstract

In this chapter we explain variance reduction techniques from the Hilbert space standpoint, in the terminating simulation context. We use projection ideas to explain how variance is reduced, and to link different variance reduction techniques. Our focus is on the methods of control variates, conditional Monte Carlo, weighted Monte Carlo, stratification, and Latin hypercube sampling.

1 Introduction

The goal of this chapter is to describe variance reduction techniques from a Hilbert space perspective in the terminating simulation setting, with the focal point lying on the method of control variates. Several variance reduction techniques have an intuitive geometric interpretation in the Hilbert space setting, and it is often possible to obtain rather deep probabilistic results with relatively little effort by framing the relevant mathematical objects in an appropriate Hilbert space. The procedure employed to reach most results in this context consists of three stages:

(1) Find a pertinent space endowed with an inner product.
(2) Apply Hilbert space results towards a desired conclusion.
(3) Translate the conclusions back into probabilistic language.

The key geometric idea used in Stage 2 is that of "projection": Given an element y in a Hilbert space H and a subset M of the Hilbert space, it holds under mild conditions that there exists a unique element in M that is closest to y. The characterization of this element depends on the space H determined

Email address: rszechtm@nps.edu (Roberto Szechtman).

by Stage 1, on y, and on M; in essence it is found by dropping a perpendicular from y to M.

The Hilbert space explanation of control variates, and to a somewhat lesser extent that of conditional Monte Carlo, is closely related to that of other variance reduction techniques; in this chapter we bridge these connections whenever they arise. From the projection perspective, it is often possible to link variance reduction techniques for which the relevant Hilbert space H and element $y \in H$ are the same. The articulation is done by judiciously choosing the subset M for each particular technique.

We do not attempt to provide a comprehensive survey of control variates or of the other techniques covered in this chapter. As to control variates, several publications furnish a broader picture; see, for example, Lavenberg and Welch (1981), Lavenberg et al. (1982), Wilson (1984), Rubinstein and Marcus (1985), Venkatraman and Wilson (1986), Law and Kelton (2000), Nelson (1990), Loh (1995), and Glasserman (2004). For additional material on other variance reduction techniques examined here, refer to the items in the References section and to references therein.

This chapter is organized as follows: Section 2 is an overview of control variates. In Section 3 we review Hilbert space theory and present several examples that serve as a foundation for the rest of the chapter. Section 4 is about control variates in Hilbert space. The focus of Section 5 is on the method of conditional Monte Carlo, and on combinations of conditional Monte Carlo with control variates. Section 6 describes how control variates and conditional Monte Carlo can reduce variance cooperatively. The subject of Section 7 is the method of weighted Monte Carlo. In Sections 8 and 9 we describe stratification techniques and Latin hypercube sampling, respectively. The last section presents an application of the techniques we investigate. As stated above, the focus of this chapter is in interpreting and connecting various variance reduction techniques in a Hilbert space framework.

2 Problem Formulation and Basic Results

We study efficiency improvement techniques for the computation of an unknown scalar parameter α that can be represented as $\alpha = EY$, where Y is a random variable called the response variable, in the terminating simulation setting. Given n independent and identically distributed (i.i.d.) replicates $Y_1, \ldots, Y_n$ of Y produced by the simulation experiment, the standard estimator for α is the sample mean

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^{n} Y_k.$$

The method of control variates (CVs) arises when the simulationist has available a random column vector $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$, called the control, such that $\mathbf{X}$ is jointly distributed with Y, $E\mathbf{X} = \boldsymbol{\mu}_{\mathbf{x}}$ is known, and it is possible to obtain i.i.d. replicates $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ of $(Y, \mathbf{X})$ as a simulation output. Under these conditions, the CV estimator is defined by

$$\hat{Y}_{CV}(\boldsymbol{\lambda}) = \bar{Y} - \boldsymbol{\lambda}^T(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{x}}), \tag{1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$ is the vector of control variates coefficients; $\cdot^T$ denotes transpose, vectors are defined as columns, and vectors and matrices are written in bold.

The following holds throughout this chapter:

Assumption 1 *$E(Y^2 + \sum_{i=1}^d X_i^2) < \infty$ and the covariance of $(Y, \mathbf{X})$, defined by*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\Sigma}_{y\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{x}y} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{pmatrix},$$

is non-singular.

The first part of Assumption 1 is satisfied in most settings of practical interest. Furthermore, when $\boldsymbol{\Sigma}$ is singular it is often possible to make it non-singular by reducing the number of controls $\mathbf{X}$; see the last paragraph of Example 5.

Naturally, $\boldsymbol{\lambda}$ is chosen to minimize $\operatorname{Var} \hat{Y}_{CV}(\boldsymbol{\lambda})$, which is the same as

$$\text{minimizing } \sigma_y^2 - 2\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{x}y} + \boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\lambda}. \tag{2}$$

The first and second-order optimality conditions for this problem imply that there exists a unique optimal solution given by

$$\boldsymbol{\lambda}^* = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}y}. \tag{3}$$

With this choice of $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ the CV estimator variance is

$$\operatorname{Var} \hat{Y}_{CV}(\boldsymbol{\lambda}^*) = \operatorname{Var} \bar{Y}(1 - R_{y\mathbf{x}}^2), \tag{4}$$

where

$$R_{y\mathbf{x}}^2 = \frac{\boldsymbol{\Sigma}_{y\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}y}}{\sigma_y^2}$$

is the square of the multiple correlation coefficient between Y and $\mathbf{X}$. CVs reduce variance because $0 \le R_{y\mathbf{x}}^2 \le 1$ implies $\operatorname{Var} \hat{Y}_{CV} \le \operatorname{Var} \bar{Y}$ in (4). The central limit theorem (CLT) for $\hat{Y}_{CV}$ asserts that, under Assumption 1,

$$n^{1/2}(\hat{Y}_{CV}(\boldsymbol{\lambda}^*) - \alpha) \Rightarrow N(0, \sigma_{CV}^2),$$

where $\sigma^2_{CV} = \sigma^2_y(1-R^2_{y\mathbf{x}})$, $\Rightarrow$ denotes convergence in distribution, and $N(0,\sigma^2)$ is a zero-mean Normal random variable with variance σ^2.

In general, however, the covariance structure of the random vector $(Y,\mathbf{X})$ may not be fully known prior to the simulation. This difficulty can be overcome by using the available samples to estimate the unknown components of $\boldsymbol{\Sigma}$, which can then be used to estimate $\boldsymbol{\lambda}^*$. Let $\boldsymbol{\lambda}_n$ be an estimator of $\boldsymbol{\lambda}^*$ and suppose that $\boldsymbol{\lambda}_n \Rightarrow \boldsymbol{\lambda}^*$ as $n \to \infty$. Then, under Assumption 1,

$$n^{1/2}(\hat{Y}_{CV}(\boldsymbol{\lambda}_n) - \alpha) \Rightarrow N(0, \sigma^2_{CV}), \tag{5}$$

as $n \to \infty$; see Glynn and Szechtman (2002) for details. Equation (5) means that estimating $\boldsymbol{\lambda}^*$ causes no loss of efficiency as $n \to \infty$, if $\boldsymbol{\lambda}_n \Rightarrow \boldsymbol{\lambda}$.

Thus far we have only considered linear control variates of the form $\bar{Y} - \boldsymbol{\lambda}^T(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{x}})$. In some applications, however, the relationship between the response and the CVs is non-linear, examples of which are: $\bar{Y}\exp(\boldsymbol{\lambda}^T(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{x}}))$, $\bar{Y}\bar{X}/\mu_x$, and $\bar{Y}^{\mu_x/\bar{X}}$. In order to have a general representation of CVs we introduce a function $f : \mathbb{R}^{d+1} \to \mathbb{R}$ that is continuous at $(y, \boldsymbol{\mu}_{\mathbf{x}})$ with $f(y, \boldsymbol{\mu}_{\mathbf{x}}) = y$. This property ensures that $f(\bar{Y}, \bar{\mathbf{X}}) \to \alpha$ a.s. if $(\bar{Y}, \bar{\mathbf{X}}) \to (\alpha, \boldsymbol{\mu}_{\mathbf{x}})$ a.s., so we only consider such functions.

The limiting behavior of $f(\bar{Y}, \bar{\mathbf{X}})$ is characterized in Glynn and Whitt (1989, Theorem 9) under the assumption that the i.i.d. sequence $\{(Y_n, \mathbf{X}_n) : n \geq 1\}$ satisfies the CLT $\sqrt{n}((\bar{Y}, \bar{\mathbf{X}}) - (\alpha, \boldsymbol{\mu}_{\mathbf{x}})) \Rightarrow N(\mathbf{0}, \boldsymbol{\Sigma})$, and that f is continuously differentiable in a neighborhood of $(\alpha, \boldsymbol{\mu}_{\mathbf{x}})$ with first partial derivatives not all zero at $(\alpha, \boldsymbol{\mu}_{\mathbf{x}})$. Then

$$\sqrt{n}(f(\bar{Y}, \bar{\mathbf{X}}) - \alpha) \Rightarrow N(0, \sigma^2_f), \tag{6}$$

as $n \to \infty$, where σ^2_f is given by

$$\sigma^2_f = \sigma^2_y + 2\boldsymbol{\nabla}_{\mathbf{x}} f(\alpha, \boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Sigma}_{\mathbf{x}y} + \boldsymbol{\nabla}_{\mathbf{x}} f(\alpha, \boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Sigma}_{\mathbf{xx}} \boldsymbol{\nabla}_{\mathbf{x}} f(\alpha, \boldsymbol{\mu}_{\mathbf{x}}), \tag{7}$$

and $\boldsymbol{\nabla}_{\mathbf{x}} f(y, \mathbf{x}) \in \mathbb{R}^d$ is the vector with ith component $\partial f / \partial x_i(y, \mathbf{x})$.

The asymptotic variance σ^2_f is minimized, according to Equation (2) with $\boldsymbol{\nabla}_{\mathbf{x}} f(\alpha, \boldsymbol{\mu}_{\mathbf{x}})$ in lieu of $\boldsymbol{\lambda}$, by selecting f^* such that $\boldsymbol{\nabla}_{\mathbf{x}} f^*(\alpha, \boldsymbol{\mu}_{\mathbf{x}}) = -\boldsymbol{\Sigma}^{-1}_{\mathbf{xx}} \boldsymbol{\Sigma}_{\mathbf{x}y}$ in (7); that is, $\sigma^2_{f^*} = \sigma^2_y(1 - R^2_{y\mathbf{x}})$. In other words, non-linear CVs have at best the same asymptotic efficiency as $\hat{Y}_{CV}(\boldsymbol{\lambda}^*)$. Notice, however, that for small sample sizes it could happen that non-linear CVs achieve more (or less) variance reduction than linear CVs.

The argument commonly used to prove this type of result is known as the Delta method; see Chapter 2 or, for a more detailed treatment, Serfling (1980, p. 122). The reason why $\sqrt{n}(f(\bar{Y}, \bar{\mathbf{X}}) - \alpha)$ converges in distribution to a normal

random variable is that f is linear in a neighborhood of $(\alpha, \boldsymbol{\mu}_{\mathbf{x}})$ because f is differentiable there, and a linear function of a normal random variable is again normal.

To conclude this section, for simplicity let the dimension $d = 1$ and suppose that only an approximation of μ_x, say $\gamma = \mu_x + \epsilon$ for some scalar ϵ, is known. This is the setting of biased control variates (BCV). The BCV estimator is given by

$$\hat{Y}_{BCV}(\lambda) = \bar{Y} - \lambda(\bar{X} - \gamma).$$

The bias of $\hat{Y}_{BCV}(\lambda)$ is $\lambda\epsilon$, and the mean-squared error is

$$E(\hat{Y}_{BCV}(\lambda) - \alpha)^2 = \operatorname{Var} \bar{Y} + \lambda^2 E(\bar{X} - \gamma)^2 - 2\lambda \operatorname{Cov}(\bar{Y}, \bar{X}).$$

Mean-squared error is minimized by

$$\lambda_n = \operatorname{Cov}(\bar{Y}, \bar{X}) / E(\bar{X} - \gamma)^2, \tag{8}$$

and

$$E(\hat{Y}_{BCV}(\lambda_n) - \alpha)^2 = \operatorname{Var} \bar{Y} \left(1 - \frac{\operatorname{Cov}(\bar{Y}, \bar{X})^2}{\operatorname{Var} \bar{Y} E(\bar{X} - \gamma)^2}\right),$$

which is (4) when $\epsilon = 0$; see Schmeiser et al. (2001) for more details on BCVs.

3 Hilbert Spaces

We present basic ideas about Hilbert spaces, mainly drawn from Kreyszig (1978), Bollobas (1990), Zimmer (1990), Williams (1991), and Billingsley (1995). We encourage the reader to consult those references for proofs, and for additional material. We illustrate the concepts with a series of examples that serve as foundational material for the rest of the chapter.

An inner product space is a vector space V with an inner product $\langle x, y \rangle$ defined on it. An inner product on V is a mapping of $V \times V$ into $\mathbb{R}$ such that for all vectors x, y, z and scalars α, β we have

(i) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$.
(ii) $\langle x, x \rangle \geq 0$, with equality if and only if $x = 0$.
(iii) $\langle x, y \rangle = \langle y, x \rangle$.

An inner product defines a norm on X given by

$$\|x\| = \sqrt{\langle x, x \rangle}. \tag{9}$$

A Hilbert space H is a complete inner product space, complete meaning that every Cauchy sequence in H has a limit in H.

The next three examples present the Hilbert spaces that will be employed throughout this chapter.

Example 1 *Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and*

$$\mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P}) = \left\{ Y \in (\Omega, \mathcal{F}, \mathcal{P}) : EY^2 = \int_\Omega Y(\omega)^2 d\mathcal{P}(\omega) < \infty \right\}$$

the space of square-integrable random variables defined on $(\Omega, \mathcal{F}, \mathcal{P})$. For $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, the inner product is defined by

$$\langle X, Y \rangle = E(XY) = \int_\Omega X(\omega) Y(\omega) d\mathcal{P}(\omega), \tag{10}$$

and the norm is given by

$$\|Y\| = \sqrt{EY^2} = \left(\int_\Omega Y(\omega)^2 d\mathcal{P}(\omega) \right)^{1/2}, \tag{11}$$

by (9). It can be easily verified that the inner product defined by (10) has properties (i), (ii), and (iii). The space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ is complete under this norm (Billingsley, 1995, p. 243). Note that

$$\operatorname{Var} Y = \|Y - EY\|^2. \tag{12}$$

Example 2 *The space $\mathbb{R}^n$ is the set of vectors $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$, and can be made into a Hilbert space by defining the inner product for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ as*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^n x_j y_j. \tag{13}$$

The norm induced by (13) is

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \left(\sum_{j=1}^n x_j^2 \right)^{1/2}.$$

The space $\mathbb{R}^n$ is complete under this norm (Bollobas, 1990, p. 133).

Example 3 *Consider independent random variables X_i with distribution function $F_i(x_i)$, $1 \leq i \leq d$, and define $F(\mathbf{x}) = \prod_{i=1}^d F_i(x_i)$, for $\mathbf{x} = (x_1, \ldots, x_d)$. Write $\mathbf{X} = (X_1, \ldots, X_d)$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a Borel-measurable function in $\mathcal{L}^2(dF)$, the space of square integrable functions with respect to F. This space can be made into a Hilbert space by defining the inner product:*

$$\langle f, g \rangle = \int f(\mathbf{x}) g(\mathbf{x}) dF(\mathbf{x}), \tag{14}$$

for any $f, g \in \mathcal{L}^2(dF)$. The norm induced by (14) is

$$\|f\| = \left(\int f(\mathbf{x})^2 dF(\mathbf{x}) \right)^{1/2},$$

and $\mathcal{L}^2(dF)$ is complete under this norm (Billingsley, 1995, p. 243).

The notion of orthogonality among elements lies at the heart of Hilbert space theory, and extends the notion of perpendicularity in Euclidean space. Two elements x, y are orthogonal if

$$\langle x, y \rangle = 0.$$

From here, there is just one step to the Pythagorean theorem:

Result 1 *(Pythagorean Theorem) Kreyszig (1978). If $x_1, \ldots, x_n$ are pairwise orthogonal vectors of an inner product space V then*

$$\left\| \sum_{i=1}^{n} x_i \right\|^2 = \sum_{i=1}^{n} \|x_i\|^2. \tag{15}$$

Let us write

$$x^{\perp} = \{y \in V : \langle x, y \rangle = 0\}$$

for the set of orthogonal vectors to $x \in V$, and

$$S^{\perp} = \{y \in V : \langle x, y \rangle = 0, \forall x \in S\}$$

for $S \subset V$. Finally, a set $(x_1, \ldots, x_n)$ of elements in V is orthogonal if all its elements are pairwise orthogonal.

We often work with a subspace S of a Hilbert space H defined on X, by which we mean a vector subspace of X with the inner product restricted to $S \times S$. It is important to know when S is complete, and therefore a Hilbert space. It is easy to prove that S is complete if and only if S is closed in H.

Consider a non-empty subset M of an inner product space V. Central to the concepts discussed in this chapter is to know when, given a point y in V, there exists a unique point $x \in M$ that minimizes the distance from y to M, where the distance from y to M is defined to be $d(y, M) = \inf_{v \in M} \|y - v\|$. The following result provides an answer to this problem.

Result 2 *(Projection Theorem) Kreyszig (1978). Let M be a complete subspace of an inner product space V, and let $y \in V$ be fixed. Then there exists a unique $x = x(y) \in M$ such that*

$$d(y, M) = \|y - x\|.$$

Furthermore, every $y \in V$ has a unique representation of the form

$$y = x + z,$$

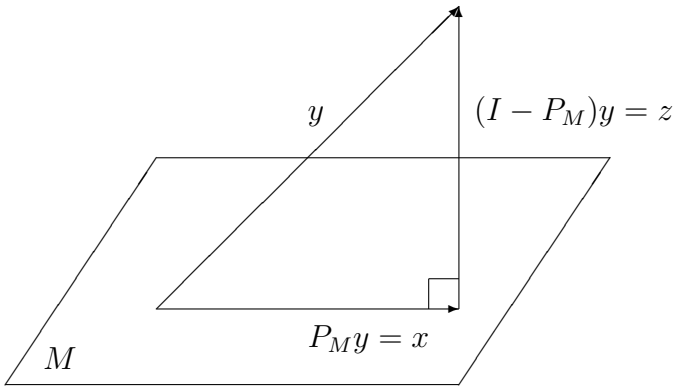

Fig. 1. Orthogonal Projection

where $x \in M$ *and* $z \in M^{\perp}$. *Then*

$$\langle x, y - x \rangle = \langle x, z \rangle = 0. \tag{16}$$

The second part of Result 2 implies that if M is a complete subspace of an inner product space V, then there exists a function $P_M : V \to M$ defined by $P_M y = x$. We call P_M the orthogonal projection of V onto M, see Figure 1.

Among the properties that the projection functional enjoys are:

a) $P_M x = x$, for all $x \in M$.
b) $P_M z = 0$, for all $z \in M^{\perp}$.
c) $\|I - P_M\| \leq 1$.

Applying Result 2 to Examples 1, 2, and 3 leads to several variance reduction techniques in the Hilbert space setting. The following example forms the basis for the method of conditional Monte Carlo.

Example 4 *In the setting of Example 1, consider a sub-*σ*-algebra* $\mathcal{G}$ *of* $\mathcal{F}$. *The set of square integrable random variables defined on* $\mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ *is a complete subspace of* $\mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$. *Therefore, for* $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ *fixed, there exists an element* $Z \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ *that is the closest point to* Y *in* $\mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ *and for which*

$$\langle Y - Z, W \rangle = 0, \tag{17}$$

for $W \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathcal{P})$ *arbitrary. Choosing* $W = I_B$, $B \in \mathcal{G}$, *Equation (17) shows that* Z *is the conditional expectation of* Y *given* $\mathcal{G}$, $Z = E(Y|\mathcal{G})$. *We also can write* $P_{\mathcal{G}} Y = E(Y|\mathcal{G})$; *see Williams (1991) for more details.*

Observe that Equation (17) and the Pythagorean theorem imply that

$$\|Y\|^2 = \|Y - Z\|^2 + \|Z\|^2. \tag{18}$$

Using Equation (12), centering Y so that $EY = 0$, we have that Equation (18) is the variance decomposition formula (Law and Kelton, 2000):

$$\operatorname{Var} Y = E \operatorname{Var}(Y|\mathcal{G}) + \operatorname{Var} E(Y|\mathcal{G}), \tag{19}$$

where $\operatorname{Var}(Y|\mathcal{G}) = E(Y^2|\mathcal{G}) - (E(Y|\mathcal{G}))^2$.

We continue with an example with a view towards control variates.

Example 5 *For elements $X_1, \ldots, X_d \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ with zero mean (otherwise re-define $X_i := X_i - EX_i$), define $M = \{Z \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P}) : Z = \sum_{i=1}^d \beta_i X_i,$ for all $\beta_i \in \mathbb{R}, i = 1, \ldots, d\}$. For $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, Result 2 then guarantees the existence of constants $\beta_1^* = \beta_1^*(Y), \ldots, \beta_d^* = \beta_d^*(Y)$ such that,*

$$P_M(Y - EY) = \sum_{i=1}^d \beta_i^* X_i, \text{ and } (I - P_M)(Y - EY) = (Y - EY) - \sum_{i=1}^d \beta_i^* X_i. \tag{20}$$

If $Y - EY \in M$, using property a) of the projection operator we obtain $(I - P_M)(Y - EY) = 0$, so that

$$\operatorname{Var}\left(Y - \sum_{i=1}^d \beta_i^* X_i\right) = 0. \tag{21}$$

If the elements $X_1, \ldots, X_d$ form an orthogonal set, applying Equation (15) we have

$$\begin{aligned} \|(I - P_M)(Y - EY)\|^2 &= \|Y - EY\|^2 - \|P_M(Y - EY)\|^2 \\ &= \|Y - EY\|^2 - \sum_{i=1}^d \beta_i^{*2} \|X_i\|^2, \end{aligned}$$

so that

$$\operatorname{Var}\left(Y - \sum_{i=1}^d \beta_i^* X_i\right) = \operatorname{Var} Y - \sum_{i=1}^d \beta_i^{*2} \operatorname{Var} X_i. \tag{22}$$

When the elements $X_1, \ldots, X_d$ are not mutually orthogonal but linearly independent, the Gram-Schmidt process (Billingsley, 1995, p. 249) yields an orthogonal set with the same span as $X_1, \ldots, X_d$. In case the X_i's are linearly dependent, at least one X_i can be expressed as a linear combination of the others, and eliminated from the set $X_1, \ldots, X_d$. By noticing that $\operatorname{Cov}(X_i, X_j) = \langle X_i, X_j \rangle$ we gather that $X_1, \ldots, X_d$ are linearly independent if and only if their covariance matrix is non-singular, and $X_1, \ldots, X_d$ are mutually orthogonal if and only if their covariance matrix has all its entries equal to zero except for positive numbers on the diagonal.

We can extend Example 5 to the setting of biased control variates.

Example 6 *Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, and $M = \{Z \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P}) : Z = \beta(X - \gamma), \forall \beta \in \mathbb{R}\}$, $\gamma \neq EX$. Fix an element $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ and let $\alpha = EY$. Project $Y - \alpha$ on M:*

$$P_M(Y - \alpha) = \beta^*(X - \gamma), \text{ and } (I - P_M)(Y - \alpha) = Y - \alpha - \beta^*(X - \gamma),$$

for some $\beta^* = \beta^*(Y) \in \mathbb{R}$. As in the last example we have

$$\|(I - P_M)(Y - \alpha)\|^2 = \|Y - \alpha\|^2 - \|P_M(Y - \alpha)\|^2$$

and, since $\langle (I - P_M)(Y - \alpha), X - \gamma \rangle = 0$, it follows that

$$\beta^* = \frac{\langle Y - \alpha, X - \gamma \rangle}{\|X - \gamma\|^2}.$$

The Pythagorean theorem applied to the denominator in the last equation yields

$$\|X - \gamma\|^2 = \|X - EX\|^2 + \|EX - \gamma\|^2, \tag{23}$$

which is known as the bias-variance decomposition formula.

The following example is geared to the method of control variates when the optimal control coefficient is estimated from the sample data.

Example 7 *Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be elements of $\mathbb{R}^n$ (cf. Example 2), and define $S = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z} = \beta\mathbf{x}, \forall \beta \in \mathbb{R}\}$. By Result 2 we have*

$$P_S\mathbf{y} = \beta_n\mathbf{x}, \text{ and } (I - P_S)\mathbf{y} = \mathbf{y} - \beta_n\mathbf{x},$$

for some $\beta_n = \beta_n(\mathbf{y})$. Because $\langle \mathbf{y} - \beta_n\mathbf{x}, \mathbf{x} \rangle = 0$, for $\|\mathbf{x}\| > 0$,

$$\beta_n = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|^2}.$$

We now set the stage for the method of Latin hypercube sampling.

Example 8 *Building on Example 3, let $M = \{h \in \mathcal{L}^2(dF) : h(\mathbf{x}) = \sum_{i=1}^d h_i(x_i)\}$ be the subspace of $\mathcal{L}^2(dF)$ spanned by the linear combinations of univariate functions $h_1, \ldots, h_d$. Because M is complete, appealing to Result 2 establishes the existence of an element $h^* = h^*(f) \in M$, $h^*(\mathbf{x}) = \sum_{i=1}^d h_i^*(x_i)$, such that $\|f - h^*\| = \inf_{h \in M} \|f - h\|$ for $f \in \mathcal{L}^2(dF)$ fixed, and of a projection operator $P_M : P_M f = h^*$. Similarly, for $M_i = \{h \in \mathcal{L}^2(dF) : h(\mathbf{x}) = h_i(x_i)\}$, $1 \leq i \leq d$, there exists $g_i^* = g_i^*(f) \in M_i$: $\|f - g_i^*\| = \inf_{h \in M_i} \|f - h\|$ and a projection $P_i : P_i f = g_i^*$. To complete the picture, define the subspace $M_0 = \{\beta \in \mathbb{R} : |\beta| < \infty\}$ which induces the projection $P_0 : P_0 f = g_0^*$, for $g_0^* : \|f - g_0^*\| = \inf_{\beta \in M_0} \|f - \beta\|$. We now have:*

- *Let* $\mathcal{F}_i = \sigma(\{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R} : B \in \mathcal{B}\})$*, where* $\mathcal{B}$ *is the Borel* $\sigma-$*field, and let* $\mathcal{F}_0$ *be the trivial* σ*-algebra* $\{\emptyset, \mathbb{R}\}$*. For each* P_i*, we know that* $\langle f - P_i f, h_\rangle = 0$ *for any* $h \in M_i$*; choosing* $h(\mathbf{x}) = h_i(x_i) = I_B(x_i)$*,* $B \in \mathcal{B}$*, shows that* $P_i f = E(f(\mathbf{X})|\mathcal{F}_i)$ *for* $1 \leq i \leq d$*, and* $P_0 f = E(f(\mathbf{X})|\mathcal{F}_0) = Ef(\mathbf{X})$*.*
- *Suppose* $P_0 f = 0$*. Then* $(I - P_M)f \in M_i^\perp$ *implies* $g_i^* = P_i f = P_i P_M f = h_i^*$*, which results in*

$$P_M = \sum_{i=1}^{d} P_i, \text{ and } P_M f = \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_i). \tag{24}$$

For general $P_0 f \neq 0$*, (24) becomes*

$$P_M = P_0 + \sum_{i=1}^{d}(P_i - P_0), \text{ and } P_M f = Ef(\mathbf{X}) + \sum_{i=1}^{d}(E(f(\mathbf{X})|\mathcal{F}_i) - Ef(\mathbf{X})). \tag{25}$$

The next result, a variant of Result 2, will be useful when we consider the method of weighted Monte Carlo.

Result 3 *Kreyszig (1978) Suppose* $M \neq \emptyset$ *is a closed convex subset of a Hilbert space* H*. Then for* $x \in H$ *fixed,* $x_1 = P_M(x)$ *is the (unique) closest point in* M *to* x *if and only if*

$$\langle x - x_1, y - x_1\rangle \leq 0, \forall y \in M. \tag{26}$$

Later in the chapter we will deal with sequences of projections, say (P_n), defined on a Hilbert space H that are monotone increasing in that

$$\|P_i x\| \leq \|P_{i+1} x\|, \text{ for } i = 1, 2, \ldots,$$

and $x \in H$ arbitrary. Using the completeness of H it can be shown that (P_n) converges in the following sense:

Result 4 *Kreyszig (1978) Let* (P_n) *be a monotone increasing sequence of projections* P_n *defined on a Hilbert space* H*. Then, for any* $x \in H$*,*

$$\|P_n x - Px\| \to 0,$$

and the limit operator P *is a projection on* H*.*

An immediate application of this result is the following example.

Example 9 *Suppose that* $(\mathcal{F}_n)$ *is an increasing sequence* $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_\infty$ *of* σ*-algebras such that* $\mathcal{F}_\infty = \sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$*. Then associated with every* $\mathcal{L}^2(\Omega, \mathcal{F}_n, \mathcal{P})$ *there exists a projection* $P_n : \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P}) \to \mathcal{L}^2(\Omega, \mathcal{F}_n, \mathcal{P})$*, and*

the sequence of projections (P_n) is monotone increasing. Let P_∞ be the projection that results from applying Result 4: $\|P_n W - P_\infty W\| \to 0$ for any $W \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{F}_\infty \subseteq \mathcal{F}$. Because $(I - P_\infty)W \in \mathcal{L}^2(\Omega, \mathcal{F}_n, \mathcal{P})^\perp$, we have

$$\int_B W d\mathcal{P} = \int_B P_\infty W d\mathcal{P} \tag{27}$$

for any $B \in \mathcal{F}_n$. A standard $\pi - \lambda$ argument (Durrett, 1996, p. 263) shows that (27) holds for $B \in \mathcal{F}_\infty$ arbitrary; that is, $P_\infty W = E(W|\mathcal{F}_\infty)$. The conclusion is

$$\|E(W|\mathcal{F}_\infty) - E(W|\mathcal{F}_n)\| \to 0, \tag{28}$$

as $n \to \infty$.

We will appeal to Example 9 when dealing with stratification techniques. A variation of the last example is

Example 10 *Suppose that X is random variable with known and finite moments $EX^i, i = 1, 2, \ldots$. Define a sequence of complete subspaces (M_d) of $\mathcal{L}^2(\Omega, \sigma(X), \mathcal{P})$ by*

$$M_d = \left\{ Z \in \mathcal{L}^2(\Omega, \sigma(X), \mathcal{P}) : Z = \sum_{i=1}^{d} \beta_i (X^i - EX^i), \forall \beta_i \in \mathbb{R}, 1 \leq i \leq d \right\},$$

for $d = 1, 2, \ldots$. Clearly $M_1 \subseteq M_2 \subseteq \ldots \subseteq M_\infty$, where $M_\infty = \cup_{i=1}^{\infty} M_i$. Associated with each M_d there is, by Result 2, a projection operator P_d with range on M_d such that for $W \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, $\sigma(X) \subseteq \mathcal{F}$, with $EW = 0$:

$$P_d W = \sum_{i=1}^{d} \beta_i^* (X^i - EX^i), \tag{29}$$

for some constants $\beta_i^* = \beta_i^*(W), 1 \leq i \leq d$, possibly dependent on d (although this is not apparent from the notation). Because the sequence of operators (P_d) is (clearly) monotone increasing, Result 4 ensures the existence of a projection P_∞ in $\mathcal{L}^2(\Omega, \sigma(X), \mathcal{P})$ that satisfies

$$\|P_\infty W - P_d W\| \to 0, \tag{30}$$

as $d \to \infty$. Proceeding like in the last example it follows that $P_\infty W = E(W|X)$, for $W \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ arbitrary. In other words,

$$\left\| E(W|X) - \sum_{i=1}^{d} \beta_i^* (X^i - EX^i) \right\| \to 0, \tag{31}$$

as $d \to \infty$, by Equations (29) and (30).

We will use the last example in Section 6 to show how conditional Monte Carlo and control variates reduce variance cooperatively.

The rest of the chapter is devoted to provide an interpretation of this section's examples in terms of variance reduction techniques. We start with the method of control variates.

4 A Hilbert Space Approach to Control Variates

We build on the sequence of examples from the previous section; Glynn and Szechtman (2002) is a relevant reference for the issues discussed in this section.

Consider the setting of Example 5: $X_1, \ldots, X_d$ are zero-mean square integrable random variables with non-singular covariance matrix $\mathbf{\Sigma_{xx}}$ (although this was not needed in Example 5), and defined on the same probability space as the response Y, $EY^2 < \infty$. The goal is to estimate $\alpha = EY$ by averaging n i.i.d. replicates of $Y - \sum_{i=1}^d \lambda_i X_i$ to obtain $\hat{Y}_{CV}(\boldsymbol{\lambda})$ as in Equation (1). Clearly, $\operatorname{Var} \hat{Y}_{CV}(\boldsymbol{\lambda}) = 1/n \operatorname{Var}(Y - \sum_{i=1}^d \lambda_i X_i)$.

From Example 5, $\hat{Y}_{CV}(\boldsymbol{\lambda}^*)$ is the remainder from projecting Y on M; as M "grows" the norm of the remainder decreases. Also, because the scalars $\lambda_1^*, \ldots, \lambda_d^*$ that minimize $\operatorname{Var}(Y - \sum_{i=1}^d \lambda_i X_i)$ are also the numbers that result from projecting Y into M, from Result 2 and Equation (16) we know that

$$\langle Y - \sum_{i=1}^d \lambda_i^* X_i, Z \rangle = 0, \forall Z \in M.$$

In particular,

$$\langle Y - \sum_{i=1}^d \lambda_i^* X_i, \lambda_k^* X_k \rangle = 0, \text{ for } k = 1, \ldots, d. \tag{32}$$

Therefore, since $\operatorname{Cov}(Y, X_j) = \langle Y, X_j \rangle$, and $\operatorname{Cov}(X_i, X_j) = \langle X_i, X_j \rangle$ for $1 \leq i, j \leq d$,

$$\lambda_i^* = (\mathbf{\Sigma_{xx}^{-1}} \mathbf{\Sigma_{xy}})_i, \text{ for } i = 1, \ldots, d,$$

in concordance with Equation (3). When $X_1, \ldots, X_d$ is an orthogonal set, Equation (32) yields

$$\lambda_i^* = \frac{\langle Y, X_i \rangle}{\langle X_i, X_i \rangle} = \frac{\operatorname{Cov}(Y, X_i)}{\operatorname{Var} X_i}, i = 1, \ldots, d. \tag{33}$$

We re-interpret the results from Examples 5 through 7 in the CV context:

a) $\operatorname{Var} \hat{Y}_{CV}(\boldsymbol{\lambda}^*) \leq \operatorname{Var} \bar{Y}$ because

$$\begin{aligned}\operatorname{Var}\left(Y - \sum_{i=1}^{d} \lambda_i^* X_i\right) &= \|(I - P_M)(Y - EY)\|^2, \text{by Equation (20)} \\ &\leq \|I - P_M\|^2 \|Y - EY\|^2 \\ &\leq \operatorname{Var} Y,\end{aligned}$$

by using the projection operator properties of the last section.

b) If Y can be expressed as a linear combination of $X_1, \ldots, X_d$, then $\operatorname{Var}(Y - \sum_{i=1}^{d} \lambda_i X_i) = 0$ for some $\lambda_1, \ldots, \lambda_d$; this is Equation (21).

c) If the controls $X_1, \ldots, X_d$ are mutually orthogonal, then

$$\begin{aligned}\operatorname{Var}\left(Y - \sum_{i=1}^{d} \lambda_i X_i\right) &= \operatorname{Var} Y - \sum_{i=1}^{d} \frac{\operatorname{Cov}(Y, X_i)^2}{\operatorname{Var} X_i} \\ &= \operatorname{Var} Y \left(1 - \sum_{i=1}^{d} \rho_{yx_i}^2\right),\end{aligned}$$

by Equations (22) and (33), where ρ_{yx_i} is the correlation coefficient between Y and X_i, $i = 1, \ldots, d$.

d) With biased control variates in mind, apply Example 6 to the elements $(\bar{Y} - \alpha)$ and $(\bar{X} - \gamma)$ to get the optimal BCV coefficient

$$\lambda_n = \operatorname{Cov}(\bar{Y}, \bar{X}) / E(\bar{X} - \gamma)^2,$$

as expected from (8). That is, BCVs as presented in Section 2 arise from taking the remainder of the projection of $\bar{Y} - \alpha$ on the span of $\bar{X} - \gamma$. Because of (23) we have

$$\operatorname{Var} \hat{Y}_{CV}(\lambda^*) \leq \operatorname{MSE} \hat{Y}_{BCV}(\lambda_n).$$

e) Consider the setting of Example 7: There exists a zero-mean control variate $X \in \mathbb{R}$, and the output of the simulation are the sample points $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$. Let $\tilde{\mathbf{y}} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$, and define the estimator

$$\hat{Y}_{CV}(\lambda_n) = \frac{1}{n} \sum_{j=1}^{n} (y_j - \lambda_n x_j),$$

where $\lambda_n = \langle \mathbf{x}, \tilde{\mathbf{y}} \rangle / \|\mathbf{x}\|^2$. From Example 7 we know that λ_n arises from

projecting $\tilde{\mathbf{y}}$ on the span of $\mathbf{x}$: $P_S\tilde{\mathbf{y}} = \lambda_n \mathbf{x}$. Now, the sample variance is

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(y_j - \lambda_n x_j - \hat{Y}_{CV}(\lambda_n))^2 &= \frac{1}{n}\|(I - P_S)\tilde{\mathbf{y}}\|^2 - (\hat{Y}_{CV}(\lambda_n) - \bar{y})^2 \\ &= \frac{1}{n}(\|\tilde{\mathbf{y}}\|^2 - \lambda_n^2\|\mathbf{x}\|^2) + O(n^{-1}) \\ &= \frac{1}{n}\|\tilde{\mathbf{y}}\|^2(1 - \rho^2_{\tilde{\mathbf{y}},\mathbf{x}}) + O(n^{-1}),\end{aligned}$$

where $\rho_{\tilde{\mathbf{y}},\mathbf{x}} = \langle \mathbf{x}, \tilde{\mathbf{y}} \rangle / \|\mathbf{x}\|\|\tilde{\mathbf{y}}\|$, which makes precise the variance reduction achieved by projecting $\tilde{\mathbf{y}}$ on the span of $\mathbf{x}$ relative to the crude estimator sample variance $\|\tilde{\mathbf{y}}\|^2/n$.

Finally, we remark that there is no impediment in extending items a) through e) to the multi-response setting, where Y is a random vector.

5 Conditional Monte Carlo in Hilbert Space

In this section we address the method of conditional Monte Carlo, paying special attention to its connection with control variates; we follow Avramidis and Wilson (1996), and Loh (1995).

Suppose $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ and that we wish to compute $\alpha = EY$. Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ be such that $E(Y|X)$ can be analytically or numerically computed. Then

$$\hat{Y}_{CMC} = \frac{1}{n}\sum_{j=1}^{n} E(Y|X_i)$$

is an unbiased estimator of α, where the $E(Y|X_i)$ are found by first obtaining i.i.d. samples X_i and then computing $E(Y|X_i)$. We call $\hat{Y}_{CMC}$ the conditional Monte Carlo (CMC) estimator of α; remember that according to Example 4, $\hat{Y}_{CMC}$ results by projecting Y on $\mathcal{L}^2(\Omega, \sigma(X), \mathcal{P})$. The variability of $\hat{Y}_{CMC}$ is given by

$$\operatorname{Var}\hat{Y}_{CMC} = \frac{1}{n}\operatorname{Var} E(Y|X),$$

with Equation (19) implying that $\operatorname{Var}\hat{Y}_{CMC} \leq \operatorname{Var}\bar{Y}$. Specifically, CMC eliminates the $E\operatorname{Var}(Y|X)$ term from the variance of Y.

Sampling from $Y - \lambda(Y - E(Y|X))$ also provides an unbiased estimator of α

for any $\lambda \in \mathbb{R}$. By Equation (3),

$$\begin{aligned}\lambda^* &= \frac{\langle Y, (I - P_{\sigma(X)})Y\rangle}{\|(I - P_{\sigma(X)})Y\|^2} \\ &= \frac{\langle P_{\sigma(X)}Y + (I - P_{\sigma(X)})Y, (I - P_{\sigma(X)})Y\rangle}{\|(I - P_{\sigma(X)})Y\|^2} \\ &= 1.\end{aligned} \tag{34}$$

This shows that CMC is optimal from a CV perspective. Avramidis and Wilson (1996), and Loh (1995), generalize this approach: Let Z be a zero-mean random variable in $\mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, and X a random variable in $\mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ for which both $E(Y|X)$ and $E(Z|X)$ can be determined. Then sampling from

$$Y - \lambda_1(Y - E(Y|X)) - \lambda_2 E(Z|X) - \lambda_3 Z \tag{35}$$

can be used to form the standard means based estimator for α, for all $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. Repeating the logic leading to (34) we obtain

$$\lambda_1^* = 1, \lambda_2^* = \frac{\mathrm{Cov}(E(Y|X), E(Z|X))}{\mathrm{Var}\, E(Z|X)}, \text{ and } \lambda_3^* = 0.$$

The conclusion is that,

$$\begin{aligned}&\mathrm{Var}\left(E(Y|X) - \frac{\mathrm{Cov}(E(Y|X), E(Z|X))}{\mathrm{Var}\, E(Z|X)} E(Z|X)\right) \\ &\qquad \leq \begin{cases} \mathrm{Var}\, Y, \\ \mathrm{Var}\, E(Y|X), \\ \mathrm{Var}\left(Y - \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Var}\, Z} Z\right), \\ \mathrm{Var}\left(Y - \frac{\mathrm{Cov}(Y,E(Z|X))}{\mathrm{Var}\, E(Z|X)} E(Z|X)\right), \\ \mathrm{Var}\left(E(Y|X) - \frac{\mathrm{Cov}(E(Y|X),Z)}{\mathrm{Var}\, Z} Z\right). \end{cases}\end{aligned}$$

In particular, Loh (1995) considers the case of $Z = X$ almost surely in (35), and Avramidis and Wilson (1996) fix $\lambda_1 = 1$ and $\lambda_3 = 0$ in (35). From the norm perspective,

$$\|E(Y|X) - \alpha - \lambda_2^* E(Z|X)\|^2 = \|Y - \alpha\|^2 - \|Y - E(Y|X)\|^2 - \|\lambda_2^* E(Z|X)\|^2$$

makes precise the variance eliminated when sampling from $E(Y|X) - \lambda_2^* E(Z|X)$.

6 Control Variates and Conditional Monte Carlo from a Hilbert Space Perspective

We now discuss how CMC and CV can be combined to reduce variance co-operatively; the results of this section appear in Loh (1995), and Glynn and Szechtman (2002).

Suppose the setting of Example 10: There exists a random variable $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$ such that the moments EX^i are known with $E|X|^i < \infty$ for all $i = 1, 2, \ldots$. Given a random variable $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$, the goal is to find $\alpha = EY$ by running a Monte Carlo simulation that uses the knowledge about the moments of X to increase simulation efficiency. Suppose we can sample from either

a) $E(Y|X) - \sum_{i=1}^{d} \lambda_i^*(X^i - EX^i)$

or

b) $Y - \sum_{i=1}^{d} \lambda_i^*(X^i - EX^i)$,

to form the standard estimator for α, where the λ_i^* are determined by applying Equation (3) on $E(Y|X)$ and on the controls $(X^1 - EX^1, \ldots, X^d - EX^d)$. From the developments of Example 10 it is a short step to:

a) Take $W = E(Y|X) - \alpha$ in Equation (31), consequently:

$$\operatorname{Var}\left(E(Y|X) - \sum_{i=1}^{d} \lambda_i^*(X^i - EX^i)\right) \to 0, \tag{36}$$

as $d \to \infty$.

b) The triangle inequality and $W = Y - \alpha$ in Equation (31) result in

$$\operatorname{Var}\left(Y - \sum_{i=1}^{d} \lambda_i^*(X^i - EX^i)\right) \to E \operatorname{Var}(Y|X), \tag{37}$$

as $d \to \infty$.

The interpretation of a) is that CV and CMC reduce variance concurrently: $E(Y|X)$ eliminates the $E \operatorname{Var}(Y|X)$ part of $\operatorname{Var} Y$, while $\sum_{i=1}^{d} \lambda_i^*(X^i - EX^i)$ asymptotically cancels $\operatorname{Var} E(Y|X)$. The effect of $\sum_{i=1}^{d} \lambda_i^*(X^i - EX^i)$ in part b) is to asymptotically eliminate the variance component due to $E(Y|X)$ when using $E(Y|X)$ in the simulation is not possible.

7 Weighted Monte Carlo

In this section we consider the asymptotic behavior of weighted Monte Carlo (WMC) estimators, for a large class of objective functions. We rely on Glasserman and Yu (2005), and Glasserman (2004), which make precise the connection between WMC and CVs for separable convex objective functions. Initial results, under weaker assumptions and just for one class of objectives were obtained in Szechtman and Glynn (2001), and in Glynn and Szechtman (2002). Applications of weighted estimators to model calibration in the finance context are presented in Avellaneda et al. (2001), and in Avellaneda and Gamba (2000).

Consider the standard CV setting: $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ are i.i.d. samples of jointly distributed random elements $(Y, \mathbf{X}) \in (\mathbb{R}, \mathbb{R}^d)$ with non-singular covariance matrix $\mathbf{\Sigma}$ and, without loss of generality, $E\mathbf{X} = \mathbf{0}$ componentwise. The goal is to compute $\alpha = EY$ by Monte Carlo simulation, using information about the means $E\mathbf{X}$ to reduce estimator variance. Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly convex and continuously differentiable function, and suppose that the weights $w^*_{1,n}, \ldots, w^*_{n,n}$

$$\text{minimize} \sum_{k=1}^{n} f(w_{k,n}) \tag{38}$$

$$\text{subject to } \frac{1}{n} \sum_{k=1}^{n} w_{k,n} = 1 \tag{39}$$

$$\frac{1}{n} \sum_{k=1}^{n} w_{k,n} \mathbf{X}_k = \mathbf{0}. \tag{40}$$

Then, the WMC estimator of α takes the form

$$\hat{Y}_{WMC} = \frac{1}{n} \sum_{k=1}^{n} w^*_{k,n} Y_k.$$

The following observations review some key properties of WMC: The weight applied to each replication i is $w^*_{i,n}/n$, rather than the weight $1/n$ used to form the sample mean $\bar{Y}$. A feasible set of weights is one that makes $\hat{Y}_{WMC}$ unbiased (cf. constraint (39)), and that forces the weighted average of the control samples to match their known mean (cf. constraint (40)). For every n sufficiently large $\mathbf{0}$ $(=E\mathbf{X})$ belongs to the convex hull of the replicates $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and therefore the constraint set is non-empty. The objective function in (38), being strictly convex, ensures uniqueness of the optimal solution if the optimal solution is finite. If $w_{k,n} \geq 0, 1 \leq k \leq n$, were additional constraints, a feasible set of weights $w_{1,n}, \ldots, w_{n,n}$ would determine a probability mass function $1/n \sum_{k=1}^{n} \delta_{\mathbf{X}_k}(\cdot) w_{k,n}$, where $\delta_{\mathbf{x}}(\mathbf{z}) = 1$ if $\mathbf{z} = \mathbf{x}$ and is equal to zero otherwise. However, as discussed in Hesterberg and Nelson (1998), $P(w_{k,n} < 0) = o(n^{-p})$

uniformly in $1 \leq k \leq n$ if $E(\|\mathbf{X}\|^p) < \infty$ indicates that the non-negativity constraints are asymptotically not binding; see Szechtman and Glynn (2001) for an example of this scenario.

There are different f's depending on the application setting. For example: $f(w) = -\log w$ results in maximizing empirical likelihood; discussed in Szechtman and Glynn (2001). The function $f(w) = -w \log w$ yields an entropy maximization objective; this is the subject of Avellaneda and Gamba (2000), and Avellaneda et al. (2001). The important case of $f(w) = w^2$ is considered next, the optimization problem being to

$$\begin{aligned} &\text{minimize} \sum_{k=1}^{n} w_{k,n}^2 \\ &\text{subject to} \frac{1}{n} \sum_{k=1}^{n} w_{k,n} = 1 \\ &\frac{1}{n} \sum_{k=1}^{n} w_{k,n} \mathbf{X}_k = \mathbf{0}. \end{aligned} \tag{41}$$

Solving the optimization problem yields (Glasserman and Yu, 2005) optimal weights given by

$$w_{k,n}^* = 1 - \bar{\mathbf{X}}^T \mathbf{M}^{-1} (\mathbf{X}_k - \bar{\mathbf{X}}), \text{ for } k = 1, \ldots, n, \tag{42}$$

where $\mathbf{M} \in \mathbb{R}^{d \times d}$ is the matrix with elements $M_{i,j} = 1/n \sum_{k=1}^{n} (X_{i,k} - \bar{X})(X_{j,k} - \bar{X})$; $\mathbf{M}^{-1}$ exists for all n large enough because $\mathbf{M} \to \mathbf{\Sigma_{xx}}$ a.s. componentwise. Rearranging terms immediately gives

$$\frac{1}{n} \sum_{k=1}^{n} w_{k,n}^* Y_k = \hat{Y}_{CV}(\boldsymbol{\lambda}_n), \tag{43}$$

with the benefit that the optimal weights do not depend on the Y_k, which makes this approach advantageous when using CVs for quantile estimation; see Hesterberg and Nelson (1998) for details.

Regarding Hilbert spaces, consider the space $\mathbb{R}^n$ (cf. Example 2) and the set

$$A(n) = \left\{ \mathbf{w}(n) = (w_{1,n}, \ldots, w_{n,n}) \in \mathbb{R}^n : \frac{1}{n} \sum_{k=1}^{n} w_{k,n} = 1 \text{ and } \frac{1}{n} \sum_{k=1}^{n} w_{k,n} \mathbf{X}_k = \mathbf{0} \right\}.$$

It can be verified that $A(n)$ meets the conditions of Result 3 for every n sufficiently large, and consequently it is fair to ask: What element in $A(n)$ is closest to $\mathbf{1}(n) = (1, \ldots, 1) \in \mathbb{R}^n$? That is, which element $\mathbf{w}^*(n) = (w_{1,n}^*, \ldots, w_{n,n}^*) \in \mathbb{R}^n$

$$\begin{aligned} &\text{minimizes} \|\mathbf{1}(n) - \mathbf{w}(n)\| \\ &\text{subject to } \mathbf{w}(n) \in A(n)? \end{aligned}$$

This problem yields the same solution as problem (41); doing simple algebra it is easy to verify that $\mathbf{w}^*(n)$ with components as in Equation (42) satisfies condition (26). The conclusion is that $\mathbf{w}^*(n)$ is the closest point in $A(n)$ to $\mathbf{1}(n)$, the vector of crude sample weights. Would this Hilbert space approach to WMC work with f's that are not quadratic? Yes, as long as the metric induced by the inner product meets the defining properties of a metric.

The main result concerning WMC and CV, proved in Glasserman and Yu (2005) under certain conditions on $\mathbf{X}$, Y, f, and the Lagrange multipliers associated with constraints (39) and (40), is that

$$\hat{Y}_{WMC} = \hat{Y}_{CV} + O_p(n^{-1}),$$

and,

$$\sqrt{n}(\hat{Y}_{WMC} - \alpha) \Rightarrow N(0, \sigma^2_{WMC}),$$

as $n \to \infty$ where

$$\sigma^2_{WMC} = \sigma^2_{CV},$$

and $O_p(a_n)$ stands for a sequence of random variables $(\xi_n : n \geq 1)$ such that for all $\epsilon > 0$ and some constant δ, $P(|\xi_n| \geq a_n\delta) < \epsilon$. The last result provides support to the statement that $\hat{Y}_{WMC}$ and $\hat{Y}_{CV}$ are asymptotically identical.

8 Stratification Techniques

In this section we discuss stratification methods emphasizing the connection with the Hilbert space and CVs ideas already developed. Refer to Fishman (1996), Glasserman et al. (1999), and Glynn and Szechtman (2002) for more details.

Suppose that we wish to compute $\alpha = EY$, for some random variable $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$. Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathcal{P})$. The method of stratification arises when there is a collection of disjoint sets ("strata") $(\mathcal{A}_i : 1 \leq i \leq d)$ in the range of X such that $P(X \in \cup_{i=1}^d \mathcal{A}_i) = 1$ and $P(X \in \mathcal{A}_i) = p_i$ is known for every $1 \leq i \leq d$. Then, assuming that one can obtain i.i.d. replicates $(Y_{i,k} : 1 \leq k \leq n_i)$ from $P(Y \in \cdot | X \in \mathcal{A}_i)$, $1 \leq i \leq d$, the estimator of α given by

$$\sum_{i=1}^{d} \frac{p_i}{n_i} \sum_{k=1}^{n_i} Y_{i,k} \tag{44}$$

is unbiased, where n_i is the number of replicates sampled from $P(Y \in \cdot | X \in \mathcal{A}_i)$.

For a total number of replications $n = \sum_{i=1}^d n_i$, proportional stratification allocates $n_i = np_i$ samples to strata $\mathcal{A}_i, 1 \leq i \leq d$, where for simplicity we

assume that the np_i are integers. The estimator of Equation (44) is then called the proportional stratification (PS) estimator

$$\hat{Y}_{PS} = \frac{1}{n} \sum_{i=1}^{d} \sum_{k=1}^{np_i} Y_{i,k},$$

with variance given by

$$\begin{aligned} \operatorname{Var} \hat{Y}_{PS} &= \frac{1}{n} \sum_{i=1}^{d} p_i \operatorname{Var}(Y|Z=i) \\ &= \frac{1}{n} E \operatorname{Var}(Y|Z), \end{aligned} \tag{45}$$

where the random variable $Z = \sum_{i=1}^{d} iI(X \in \mathcal{A}_i)$.

One implication of Equation (45) is that if Y is not constant inside each strata, then $\operatorname{Var} \hat{Y}_{PS} > 0$, so that proportional stratification does not eliminate the variability of Y inside strata, but rather the variability of $E(Y|Z)$ across strata. In addition, Equation (45), jointly with the variance decomposition formula (19), quantifies the per-replication variance reduction achieved by proportional stratification: $E \operatorname{Var}(Y|Z) = \operatorname{Var} Y - \operatorname{Var} E(Y|Z)$. Observe that although PS is relatively simple to implement, it does not provide the optimal sample allocation n_i per strata; see Glasserman (2004, p. 217) for more details.

From a CV perspective, PS acts like applying $E(Y|Z) - \alpha$ as a CV on Y; $\hat{Y}_{PS}$ achieves the same variance reduction as that obtained by averaging i.i.d. replications of $Y - (E(Y|Z) - \alpha)$. Of course, sampling from the distribution of $Y - (E(Y|Z) - \alpha)$ is impractical because α is unknown.

Regarding Hilbert spaces, Equation (45) is simply

$$n \operatorname{Var} \hat{Y}_{PS} = \|(I - P_{\sigma(Z)})Y\|^2. \tag{46}$$

In addition, $\hat{Y}_{PS}$ satisfies the following CLT:

$$n^{1/2}(\hat{Y}_{PS} - \alpha) \Rightarrow N(0, \sigma_{PS}^2) \text{ as } n \to \infty,$$

where $\sigma_{PS}^2 = E \operatorname{Var}(Y|Z)$, which enables the construction of asymptotically valid confidence intervals for α.

Post-stratification offers an alternative to proportional stratification when sampling from $P(Y \in \cdot | X \in \mathcal{A}_i)$ is not possible, but when it is possible to sample from the distribution of (X, Y). Specifically, we construct the unbiased estimator

$$\hat{Y}_{pST} = \sum_{i=1}^{d} p_i \frac{\sum_{k=1}^{n} Y_k I(X_k \in \mathcal{A}_i)}{\sum_{j=1}^{n} I(X_j \in \mathcal{A}_i)}.$$

Using the Delta method (cf. Section 2) it is easy to prove a CLT for $\hat{Y}_{pST}$:

$$n^{1/2}(\hat{Y}_{pST} - \alpha) \Rightarrow N(0, \sigma^2_{pST}),$$

as $n \to \infty$, where $\sigma^2_{pST} = E\,\mathrm{Var}(Y|Z)$. However, for every stratum we know a priori that $EI(X \in \mathcal{A}_i) = p_i$, which suggests the use of the vector $(I(X \in \mathcal{A}_i) - p_i : 1 \le i \le d)$ as a control. More specifically, an unbiased CV estimator is given by

$$\hat{Y}_{CV}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \sum_{i=1}^{d}\lambda_i(I(X_j \in \mathcal{A}_i) - p_i)\right).$$

Using Equation (3), the optimal coefficients $\lambda_i^*, 1 \le i \le d$ are immediately found to be

$$\lambda_i^* = E(Y|Z = i), \text{ for } 1 \le i \le d.$$

That is,

$$\hat{Y}_{CV}(\boldsymbol{\lambda}^*) = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - (E(Y|Z_j) - \alpha)\right),$$

and

$$\mathrm{Var}\,\hat{Y}_{CV}(\boldsymbol{\lambda}^*) = \frac{1}{n}E\,\mathrm{Var}(Y|Z).$$

Therefore, $n(\mathrm{Var}\,\hat{Y}_{CV}(\boldsymbol{\lambda}^*) - \mathrm{Var}\,\hat{Y}_{pST}) \to 0$ as $n \to \infty$. With a little more effort, it can be shown that

$$n^{1/2}\left(\hat{Y}_{pST} - \hat{Y}_{CV}(\boldsymbol{\lambda}^*)\right) \Rightarrow 0,$$

as $n \to \infty$. In other words, $\hat{Y}_{pST}$ and $\hat{Y}_{CV}(\boldsymbol{\lambda}^*)$ have the same distribution up to an error of order $o_p(n^{-1/2})$, as $n \to \infty$; where $o_p(a_n)$ denotes a sequence of random variables $(\xi_n : n \ge 1)$ such that $a_n^{-1}\xi_n \Rightarrow 0$ as $n \to \infty$.

Given the strata $(\mathcal{A}_i : 1 \le i \le d)$, it is always possible to find finer strata that further reduce estimator variance. In the case of proportional stratification, suppose that it is possible to split each stratum $\mathcal{A}_i$ into integer $n_i = np_i$ strata $(\mathcal{A}_{i,k} : 1 \le k \le n_i)$ such that $P(X \in \mathcal{A}_{i,k}) = 1/n$; i.e., the bivariate random vector $V_n = \sum_{i=1}^{d}\sum_{k=1}^{n_i}(i,k)I(X \in \mathcal{A}_{i,k})$ is uniformly distributed on the lattice $\{(i,k) : 1 \le i \le d, 1 \le k \le n_i\}$. Assume in addition that it is possible to sample from $P(Y \in \cdot|X \in \mathcal{A}_{i,k})$. Then the refined proportional stratification (rST) estimator is

$$\hat{Y}_{rST} = \frac{1}{n}\sum_{i=1}^{d}\sum_{k=1}^{np_i}Y_{i,k},$$

where the $Y_{i,k}$ are sampled from $P(Y \in \cdot|X \in \mathcal{A}_{i,k})$. Proceeding as in (45), we arrive at

$$\mathrm{Var}\,\hat{Y}_{rST} = \frac{1}{n}E\,\mathrm{Var}(Y|V_n).$$

The fact that $\|E(Y|V_n) - EY\|^2 = \|E(Y|V_n) - E(Y|Z)\|^2 + \|E(Y|Z) - EY\|^2$ shows that $\operatorname{Var} E(Y|V_n) \geq \operatorname{Var} E(Y|Z)$, and therefore

$$\operatorname{Var} \hat{Y}_{rST} \leq \operatorname{Var} \hat{Y}_{PS}.$$

With regards to Example 9, the conditions leading to Equation (28) apply, so that as $n \to \infty$

$$\operatorname{Var}(E(Y|X) - E(Y|V_n)) \to 0,$$

and

$$\operatorname{Var}(Y - E(Y|V_n)) \to E \operatorname{Var}(Y|X).$$

In particular, $n \operatorname{Var} \hat{Y}_{rST} \to E \operatorname{Var}(Y|X)$. This result should come as no surprise because as n grows we get to know the full distribution of X, not unlike the setting of Equations (36) – (37): rST presumes knowledge of an increasing sequence of σ-algebras that converge to $\sigma(X)$, whereas in Equations (36) – (37) we have information about the full sequence of moments of X.

As to control variates, rST produces the same estimator variance as the standard CV estimator formed by i.i.d. sampling from $Y - (E(Y|X) - \alpha)$, as $n \to \infty$. Similar to (46), we can write $n \operatorname{Var} \hat{Y}_{rST} \to \|(I - P_{\sigma(X)})Y\|^2$, as $n \to \infty$. Finally, the CLT satisfied by $\hat{Y}_{rST}$ is

$$n^{1/2}(\hat{Y}_{rST} - \alpha) \Rightarrow N(0, \sigma^2_{rST}),$$

as $n \to \infty$, where $\sigma^2_{rST} = E \operatorname{Var}(Y|X)$.

To conclude this section, we mention the link between post-stratification and WMC. Write

$$w^*_{k,n} = \sum_{i=1}^{d} \frac{p_i I(X_k \in \mathcal{A}_i)}{\sum_{j=1}^{n} I(X_j \in \mathcal{A}_i)}, \tag{47}$$

for $1 \leq k \leq n$, then the WMC estimator $\hat{Y}_{WMC} = \sum_{k=1}^{n} w^*_{k,n} Y_k$ equals $\hat{Y}_{pST}$. It can be confirmed that the weights given in Equation (47) are the solution of the optimization problem with objective function $\min \sum_{k=1}^{n} w^2_{k,n}$ and constraints $\sum_{k=1}^{n} w_{k,n} I(X_k \in \mathcal{A}_i) = p_i, 1 \leq i \leq d$, and $\sum_{k=1}^{n} w_{k,n} = 1$. Interpreting at face value, $\mathbf{w}^*_n$ with elements as in (47) is the closest point in the set determined by the constraints to the vector consisting of n ones.

9 Latin Hypercube Sampling

We now discuss the method of Latin hypercube sampling (LHS) from a Hilbert space and CV perspective. McKay et al. (1979), Stein (1987), Owen (1992), and Loh (1996) are standard references for LHS. We rely on Mathé (2000), which gives a good account of LHS from a Hilbert space point of view.

Avramidis and Wilson (1996) is also a valuable reference for the issues we consider.

Suppose the setting of Examples 3 and 8: We have mutually independent random variables $X_1, \ldots, X_d$, each with known distribution function F_i, and the goal is to compute

$$\alpha = Ef(\mathbf{X}) = \int f(\mathbf{x}) dF(\mathbf{x})$$

via simulation, where $f : \mathbb{R}^d \to \mathbb{R}$ is a square integrable function with respect to $F(\mathbf{x}) = \prod_{d=1}^{n} F_i(x_i)$, $\mathbf{x} = (x_1, \ldots, x_d)$, and $\mathbf{X} = (X_1, \ldots, X_d)$.

LHS generates samples of $\mathbf{X}$ as follows:

(i) Tile $[0,1)^d$ into n^d hypercubes $\Delta_{l_1,\ldots,l_d} = \prod_{i=1}^{d} [\frac{l_i - 1}{n}, \frac{l_i}{n})$, $l_i = 1, \ldots, n$, $i = 1, \ldots, d$, each of volume n^{-d}.
(ii) Generate d uniform independent permutations $(\pi_1(\cdot), \ldots, \pi_d(\cdot))$ of $\{1, \ldots, n\}$.
(iii) Use the output of (ii) to choose n hypercubes from (i): $\Delta_{\pi_1(k),\ldots,\pi_d(k)}$ for the k'th tile, $k = 1, \ldots, n$.
(iv) Uniformly select a point from within each $\Delta_{\pi_1(k),\ldots,\pi_d(k)}$, and generate $X_{i,k}$ by inverting F_i at that point.

Notice that (i) – (iv) are

$$X_{i,k} = F_i^{-1}\left(\frac{\pi_i(k) - 1 + U_i(k)}{n}\right), 1 \le i \le d, \text{ and } 1 \le k \le n, \qquad (48)$$

where the $U_i(k)$ are i.i.d. uniform on $[0,1]$. The LHS estimator is the average of the n samples $f(\mathbf{X}_k)$, each $\mathbf{X}_k = (X_{1,k}, \ldots, X_{d,k})$ obtained according to (48):

$$\hat{Y}_{LHS} = \frac{1}{n} \sum_{k=1}^{n} f(\mathbf{X}_k).$$

As in refined stratification, given a sample size n, LHS assigns one sample to each strata $\mathcal{A}_{i,k}$ given by

$$\mathcal{A}_{i,k} = \left[F_i^{-1}\left(\frac{k-1}{n}\right), F_i^{-1}\left(\frac{k}{n}\right)\right), 1 \le i \le d, 1 \le k \le n,$$

with the sample uniformly distributed within the strata. Where refined proportional stratification applied to a particular X_i asymptotically eliminates the variance due to $E(f(\mathbf{X})|\mathcal{F}_i)$ (cf. Example 8 for the definition of $\mathcal{F}_i$) along just one dimension i, LHS asymptotically eliminates $\mathrm{Var} \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_i)$ at the same rate used by rST used to eliminate only $\mathrm{Var}\, E(f(\mathbf{X})|\mathcal{F}_i)$.

Figure 2 illustrates LHS with $d = 2$, $n = 4$. Each dot in the lower left square is a sample from $(\pi_i(k) - 1 + U_i(k))/n$. The position of each dot within a square

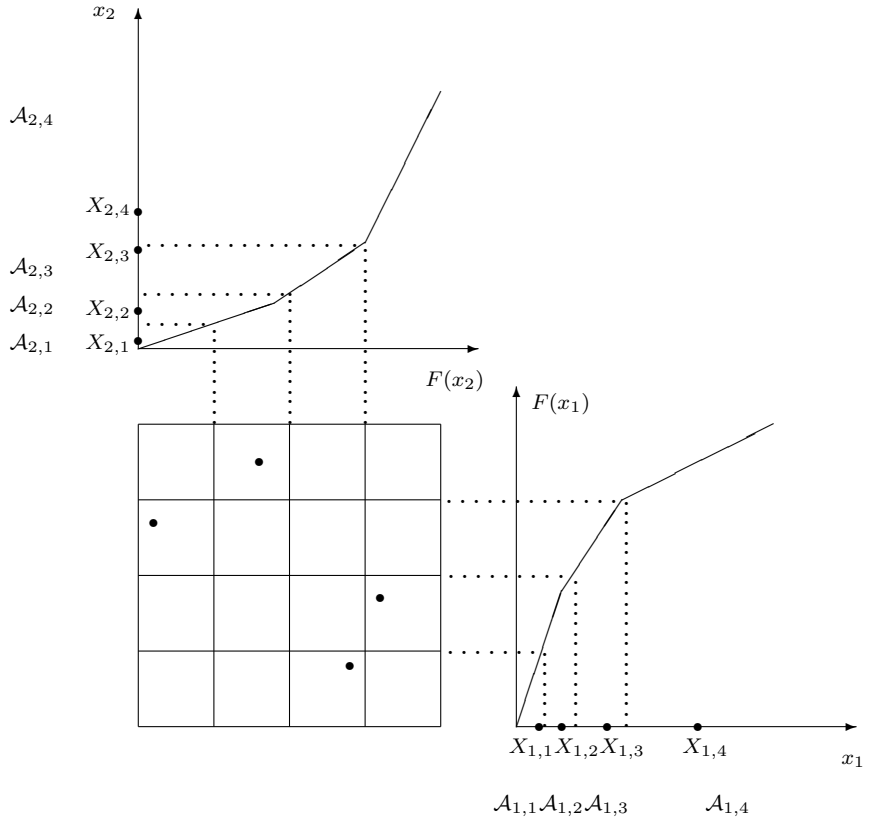

Fig. 2. Latin hypercube sampling

is uniformly distributed according to the $U_i(k)$; the permutations $\pi_i(k)$ ensure that there is just one dot per row and per column. The lower right region depicts F_1 and the four equiprobable strata for X_1, with one sample point $F_1^{-1}((\pi_1(k) - 1 + U_1(k))/n)$ per strata $\mathcal{A}_{1,k}$; the final output are the samples $X_{1,1}, \ldots, X_{1,4}$. In the upper left F_2 is pictured with the axes inverted; it has the same explanation as that of F_1, with the final output being the samples $X_{2,1}, \ldots, X_{2,4}$.

Stein (1987) demonstrates that when $f \in \mathcal{L}^2(dF)$,

$$\operatorname{Var} \hat{Y}_{LHS} = \frac{1}{n} \operatorname{Var} \left(f(\mathbf{X}) - \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_i) \right) + o(n^{-1}), \tag{49}$$

as $n \to \infty$, which makes precise the variance reduction achieved by LHS, up to order $o(n^{-1})$.

The CLT satisfied by $\hat{Y}_{LHS}$, proved in Owen (1992) under the condition that $f(F^{-1}(\cdot))$ is bounded on $[0,1]^d$ is

$$n^{1/2}(\hat{Y}_{LHS} - \alpha) \Rightarrow N(0, \sigma^2_{LHS}), \tag{50}$$

as $n \to \infty$, where

$$\sigma^2_{LHS} = \operatorname{Var}\left(f(\mathbf{X}) - \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_i)\right).$$

Equation (50) provides theoretical support for the construction of a valid confidence interval for α, whose width depends on σ_{LHS}. This term is generally not known prior to the simulation, nor easily estimated from the simulation data. Section 3 of Owen (1992) deals with the estimation of σ^2_{LHS} from LHS data, and shows how to use the LHS samples to find an estimator for σ^2_{LHS} which is within $n^{-1/2}$ in probability of σ^2_{LHS} as $n \to \infty$. This permits the formation of an asymptotically valid confidence interval for α.

As to standard Monte Carlo, Owen (1997) proves that LHS is no less efficient because

$$\operatorname{Var} \hat{Y}_{LHS} \leq \frac{\operatorname{Var} f(\mathbf{X})}{n-1},$$

for all $n \geq 2$ and $d \geq 2$. In other words, even if the variance eliminated by LHS, $\operatorname{Var} \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_i)$, is small, LHS with n samples is no less efficient than standard Monte Carlo with $n-1$ replications. Of course, the only measure of efficiency in this argument is variance, a more complete analysis would take into account the computational cost of generating sample variates.

Example 8 leads into the Hilbert interpretation of LHS. In particular, Equations (25) and (49) imply

$$\lim_{n\to\infty} n \operatorname{Var} \hat{Y}_{LHS} = \operatorname{Var}\left(f(\mathbf{X}) - \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_l)\right) = \|(I - P_M)f\|^2. \quad (51)$$

That is, LHS takes the remainder from projecting f on the subspace M determined by the span of the linear combinations of univariate functions. Using the last equation, LHS eliminates variance because

$$\begin{aligned}
\operatorname{Var} f(\mathbf{X}) &= \|(I - P_0)f\|^2 \\
&= \|P_M(I - P_0)f\|^2 + \|(I - P_M)(I - P_0)f\|^2 \\
&\geq \|(I - P_M)(I - P_0)f\|^2 \\
&= \|(I - P_M)f\|^2,
\end{aligned}$$

where

$$\|P_M(I - P_0)f\|^2 = \sum_{i=1}^{d} \|P_i(I - P_0)f\|^2 = \sum_{i=1}^{d} \operatorname{Var} E(f(\mathbf{X})|\mathcal{F}_i)$$

is the variance eliminated by LHS. If $f \in M$, then Equation (51) also shows that $n \operatorname{Var} \hat{Y}_{LHS} \to 0$ as $n \to \infty$; in other words, LHS asymptotically eliminates all the variance of f if f is a sum of univariate functions.

Much more can be said about Var $f(\mathbf{X})$. Suppose for simplicity that the X_i are Uniform $[0,1]$ random variables, so that $\alpha = \int_0^1 f(\mathbf{x})d\mathbf{x}$. Let $u \subseteq \{1,2,\ldots,d\}$ and define $d\mathbf{x}^{-u} = \prod_{j\notin u} dx_j$. Then, if f is square integrable, there is a unique recursion

$$f_u(\mathbf{x}) = \int f(\mathbf{x})d\mathbf{x}^{-u} - \sum_{v\subset u} f_v(\mathbf{x}) \tag{52}$$

with the property that $\int_0^1 f_u(\mathbf{x})dx_j = 0$ for every $j \in u$, and

$$f(\mathbf{x}) = \sum_{u\subseteq\{1,2,\ldots,d\}} f_u(\mathbf{x});$$

see, for example, Jiang (2003) for a proof. Recursion (52) actually is the Gram-Schmidt process, and it splits f into 2^d orthogonal components such that

$$\int f_u(\mathbf{x})f_v(\mathbf{x})d\mathbf{x} = 0,$$

for $u \neq v$, and

$$\sigma^2 = \sum_{|u|>0} \sigma_u^2, \tag{53}$$

where $\sigma^2 = \int(f(\mathbf{x}) - \alpha)^2 d\mathbf{x}$ is the variance of $f(\mathbf{X})$, and $\sigma_u^2 = \int f_u^2(\mathbf{x})d\mathbf{x}$ is the variance of $f_u(\mathbf{X})$. The conclusion in this context is that LHS eliminates the variance of the f_u for all u with $|u| = 1$. The setting of this paragraph is known as functional ANOVA; see Chapter 13 for more details on this topic.

Considering control variates, we can say that the estimator formed by averaging i.i.d. replicates of $f(\mathbf{X}) - \sum_{i=1}^d (E(f(\mathbf{X})|\mathcal{F}_i) - \alpha)$ has the same asymptotic variance as $\hat{Y}_{LHS}$. Moreover, given a zero-mean control variate $h(\mathbf{X})$, h a deterministic function, obtain n Latin hypercube samples of $\mathbf{X}$ using (48), and form the combined LHS+CV estimator

$$\hat{Y}_{LHS+CV}(\lambda) = \frac{1}{n}\sum_{k=1}^n \left(f(\mathbf{X}_k) - \lambda h(\mathbf{X}_k)\right).$$

Then, $n(\operatorname{Var}\hat{Y}_{LHS+CV}(\lambda) - \operatorname{Var}\hat{Y}_{LHS}) \to 0$ for any control of the type $h(\mathbf{X}) = \sum_{i=1}^d h_i(X_i)$ because $h \in M$ and property a) of the projection operator together imply $(I - P_M)(f - \lambda h) = (I - P_M)f$ for all $\lambda \in \mathbb{R}$. Using Equations (4) and (49):

$$\operatorname{Var}\hat{Y}_{LHS+CV}(\lambda^*) = \operatorname{Var}\hat{Y}_{LHS}(1 - \rho^2),$$

where ρ^2 is the square of the correlation coefficient between

$$f(\mathbf{X}) - \sum_{i=1}^d E(f(\mathbf{X})|\mathcal{F}_i) \text{ and } h(\mathbf{X}) - \sum_{i=1}^d E(h(\mathbf{X})|\mathcal{F}_i).$$

In other words, a good CV for $f(\mathbf{X})$ in the LHS context is one that maximizes the absolute value of the correlation coefficient of its non-additive part with the non-additive part of $f(\mathbf{X})$. Notice that λ^* is the optimal CV coefficient

associated with the response $f(\mathbf{X}) - \sum_{i=1}^{d} E(f(\mathbf{X})|\mathcal{F}_i)$ and the CV $h(\mathbf{X}) - \sum_{i=1}^{d} E(h(\mathbf{X})|\mathcal{F}_i)$; refer to Owen (1992) for the estimation of λ^* from sample data.

Additional guidance for effective CVs in the LHS setting is provided by (53): Choose a CV with non-additive part that is highly correlated with f_u's, $|u| > 1$, that have σ_u^2 large.

As regards weighted Monte Carlo, consider using LHS to generate replicates $\mathbf{X_1}, \ldots, \mathbf{X}_n$, and the optimization problem

$$\begin{aligned} &\text{minimize} \sum_{k=1}^{n} w_{k,n}^2 \\ &\text{subject to } \sum_{k=1}^{n} w_{k,n} = 1 \\ &\sum_{k=1}^{n} w_{k,n} I(X_{i,k} \in \mathcal{A}_{i,j}) = \frac{1}{n}, \text{ for } i = 1, \ldots, d \text{ and } j = 1, \ldots, n. \end{aligned} \tag{54}$$

Clearly $w_{k,n}^* = 1/n$, $k = 1, \ldots, n$, is feasible for (54), and it is also optimal by the developments of Section 7, so that the WMC estimator $\sum_{k=1}^{n} w_{k,n}^* f(\mathbf{X}_k)$ coincides with $\hat{Y}_{LHS}$ for every $n \geq 1$. For $d = 1$, problem (54) furnishes a WMC estimator that equals $\hat{Y}_{rST}$; in other words, for $d = 1$ LHS yields the same variance reduction as rST in the limit as $n \to \infty$.

10 A Numerical Example

In this section we present a numerical example that supports many of the results discussed in the chapter. Consider the stochastic activity network (SAN) of Loh (1995) depicted in Figure 3 (see also the SAN discussion in Chapter 1) with arcs X_1, X_2, X_3, X_4, X_5 that are independent random variables that represent activity durations. The problem of numerically computing the expected duration of the shortest path that leads from the source node a to the sink node z involves estimating $\alpha = EY$, where $Y = \min\{X_1 + X_2, X_1 + X_3 + X_5, X_4 + X_5\}$. For the purposes of this example, we assume that the X_i's are exponentially distributed with parameters $\mu_1 = 1.1, \mu_2 = 2.7, \mu_3 = 1.1, \mu_4 = 2.5, \mu_5 = 1.2$.

Given an inner sample size n, we wish to appraise the variability of:

- The crude Monte Carlo estimator $\bar{Y}$.
- The control variates estimator $\hat{Y}_{CV}(\boldsymbol{\lambda}_n)$, using the first moments of the X_i's as control variates.

- The conditional Monte Carlo estimator $\hat{Y}_{CMC}$. Because the durations of the three paths that lead from a to z are conditionally independent given X_1 and X_5, $E(Y|X_1, X_5)$ can be found analytically; see Loh (1995, p. 103).
- The weighted Monte Carlo estimator $\hat{Y}_{WMC}$, with $f(w) = -\log w$ in Equation (38).
- The stratification estimator $\hat{Y}_{rST}$, where we stratify on X_1.
- The Latin hypercube estimator $\hat{Y}_{LHS}$ applied to $X_1, \ldots, X_5$.

In order to compare the variance of these estimators, we repeat $m = 1000$ times the simulation to obtain $\bar{Y}(m), \ldots, \hat{Y}_{LHS}(m)$ by averaging $\bar{Y}, \ldots, \hat{Y}_{LHS}$ over m . The (sample) standard deviations of these six estimators are $s(m,n)$, $s_{CV}(m,n)$, $s_{CMC}(m,n)$, $s_{WMC}(m,n)$, $s_{rST}(m,n)$, and $s_{LHS}(m,n)$.

The results are summarized in Table 1. As expected, $s_{CV}(m,n) \approx s_{WMC}(m,n)$, and $s_{LHS}(m,n) < s_{rST}(m,n)$ for each n. Notice that $s_{rST}(m,n)$ and $s_{LHS}(m,n)$ behave like a constant divided by $n^{1/2}$ for each n, which indicates that rST and LHS achieve their variance reduction potential by $n = 100$.

11 Conclusions

We presented various variance reduction techniques for terminating simulations in the Hilbert space setting, establishing connections with CV, CMC, and WMC whenever possible. It is the geometric interpretation of Result 2 that makes this approach especially tractable.

There are, however, several topics missing from our coverage where Hilbert space theory might yield valuable insights. Consider for instance the case of variance reduction techniques for steady-state simulations that have a suitable martingale representation; see, for example, Henderson and Glynn (2002). It is

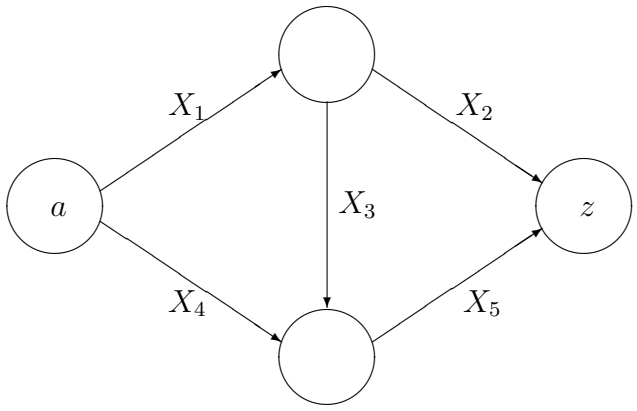

Fig. 3. Stochastic Activity Network

	Sample size n		
Parameter	100	1000	10000
$s(m,n)$	0.0529	0.0162	0.0053
$s_{CV}(m,n)$	0.0392	0.0112	0.0036
$s_{CMC}(m,n)$	0.0328	0.0102	0.0033
$s_{WMC}(m,n)$	0.0395	0.0116	0.0038
$s_{rST}(m,n)$	0.0456	0.0144	0.0044
$s_{LHS}(m,n)$	0.03	0.0093	0.0030

Table 1. SAN Numerical Example

well-known that square integrable martingale differences have a simple interpretation in the Hilbert space framework, which suggests that it might be possible to obtain additional insights when dealing with such techniques. Another area of interest is the Hilbert space formulation of CVs in the multi-response setting, where Y is a random vector; see Rubinstein and Marcus (1985) for relevant results. The combination of importance sampling (cf. Chapter 12) with CVs also can be studied in the Hilbert space setting; see, for example, Hesterberg (1995), and Owen and Zhou (1999).

Acknowledgements

I would like to thank the editors for their valuable comments and suggestions. This work was supported by NPS grant RIP BORY4.

References

Avellaneda, M., Buff, R., Friedman, C., Grandchamp, N., Kruk, L., Newman, J., 2001. Weighted Monte Carlo: A new technique for calibrating asset-pricing models. International Journal of Theoretical and Applied Finance 4, 1–29.

Avellaneda, M., Gamba, R., 2000. Conquering the Greeks in Monte Carlo: efficient calculation of the market sensitivities and hedge-ratios of financial assets by direct numerical simulation. Quantitative Analysis in Financial Markets, Vol. III. World Scientific, Singapore, pp. 336–356, M. Avellaneda (ed.).

Avramidis, A. N., Wilson, J. R., 1996. Integrated variance reduction techniques for simulation. Operations Research 44, 327–346.

Billingsley, P., 1995. Probability and Measure, 3rd Edition. John Wiley, New York.

Bollobas, B., 1990. Linear Analysis. Cambridge University Press, Cambridge.

Durrett, R., 1996. Probability: Theory and Examples, 2nd Edition. Duxbury Press, Belmont.

Fishman, G. S., 1996. Monte Carlo: Concepts, Algorithms, and Applications. Springer-Verlag, New York.

Glasserman, P., 2004. Monte Carlo Methods in Financial Engineering. Springer-Verlag, New York.

Glasserman, P., Heidelberger, P., Shahabuddin, P., 1999. Asymptotically optimal importance sampling and stratification for path-dependent options. Mathematical Finance 9, 117–152.

Glasserman, P., Yu, B., 2005. Large sample properties of weighted Monte Carlo. Operations Research 53 (2), 298–312.

Glynn, P. W., Szechtman, R., 2002. Some new perspectives on the method of control variates. Monte Carlo and Quasi-Monte Carlo Methods 2000. Springer-Verlag, Berlin, pp. 27–49, K. T. Fang, F. J. Hickernell, and H. Niederreiter (ed.).

Glynn, P. W., Whitt, W., 1989. Indirect estimation via $L = \lambda W$. Operations Research 37, 82–103.

Henderson, S., Glynn, P. W., 2002. Approximating martingales for variance reduction in Markov process simulation. Mathematics of Operations Research 27, 253–271.

Hesterberg, T. C., 1995. Weighted average importance sampling and defensive mixture distributions. Technometrics 37 (2), 185–194.

Hesterberg, T. C., Nelson, B. L., 1998. Control variates for probability and quantile estimation. Management Science 44, 1295–1312.

Jiang, T., 2003. Data driven shrinkage strategies for quasi-regression. Ph.D. thesis, Stanford University.

Kreyszig, E., 1978. Introductory Functional Analysis with Applications. John Wiley, New York.

Lavenberg, S. S., Moeller, T. L., Welch, P. D., 1982. Statistical results on control variates with application to queueing network simulation. Operations Research 30, 182–202.

Lavenberg, S. S., Welch, P. D., 1981. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. Management Science 27, 322–335.

Law, A. M., Kelton, W. D., 2000. Simulation Modeling and Analysis, 3rd Edition. McGraw-Hill, New York.

Loh, W. L., 1996. On Latin hypercube sampling. The Annals of Statistics 24 (5), 2058–2080.

Loh, W. W., 1995. On the method of control variates. Ph.D. thesis, Stanford University.

Mathé, P., 2000. Hilbert space analysis of Latin hypercube sampling. Proceedings of the American Mathematical Society 129 (5), 1477–1492.

McKay, M. D., Conover, W. J., Beckman, R. J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Nelson, B. L., 1990. Control variate remedies. Operations Research 38, 974–992.

Owen, A. B., 1992. A central limit theorem for Latin hypercube sampling. Journal of the Royal Statistical Society B 54 (2), 541–551.

Owen, A. B., 1997. Monte Carlo variance of scrambled equidistribution quadrature. SIAM Journal of Numerical Analysis 34 (5), 1884–1910.

Owen, A. B., Zhou, Y., 1999. Safe and effective importance sampling. Tech. rep., Stanford University.

Rubinstein, R. Y., Marcus, R., 1985. Efficiency of multivariate control variates in Monte Carlo simulation. Operations Research 33, 661–677.

Schmeiser, B. W., Taaffe, M. R., Wang, J., 2001. Biased control-variate estimation. IIE Transactions 33, 219–228.

Serfling, R. J., 1980. Approximation Theorems of Mathematical Statistics. John Wiley, New York.

Stein, M., 1987. Large sample properties of simulations using Latin hypercube sampling. Technometrics 29, 143–151.

Szechtman, R., Glynn, P. W., 2001. Constrained Monte Carlo and the method of control variates. In: Proceedings of the 2001 Winter Simulation Conference. Institute of Electrical and Electronic Engineers, Piscataway, NJ, pp. 394–400, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer (ed.).

Venkatraman, S., Wilson, J. R., 1986. The efficiency of control variates in multiresponse simulation. Operations Research Letters 5, 37–42.

Williams, D., 1991. Probability with Martingales. Cambridge University Press, Cambridge.

Wilson, J. R., 1984. Variance reduction techniques for digital simulation. American Journal of Mathematical and Management Sciences 4, 277–312.

Zimmer, R. J., 1990. Essential Results of Functional Analysis. The University of Chicago Press, Chicago.