

MONTE CARLO INTEGRATION WITH A GROWING NUMBER OF CONTROL VARIATES

FRANÇOIS PORTIER,* *Télécom Paris*

JOHAN SEGERS,** *UCLouvain*

Abstract

It is well known that Monte Carlo integration with variance reduction by means of control variates can be implemented by the ordinary least squares estimator for the intercept in a multiple linear regression model. A central limit theorem is established for the integration error if the number of control variates tends to infinity. The integration error is scaled by the standard deviation of the error term in the regression model. If the linear span of the control variates is dense in a function space that contains the integrand, the integration error tends to zero at a rate which is faster than the square root of the number of Monte Carlo replicates. Depending on the situation, increasing the number of control variates may or may not be computationally more efficient than increasing the Monte Carlo sample size.

Keywords: central limit theorem; control variates; multiple linear regression; ordinary least squares; post-stratification; Legendre polynomial

2010 Mathematics Subject Classification: Primary 60F05

Secondary 62J05;65C05

1. Introduction

Numerical integration algorithms can generally be characterized by (a) the *integration points* at which the integrand is evaluated and (b) the *integration weights* describing how to combine the evaluations of the integrand. Popular algorithms include the Riemann sums method, the Gaussian quadrature rule, and the classical Monte Carlo method. Those algorithms are usually compared by looking at the integration error for a given number, say $n \geq 1$, of integration points. Two types of methods can be distinguished. The ones that are based on deterministic integration points (e.g., equally spaced points) including the Riemann sums and the Gaussian quadrature, and the ones that generate randomly the integration points including the Monte Carlo method. The deterministic methods reach an accuracy of order $n^{-k/d}$ [21, Theorem 1], where k stands for the regularity of the integrand and d is the dimension of the integration domain, whereas random methods are subjected to an optimal error bound of order $n^{-k/d}n^{-1/2}$ [21, Theorem 3]. For instance, the naive Monte Carlo method, which does not use any

* Postal address: LTCI, Télécom Paris, Institut polytechnique de Paris, rue Barrault, 75013 Paris, France. Email: francois.portier@gmail.com

** Postal address: Institut de statistique, biostatistique et sciences actuariales, LIDAM, UCLouvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium. Email: johan.segers@uclouvain.be

regularity of the integrand, converges at the rate $n^{-1/2}$. Those error bounds testify to the benefits of random methods over deterministic ones especially when facing high-dimensional settings.

The method of control variates is a popular technique in Monte Carlo integration that aims at reducing the variance of the naive Monte Carlo estimate by taking advantage of the regularity of the integrand [8, 10, 26, 29]. It is based on the introduction of auxiliary functions, called control variates, with known integral. Given a fixed number of control variates, the method consists in (i) fitting a linear combination of the control variates to the integrand, and (ii) using the fitted function in a modified Monte Carlo procedure. As noted in [24], the fit of the integrand in step (i) generally uses the integration points, which makes the control variate method a post-hoc scheme, i.e., that might be done after sampling the integration points. The control variate approach is quite general as it allows to recover several famous examples from numerical integration. In dimension 1, the Newton–Cotes rule with random interpolation points can be recovered by taking the polynomials of degree smaller than $n - 1$ as control variates. The post-stratification method can also be recovered by combining the indicators of a given partition of the integration domain (see Example 1 in Section 5 below). When no control variates are used, it coincides with the classical Monte Carlo method.

An important field of application of the control variates method is financial engineering where it has been used for Asian option pricing in the Black–Scholes model [8, Example 4.1.2] or to solve backward stochastic differential equations [11]. More recently, it has been helpful in reinforcement learning to accelerate the estimation of the optimal policy [14]. As highlighted in the present paper the method is efficient when many integrals need to be computed. This is the case for instance in quantile estimation [12], option pricing [9], and likelihood computation in statistical models with latent variables [28], which arise frequently in economics [19] and medicine [18, Examples 4, 6 and 9]. Finally, note that using importance sampling permits to recover the Lebesgue measure as the reference measure [25, Theorem 2] which in turn allows the use of many control variates such as polynomials, indicators, splines and Gaussian mixtures.

As illustrated by the well-known Runge phenomenon in approximation theory, enlarging the number of control variates does not necessarily improve the method. A key question then, which will be central in the paper, is related to the number of control variates that should be used in the procedure. The possibility of letting the number of control functions tend to infinity is already alluded to in [10, Theorem 3], who show that, for control functions arising as the power sequence of a given function, the variance of the limiting normal distribution of the error of the control variate method converges to the variance of the residual of the conditional expectation of the integrand given the initial control function. However, this result is still cast within the setting of a fixed number of control variates, i.e., the number of control variates does not depend on the Monte Carlo sample size. A recent proposal in [24] is to construct the linear fit to the integrand in step (i) above as an element of a reproducing kernel Hilbert space, whose dimension grows with the sample size n . Their approach leads to a convergence rate that is at least as fast as $n^{-7/12}$ and thus improves over the Monte Carlo rate. Further

refinements are given in [22], with tighter error bounds depending on the smoothness of the integrand.

In this paper, we adopt the original control variate framework but allow the number of control variates $m = m_n$ to grow with n . Among the six control variate estimators in [10], only one possesses the property of integrating the constants and the control functions without error. This is the one we promote and study in this paper. We use the denomination ordinary least squares Monte Carlo (OLSMC) because of the well-known link [8] with the ordinary least squares estimator for the intercept in a multiple linear regression model with the integrand as dependent variable and the control variates as explanatory variables.

Our main result is that when $m_n \rightarrow \infty$ but $m_n = o(n^{1/2})$ and under reasonable conditions on the control functions and the integrand, the OLSMC estimator obeys a central limit theorem with the non-standard rate $n^{-1/2}\sigma_n$, where σ_n is the standard deviation of the error term in the aforementioned multiple regression model. Moreover, we show that the common estimator $\hat{\sigma}_n$ of the standard deviation defined via the residual sum of squares is consistent in the sense that $\hat{\sigma}_n/\sigma_n \rightarrow 1$ in probability. This fact guarantees the asymptotic coverage of the usual confidence intervals.

If $\sigma_n \rightarrow 0$, then the convergence rate of the OLSMC is faster than the $n^{-1/2}$ rate of the ordinary Monte Carlo procedure. Still, this acceleration is offset by an increased computational cost, from $O(n)$ operations for ordinary Monte Carlo to $O(nm_n^2)$ for the control variate method, a number which can be brought down to $O(nm_n)$ in certain situations. A more balanced comparison arises when we allow the naive Monte Carlo method to compete on the basis of a larger sample size, matching computation times. Whether or not the investment in m_n control variates is worth the effort then depends on the exact speed at which σ_n tends to zero, as is illustrated by examples.

In Section 2, we recall the method of control variates, highlighting a formulation in terms of projections which is useful later on. A central limit theorem when the number of control variates tends to infinity is developed in Section 3. Its formulation allows for a sequence of integrands and for a triangular array of control variates. The balance between accelerated convergence rate and increased computational cost is investigated in Section 4. Examples of families of control functions are presented in Section 5 while some concluding comments are given in Section 6. All proofs are relegated to Section 7.

2. Control variates and orthogonal projections

2.1. Control variates

Let (S, \mathcal{S}, P) be a probability space and let $f \in L^2(P)$ be a real function on S of which we would like to calculate the integral $\mu = P(f) = \int_S f(x) P(dx)$. Let X_1, \dots, X_n be an independent random sample from P on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let P_n be its empirical distribution. The Monte Carlo estimate of μ is $\hat{\mu}_n = P_n(f) = n^{-1} \sum_{i=1}^n f(X_i)$. The Monte Carlo estimator is unbiased and has variance $\text{var}(\hat{\mu}_n) = n^{-1}\sigma^2(f)$, where $\sigma^2(f) = P[\{f - P(f)\}^2]$. By the central limit theorem, $\sqrt{n}(\hat{\mu}_n - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2(f))$ as $n \rightarrow \infty$, where the arrow \rightsquigarrow denotes convergence in distribution.

The use of control variates is one of many methods to reduce the asymptotic variance

of the Monte Carlo estimator. Let $h_1, \dots, h_m \in L^2(P)$ be functions with known expectations. Without loss of generality, assume that $P(h_j) = 0$ for all $j = 1, \dots, m$. For every column vector $\beta \in \mathbb{R}^m$, we obviously have $\mu = P(f - \beta'h)$, where $h = (h_1, \dots, h_m)'$ is the column vector with the m control functions as elements. But then $\hat{\mu}_n(\beta) = P_n(f - \beta'h)$ is an unbiased estimator of μ too, with variance $\text{var}\{\hat{\mu}_n(\beta)\} = n^{-1}\sigma^2(f - \beta'h)$.

The asymptotic variance $\sigma^2(f - \beta'h)$ is minimal if β is equal to

$$\beta_{\text{opt}} = P(hh')^{-1} P(hf). \quad (1)$$

Here we assume that the functions h_1, \dots, h_m are linearly independent in $L^2(P)$, so that the $m \times m$ covariance matrix $P(hh') = (P(h_j h_k))_{j,k=1}^m$ is invertible. The minimal asymptotic variance is

$$\sigma^2(f - \beta'_{\text{opt}} h) = \sigma^2(f) - P(fh') P(hh')^{-1} P(hf). \quad (2)$$

In practice, β_{opt} in (1) is unknown and needs to be estimated. Any estimator $\hat{\beta}_n$ of β_{opt} produces a control variate estimator: $\hat{\mu}_n(\hat{\beta}_n) = P_n(f - \hat{\beta}'_n h)$. As soon as $\hat{\beta}_n \rightsquigarrow \beta_{\text{opt}}$, then [10, Theorem 1],

$$\sqrt{n}\{\hat{\mu}_n(\hat{\beta}_n) - \mu\} \rightsquigarrow \mathcal{N}(0, \sigma^2(f - \beta'_{\text{opt}} h)), \quad n \rightarrow \infty. \quad (3)$$

It is thus sufficient to estimate the vector β_{opt} consistently to obtain an integration procedure with the same asymptotic distribution as the oracle procedure $\hat{\mu}_n(\beta_{\text{opt}})$.

The asymptotic variance in (2) may be estimated by the empirical variance

$$\hat{\sigma}_n^2(\hat{\beta}_n) = P_n[\{f - \hat{\beta}'_n h\}^2] - \{P_n[f - \hat{\beta}'_n h]\}^2.$$

If $\hat{\beta}_n \rightsquigarrow \beta_{\text{opt}}$, then, by the law of large numbers and Slutsky's lemma,

$$\hat{\sigma}_n^2(\hat{\beta}_n) \rightsquigarrow \sigma^2(f - \beta'_{\text{opt}} h), \quad n \rightarrow \infty. \quad (4)$$

Equations (3) and (4) justify the usual asymptotic confidence intervals for μ .

2.2. Ordinary least squares estimator

To estimate $\beta_{\text{opt}} = P(hh')^{-1} P(hf)$, multiple options exist [10]. The more common estimator is

$$\hat{\beta}_n^{\text{OLS}} = G_n^{-1} \{P_n(hf) - P_n(h) P_n(f)\},$$

where $G_n = P_n(hh') - P_n(h) P_n(h')$ is the empirical covariance matrix of the control variates, assumed to be invertible, which is the case with large probability under the conditions in Section 3. The resulting ordinary least squares Monte Carlo estimator is

$$\hat{\mu}_n^{\text{OLS}} = \hat{\mu}_n(\hat{\beta}_n^{\text{OLS}}) = P_n(f) - \{P_n(fh') - P_n(f) P_n(h')\} G_n^{-1} P_n(h).$$

The OLSMC variance estimator is equal to the sample analogue of (2):

$$\begin{aligned} \hat{\sigma}_{n,\text{OLS}}^2 &= \hat{\sigma}_n^2(\hat{\beta}_n^{\text{OLS}}) \\ &= P_n[\{f - P_n(f)\}^2] - \{P_n(fh') - P_n(f) P_n(h')\} G_n^{-1} \{P_n(hf) - P_n(h) P_n(f)\}. \end{aligned}$$

The terminology stems from the well-known [8] property that

$$(\hat{\mu}_n^{\text{OLS}}, \hat{\beta}_n^{\text{OLS}}) = \arg \min_{(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^m} \sum_{i=1}^n \{f(X_i) - \alpha - \beta' h(X_i)\}^2. \quad (5)$$

The identity (5) is a consequence of the normal equations in the multiple linear regression model

$$f(X_i) = \mu + \beta'_{\text{opt}} h(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

with dependent variable $f(X_i)$, explanatory variables $h_1(X_i), \dots, h_m(X_i)$, and errors ε_i . The intercept is μ whereas the vector of regression coefficients is β_{opt} . The errors are $\varepsilon_i = \varepsilon(X_i)$ with $\varepsilon = f - \mu - \beta'_{\text{opt}} h \in L^2(P)$, a mean-zero function which is uncorrelated with each of the control functions, i.e., $P(\varepsilon) = 0$ and $P(h\varepsilon) = 0$. The variance of the errors is equal to the asymptotic variance of the OLSMC estimator: $P(\varepsilon^2) = \sigma^2(f - \beta'_{\text{opt}} h)$.

Equation (5) has the convenient consequence that the OLSMC estimator and the variance estimator can be computed via standard linear regression software [26, Section 8.9]. Also, it implies that the OLSMC integration rule integrates the constant function and the m control functions exactly.

2.3. Orthogonal projections

Geometric considerations lead to another, insightful representation of the OLSMC estimator, revealing properties relevant for asymptotic theory. Let $H^{(n)}$ be the $n \times m$ matrix

$$H^{(n)} = \begin{pmatrix} h_1(X_1) & \dots & h_m(X_1) \\ \vdots & & \vdots \\ h_1(X_n) & \dots & h_m(X_n) \end{pmatrix}. \quad (6)$$

Let $\Pi_{n,m}$ be the $n \times n$ projection matrix on the column space of the matrix $H^{(n)}$ in (6). If the m columns of $H^{(n)}$ are linearly independent, then

$$\Pi_{n,m} = H^{(n)} \{ (H^{(n)})' H^{(n)} \}^{-1} (H^{(n)})' = n^{-1} H^{(n)} P_n (h h')^{-1} (H^{(n)})', \quad (7)$$

the so-called hat matrix in a multiple linear regression model without intercept on the m variables $(h_j(X_i))_{i=1}^n$, $j = 1, \dots, m$. Even if the m columns of $H^{(n)}$ are not linearly independent, the projection matrix $\Pi_{n,m}$ is well-defined, for instance, by using Moore–Penrose inverses.

Write the OLSMC estimator in (5) in terms of two nested minimization problems:

$$\hat{\mu}_n^{\text{OLS}} = \arg \min_{\alpha \in \mathbb{R}} \left[\min_{\beta \in \mathbb{R}^m} \sum_{i=1}^n \{f(X_i) - \alpha - \beta' h(X_i)\}^2 \right].$$

Given $\alpha \in \mathbb{R}$, the minimum over $\beta \in \mathbb{R}^m$ is well-defined and is attained as soon as β satisfies $H^{(n)} \beta = \Pi_{n,m} (f^{(n)} - \alpha 1_n)$, where $f^{(n)} = (f(X_1), \dots, f(X_n))'$ and where 1_n is an $n \times 1$ vector with all elements equal to 1. We find that

$$\hat{\mu}_n^{\text{OLS}} = \arg \min_{\alpha \in \mathbb{R}} |(I_n - \Pi_{n,m})(f^{(n)} - \alpha 1_n)|^2, \quad (8)$$

where $|v| = (v'v)^{1/2}$ is the Euclidean norm of a vector v and I_n is the $n \times n$ identity matrix. It follows that $\alpha(I_n - \Pi_{n,m})1_n$ is equal to the orthogonal projection of $(I_n - \Pi_{n,m})f^{(n)}$ on the line passing through the origin and $(I_n - \Pi_{n,m})1_n$. A necessary and sufficient condition for the uniqueness of $\alpha \in \mathbb{R}$ is that $(I_n - \Pi_{n,m})1_n$ is not equal to the zero vector, that is, 1_n is *not* an element of the column space of $H^{(n)}$. Suppose this condition holds. Then $1_n'(I_n - \Pi_{n,m})1_n = |(I_n - \Pi_{n,m})1_n|^2 > 0$ and

$$\hat{\mu}_n^{\text{OLS}} = \frac{(f^{(n)})'(I_n - \Pi_{n,m})1_n}{1_n'(I_n - \Pi_{n,m})1_n}. \quad (9)$$

If, in addition, the columns of $H^{(n)}$ are linearly independent, then, by (7),

$$\hat{\mu}_n^{\text{OLS}} = \frac{P_n(f) - P_n(fh') P_n(hh')^{-1} P_n(h)}{1 - P_n(h') P_n(hh')^{-1} P_n(h)}. \quad (10)$$

Indeed, we have $(f^{(n)})'1_n = n P_n(f)$, $(f^{(n)})'H^{(n)} = n P_n(fh')$, and $1_n'H^{(n)} = n P_n(h')$.

We have supposed that the $n \times 1$ vector 1_n is *not* an element of the column space of $H^{(n)}$. If it is, then there obviously cannot exist a weight vector such that the corresponding linear integration rule integrates both the constant functions and the control functions exactly. Also, the minimizer α in (5) is then no longer identifiable. In that case, we recommend to reduce the number of control functions. Actually, when m is not too large with respect to n (Section 3), the denominator in (10) tends to 1 in probability, implying that, with probability tending to one, 1_n is *not* an element of the column space of $H^{(n)}$.

The representation (9) also implies that the OLSMC estimator does not change if we replace the vector h of control functions by the vector Ah , where A is an arbitrary invertible $m \times m$ matrix. Indeed, such a transformation results in changing the matrix $H^{(n)}$ in (6) into $H^{(n)}A'$, but both $n \times m$ matrices share the same column space.

The OLSMC variance estimator $\hat{\sigma}_{n,\text{OLS}}^2$ coincides with n^{-1} times the minimal sum of squares in (5) and (8):

$$\hat{\sigma}_{n,\text{OLS}}^2 = \frac{1}{n} (f^{(n)} - \hat{\mu}_n^{\text{OLS}} 1_n)' (I_n - \Pi_{n,m}) (f^{(n)} - \hat{\mu}_n^{\text{OLS}} 1_n). \quad (11)$$

Recall $f = \mu + \beta'_{\text{opt}} h + \varepsilon$, where $\varepsilon \in L^2(P)$ is centered and uncorrelated with all control functions h_j . If $P_n(hh')$ is invertible and $P_n(h') P_n(hh')^{-1} P_n(h) < 1$, we can use (7) for $\Pi_{n,m}$ and (10) for $\hat{\mu}_n^{\text{OLS}}$ to work out (11) and find (proof in Section 7)

$$\begin{aligned} \hat{\sigma}_{n,\text{OLS}}^2 &= P_n(\varepsilon^2) - P_n(\varepsilon h') P_n(hh')^{-1} P_n(h\varepsilon) \\ &\quad - (\hat{\mu}_n^{\text{OLS}} - \mu)^2 \{1 - P_n(h') P_n(hh')^{-1} P_n(h)\}. \end{aligned} \quad (12)$$

Since $\hat{\sigma}_{n,\text{OLS}}^2 \leq P_n(\varepsilon^2)$ and $\mathbb{E}\{P_n(\varepsilon^2)\} = \sigma^2$, it follows that $\hat{\sigma}_{n,\text{OLS}}^2$ has a negative bias. In view of the multiple linear regression perspective in Section 2.2 and to possibly reduce this bias, one may prefer to multiply the variance estimator by $n/(n - m - 1)$, although this particular correction is justified only in case of a linear model with fixed design and centered, uncorrelated, and homoskedastic Gaussian errors.

3. Central limit theorem for a growing number of control variates

By (3), the asymptotic variance of the OLSMC estimator $\hat{\mu}_n^{\text{OLS}}$ of $\mu = P(f)$ with a fixed number of control variates is equal to the variance of the error variable

$$\varepsilon = f - \mu - \beta'_{\text{opt}} h, \quad (13)$$

where $\mu + \beta'_{\text{opt}} h$ is the orthogonal projection in $L^2(P)$ of f on the linear space \mathcal{F}_m spanned by $\{1, h_1, \dots, h_m\}$. Suppose that the number, $m = m_n$, of control functions varies with n and tends to infinity and that f can be written as an $L^2(P)$ limit of a sequence of approximating functions in \mathcal{F}_{m_n} . Then $\sigma_n^2 = P(\varepsilon_n^2) \rightarrow 0$ as $n \rightarrow \infty$, where ε_n is the error variable ε in (13) when there are m_n control variates in use. Then we may hope that the asymptotic variance of the OLSMC estimator becomes zero too, so that its convergence rate is $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$, faster than the one of the Monte Carlo estimator. More precisely, we may hope to pin the convergence rate down to $\mathcal{O}_{\mathbb{P}}(\sigma_n/\sqrt{n})$.

3.1. Set-up

Our set-up is a triangular array of control functions. Let $h_n = (h_{n,1}, \dots, h_{n,m_n})'$ for some positive integer sequence $m_n \rightarrow \infty$, where $h_{n,j} \in L^2(P)$ and $P(h_{n,j}) = 0$ for all n and j . Assume that $h_{n,1}, \dots, h_{n,m_n}$ are linearly independent in $L^2(P)$, so that the $m_n \times m_n$ Gram matrix $P(h_n h_n') = (P(h_{n,j} h_{n,k}))_{j,k}$ is invertible. Examples of control functions we have in mind are polynomials or trigonometric functions, in which case a single sequence h_1, h_2, \dots would suffice, or spline functions on an interval with the knots forming a grid depending on m_n , an example which requires a triangular array of control functions.

There is no additional mathematical cost to let the integrands depend on n as well: we want to calculate the integral $\mu_n = P(f_n)$ of $f_n \in L^2(P)$. In doing so, we obtain results that are locally uniform in the integrand. We have $f_n = \mu_n + \beta_n' h_n + \varepsilon_n$ for some vector $\beta_n \in \mathbb{R}^{m_n}$ determined by the orthogonality equations $P(\varepsilon_n h_{n,j}) = 0$ for all $j = 1, \dots, m_n$. We have $P(\varepsilon_n) = 0$, while the error variance is $\sigma_n^2 = P(\varepsilon_n^2)$. To avoid trivialities, we assume that $\sigma_n^2 > 0$, that is, f_n is not equal to a constant plus a linear combination of the control functions, in which case its integral would be known. Of particular interest is the case where $\sigma_n^2 \rightarrow 0$ as $n \rightarrow \infty$, although we do not impose this.

3.2. Leverage condition

Consider the linear regression model without intercept term for the centered integrand on the control variates:

$$f_n(X_i) - \mu_n = h_n'(X_i) \beta_n + \varepsilon_n(X_i), \quad i = 1, \dots, n.$$

The $n \times m_n$ design matrix is $H^{(n)}$ in (6), whereas the $n \times n$ projection matrix onto the column space of $H^{(n)}$ is $\Pi_n \equiv \Pi_{n,m_n}$ in (7), assuming that the m_n columns of $H^{(n)}$ are linearly independent. In multiple linear regression theory, this projection matrix is called the *hat matrix*, and its i th diagonal element is called the *leverage* of the i th sample point:

$$\Pi_{n,ii} = n^{-1} h_n(X_i)' P_n(h_n h_n')^{-1} h_n(X_i), \quad i = 1, \dots, n.$$

The average leverage is equal to $n^{-1} \text{tr}(\Pi_n) = m_n/n$. Points for which $\Pi_{n,ii} > cm_n/n$ for some pre-determined constant $c > 1$, often $c = 2$ or $c = 3$, are commonly flagged as high-leverage points; see [31] and the references therein.

We have $\Pi_{n,ii} = n^{-1} \hat{q}_n(X_i)$ for $i = 1, \dots, n$, where $\hat{q}_n(x) = h_n(x)' P_n(h_n h_n')^{-1} h_n(x)$ is the sample version of what could be called the *leverage function*

$$q_n(x) = h_n(x)' P(h_n h_n')^{-1} h_n(x), \quad x \in S. \quad (14)$$

Note that $q_n(x)$ is the squared Mahalanobis distance of $h_n(x)$ to the center $P(h_n) = 0$ of the distribution of the m_n -dimensional random vector h_n under P . The expectation of the leverage function is equal to the dimension of the control space,

$$P(q_n) = m_n. \quad (15)$$

Recall that the OLSMC estimator does not change if we replace the vector h_n by the vector Ah_n , where A is any invertible $m_n \times m_n$ matrix. The function q_n is invariant under such transformations of the control functions, as can be easily checked. It follows that q_n is linked to the linear space spanned by the control functions $h_{n,1}, \dots, h_{n,m_n}$ rather than to the functions themselves.

To establish the rate of convergence of the OLSMC estimator, we need to prohibit the occurrence of points of which the leverage is too high. The criterion commonly used in regression diagnostics to flag high-leverage points would suggest that we impose that $\sup_{x \in S} q_n(x) = O(m_n)$ as $n \rightarrow \infty$. [By (15), a smaller bound can never be satisfied.] Instead, we impose a weaker condition, which is reminiscent of Assumption 2(ii) in [20].

Condition 1. (Leverage.) *We have*

$$\sup_{x \in S} q_n(x) = o(n/m_n), \quad n \rightarrow \infty. \quad (16)$$

Equations (15) and (16) imply

$$P(q_n^2) = o(n), \quad n \rightarrow \infty. \quad (17)$$

Since $m_n^2 = P(q_n)^2 \leq P(q_n^2)$, Equation (17) implies that $m_n = o(n^{1/2})$, restricting the dimension of the control space. As a consequence, also $m_n = o(n/m_n)$, meaning that Equation (16) is indeed weaker than $\sup_{x \in S} q_n(x) = O(m_n)$ as $n \rightarrow \infty$.

According to [13], the reciprocal of the leverage can be seen as the equivalent number of observations entering into the determination of the predicted response for the i th point. Since our condition implies that $\sup_{x \in S} n^{-1} q_n(x) = o(1/m_n)$ as $n \rightarrow \infty$, a possible interpretation of Condition 1 is that the equivalent number of observations used to predict each response is of larger order than the number of control variates, m_n .

3.3. Main results

Assume the set-up of Section 3.1.

Theorem 1. (Rate.) *If Condition 1 holds, then, as $n \rightarrow \infty$, the OLSMC estimator is well-defined with probability tending to one and*

$$\frac{\sqrt{n}}{\sigma_n} (\hat{\mu}_n^{\text{OLS}} - \mu_n) = \frac{\sqrt{n}}{\sigma_n} P_n(\varepsilon_n) + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1). \quad (18)$$

In particular, $\hat{\mu}_n^{\text{OLS}} - \mu_n = O_{\mathbb{P}}(\sigma_n/\sqrt{n})$ as $n \rightarrow \infty$.

To prove asymptotic normality of the estimation error, we apply the Lindeberg–Feller central limit theorem. The Lindeberg condition, which is both necessary and sufficient [15, Theorem 5.12], also guarantees consistency of the OLS variance estimator. A sufficient but not necessary condition as well as some intuition are provided in Remark 3 below. Recall that the arrow \rightsquigarrow denotes weak convergence.

Condition 2. (Lindeberg.) *For every $\delta > 0$, we have, as $n \rightarrow \infty$,*

$$P[(\varepsilon_n/\sigma_n)^2 \mathbf{1}\{|\varepsilon_n/\sigma_n| > \delta\sqrt{n}\}] = o(1).$$

Theorem 2. (Asymptotic normality.) *Suppose Condition 1 holds. Then Condition 2 holds if and only if*

$$\frac{\sqrt{n}}{\sigma_n} (\hat{\mu}_n^{\text{OLS}} - \mu_n) \rightsquigarrow \mathcal{N}(0, 1), \quad n \rightarrow \infty. \quad (19)$$

Moreover, under Conditions 1 and 2, the variance estimator is consistent in the sense that

$$\hat{\sigma}_{n,\text{OLS}}^2/\sigma_n^2 \rightsquigarrow 1, \quad n \rightarrow \infty. \quad (20)$$

Equation (19) thus remains true if σ_n is replaced by $\hat{\sigma}_{n,\text{OLS}}$.

Theorem 2 justifies the use of the usual asymptotic confidence intervals of nominal coverage $1 - \alpha$ of the form $\hat{\mu}_{n,\text{OLS}} \pm z_{1-\alpha/2} \hat{\sigma}_{n,\text{OLS}}/\sqrt{n}$, where z_p is the p th quantile of the standard normal distribution. As in multiple linear regression, quantiles of the Student t distribution with $n - m_n - 1$ degrees of freedom may be used instead, making the intervals a bit wider, although there is no guarantee that this will bring the real coverage closer to the nominal one when the errors are not normally distributed.

3.4. Discussion

Remark 1. (*Weakening the leverage condition.*) Equation (16) implies

$$P(q_n \varepsilon_n^2) = o\{(n/m_n)\sigma_n^2\}, \quad n \rightarrow \infty. \quad (21)$$

In fact, Theorems 1 and 2 would remain true if Condition 1 would be replaced by the weaker pair of Equations (17) and (21). In addition, by the Cauchy–Schwarz inequality, $P(q_n \varepsilon_n^2) \leq P(q_n^2)^{1/2} P(\varepsilon_n^4)^{1/2}$, so that Equation (17) together with

$$P(\varepsilon_n^4) = O\{(n/m_n^2)\sigma_n^4\}, \quad n \rightarrow \infty, \quad (22)$$

would be sufficient. However, both Equations (21) and (22) depend on the integrand through the error function ε_n and may be difficult to check. The advantage of Equation (16) is that it only depends on the control variates and not on the integrand.

Remark 2. (*Checking the leverage condition.*) When calculating q_n is complicated, the following bound may be helpful in establishing (16): we have

$$q_n \leq \lambda_{n,1}^{-1} h'_n h_n = \lambda_{n,1}^{-1} \sum_{j=1}^{m_n} h_{n,j}^2,$$

where $\lambda_{n,1} > 0$ is the smallest eigenvalue of $P(h_n h'_n)$. See also Section 5 for a number of examples in which we check the leverage condition.

Remark 3. (*On the Lindeberg condition.*) As already mentioned, Condition 2 is both necessary and sufficient for (19) to hold. In view of Hölder's inequality, the condition is implied by the Lyapunov condition that there exists $\eta > 0$ such that

$$\sup_{n \geq 1} P[|\varepsilon_n / \sigma_n|^{2+\eta}] < \infty.$$

The latter condition is equivalent to $\|\varepsilon_n\|_{2+\eta} = O(\|\varepsilon_n\|_2)$ as $n \rightarrow \infty$, where $\|\cdot\|_p$ denotes the $L_p(P)$ (semi-)norm.

Intuitively, the Lindeberg condition requires that the error sequence ε_n behaves regularly in some sense. It fails for instance if, along a subsequence, the centered integrand $f_n - P(f_n)$ is a linear combination of the control functions $h_{n,1}, \dots, h_{n,m_n}$: if the fit is perfect ($\sigma_n = 0$), the integration error is zero and cannot be normalized to be asymptotically standard Gaussian. See Example 2 in Section 5 below for an illustration on checking the Lindeberg condition.

4. Computational cost

For pure Monte Carlo integration, the main computational cost stems from the n evaluations of the integrand f . The computation time is therefore of the order $O(n)$. Decreasing the integration error then simply amounts to increase the number, n , of random evaluation points X_i .

Another way to improve the integration accuracy is by increasing the number of control variates. For fixed sample size n , this will decrease the standard deviation $\sigma_n = \{P(\varepsilon_n^2)\}^{1/2}$ of the error term $\varepsilon_n \in L^2(P)$ in the representation

$$f = \mu + \beta_{n,1} h_{n,1} + \dots + \beta_{n,m_n} h_{n,m_n} + \varepsilon_n,$$

with $\beta_n \in \mathbb{R}^{m_n}$ determined by $P(\varepsilon_n) = 0$ and $P(\varepsilon_n h_{n,j}) = 0$ for all $j = 1, \dots, m_n$.

However, the use of m_n control variates makes the number of operations go up to $O(nm_n^2)$. The bottleneck comes from the $m_n \times m_n$ empirical Gram matrix $P_n(h_n h'_n)$, each element of which requires calculating an arithmetic mean over the n sample points. The other terms in (10) require fewer operations. Indeed, evaluating the m_n control variates $h_{n,j}$ in the n sample points X_i amounts to $O(nm_n)$ operations. The vectors $P_n(h_n)$ and $P_n(fh'_n)$ contain m_n elements, each of which is an arithmetic mean over the Monte Carlo sample, requiring $O(nm_n)$ operations too. The matrix inversion and matrix multiplication in (10) represent $O(m_n^3)$ operations. Since necessarily $m_n^2 = o(n)$ by (15) and (17), the latter represents an additional cost of only $o(nm_n)$ operations.

The method of control variates thus invests $O(nm_n^2)$ operations to achieve an asymptotic standard deviation of $\sigma_n n^{-1/2}$. Alternatively, one could allocate all computation resources to augmenting the Monte Carlo sample size from n to nm_n^2 , yielding a standard deviation of the order $O(n^{-1/2}m_n^{-1})$. At equal computational budget, the method of control variates with the number of control variates tending to infinity will thus converge at a faster rate than naive Monte Carlo integration as soon as

$$\sigma_n = o(m_n^{-1}), \quad n \rightarrow \infty. \quad (23)$$

Whether or not this is the case depends on the control variates and the integrand; see the examples in Section 5.

For certain families of control variates, the computational cost can be brought down from $O(nm_n^2)$ to $O(nm_n)$. This is the case for instance for the normalized indicator functions in Example 1 below and more generally for control variates that arise from functions that, prior to centering, have localized supports, such as splines or wavelets. In such cases, only $O(m_n)$ elements of the $m_n \times m_n$ matrix $P_n(h_n h_n')$ vary with the sample, while the other elements are known in advance and thus non-random. Comparing the asymptotic standard deviation $\sigma_n n^{-1/2}$ of the control variate error with the one of the naive Monte Carlo method at sample size nm_n , which is $O(n^{-1/2}m_n^{-1/2})$, we find that, at equal computational budget, the OLSMC estimator already converges at a faster rate than the Monte Carlo estimator as soon as

$$\sigma_n = o(m_n^{-1/2}), \quad n \rightarrow \infty. \quad (24)$$

If evaluating the integrand f is expensive while evaluating the control functions h_j is cheap, then, in practice, it may still be computationally beneficial to increase the number of control variates rather than the Monte Carlo sample size, even though this is not backed up by the asymptotic considerations so far.

Computational benefits can also occur when there are multiple integrands. Indeed, it is well known that the method of control variates can be seen as a form of weighted Monte Carlo, i.e.,

$$\hat{\mu}_n^{\text{OLS}} = \sum_{i=1}^n w_{n,i} f(X_i)$$

where the expression of the weight vector $w_n \in \mathbb{R}^n$ can for instance be deduced from (9); see also [8, eq. (4.20)]. The control variates only enter the formula through these weights, which, even in case of multiple integrands, thus need to be computed only once. This feature can for instance be put to work to efficiently estimate quantiles [12], price financial options [9], and compute likelihoods arising in statistical models with latent variables [28].

5. Examples

Example 1. (*Post-stratification.*) On $S = [0, 1]$ equipped with the Lebesgue measure P , let $h_{n,j}(x) = (m_n + 1)\mathbb{1}\{x \in \mathcal{I}_{m_n,j}\} - 1$ for $j = 1, \dots, m_n$, where $\mathcal{I}_{m_n,j} = [(j-1)/(m_n+1), j/(m_n+1))$. The control variates are normalized indicator functions induced by a partition of $[0, 1]$ into $m_n + 1$ intervals of equal length. Note that the last cell

$\mathcal{I}_{m_n, m_n+1} = [m_n/(m_n + 1), 1]$ is omitted, since its normalized indicator h_{n, m_n+1} is a linear combination of $h_{n, 1}, \dots, h_{n, m_n}$.

Unless one or more cells contain no sample points X_i , the constant vector 1_n is not an element of the column space of the design matrix in (6) and the OLSMC estimator is well-defined. A particular cell being empty with probability $\{1 - (m_n + 1)^{-1}\}^n$, the probability that at least one cell is empty is bounded by $(m_n + 1)(1 - (m_n + 1)^{-1})^n$, which converges to zero as soon as $m_n \ln(m_n) = o(n)$.

The Gram matrix $P(h_n h_n') = (m_n + 1)I_{m_n} - 1_{m_n} 1_{m_n}'$ has inverse $P(h_n h_n')^{-1} = (m_n + 1)^{-1}(I_{m_n} + 1_{m_n} 1_{m_n}')$. The function $q_n = h_n' P(h_n h_n')^{-1} h_n = \{h_n' h_n + (h_n' 1_{m_n})^2\}/(m_n + 1) = m_n$ is constant. Condition 1 is satisfied as soon as $m_n = o(n^{1/2})$.

Let $f_{m_n, j} = (m_n + 1)^{-1} P(f \mathbb{1}\{\cdot \in \mathcal{I}_{m_n, j}\})$ be the average of the integrand f on the cell $\mathcal{I}_{m_n, j}$, for $j = 1, \dots, m_n + 1$. The OLSMC estimator is equal to the arithmetic mean of the Monte Carlo estimates of these $m_n + 1$ local averages $f_{m_n, j}$. This is also the value obtained by post-stratification [26, Example 8.4]. The number of operations required to calculate the OLSMC estimator is thus $O(nm_n)$ only.

The projection of f on the space spanned by $\{1, h_{n, 1}, \dots, h_{n, m_n}\}$ is equal to the piecewise constant function $f^{(n)}$ with value $f_{m_n, j}$ on $\mathcal{I}_{m_n, j}$ for $j = 1, \dots, m_n + 1$. If f is Lipschitz, then the error term $\varepsilon_n = f - f^{(n)}$ will satisfy $\sup_{x \in S} |\varepsilon_n(x)| = O(m_n^{-1})$. In particular, $\sigma_n = O(m_n^{-1})$. If, in addition, $\liminf_{n \rightarrow \infty} \sigma_n m_n > 0$, then ε_n/σ_n remains bounded uniformly, and the Lindeberg condition (Condition 2) is satisfied too.

The standard deviation of the OLSMC error at sample size n is $\sigma_n n^{-1/2}$, achieved at $O(nm_n)$ operations, while the one of the Monte Carlo integration error at sample size nm_n is $n^{-1/2} m_n^{-1/2}$. For Lipschitz functions, we have $\sigma_n = O(m_n^{-1}) = o(m_n^{-1/2})$, as in (24). At comparable computational budgets, the OLSMC estimator thus achieves a faster rate of convergence than the Monte Carlo estimator.

On the d -dimensional cube $S = [0, 1]^d$, we can employ a similar construction, starting from a partition of S into $O(m_n)$ cubes with side length $O(m_n^{1/d})$. For Lipschitz functions, the error term ε_n will then have a standard deviation σ_n of the order $O(m_n^{-1/d})$. As soon as $d \geq 2$, Equation (24) is no longer fulfilled. Given a comparable number of operations, the OLSMC estimator cannot achieve a convergence rate acceleration in comparison to ordinary Monte Carlo integration. \triangle

Example 2. (*Lindeberg condition.*) We elaborate on Example 1 to illustrate the Lindeberg condition. For ease of notation, put $k_n = m_n + 1$ and consider the integrand $f(x) = \mathbb{1}_{[u, 1]}(x)$ for $x \in [0, 1]$, for some fixed $u \in [0, 1]$.

If u is rational, then for infinitely many integer n we can write $u = \ell_n/k_n$ for some $\ell_n \in \{0, \dots, k_n\}$, and it follows that f is a member of the linear span \mathcal{F}_n of $\{1, h_{n, 1}, \dots, h_{n, m_n}\}$. In that case, $\sigma_n = 0$ for such n , and the normalized integration can obviously not converge to the standard normal distribution.

Suppose that $u \in (0, 1)$ is irrational, and for every n , let $\ell_n \in \{0, \dots, k_n - 1\}$ be such that $a_n = \ell_n/k_n \leq u < (\ell_n + 1)/k_n = b_n$. The L_2 -orthogonal projection of f on \mathcal{F}_n is

given by the piecewise constant function

$$f^{(n)}(x) = \begin{cases} 0 & \text{if } 0 \leq x < a_n, \\ v_n = k_n(b_n - u) & \text{if } x \in [a_n, b_n), \\ 1 & \text{if } b_n \leq x \leq 1. \end{cases}$$

The approximation error $\varepsilon_n = f - f^{(n)}$ is

$$\varepsilon_n(x) = \begin{cases} 0 & \text{if } x \in [0, 1] \setminus [a_n, b_n), \\ -v_n & \text{if } a_n \leq x < u, \\ 1 - v_n & \text{if } u \leq x < b_n, \end{cases}$$

with error variance $\sigma_n^2 = P(\varepsilon_n^2) = k_n(b_n - u)(u - a_n)$. The squared, standardized approximation error is thus

$$\varepsilon_n^2(x)/\sigma_n^2 = \begin{cases} 0 & \text{if } x \in [0, 1] \setminus [a_n, b_n), \\ k_n \frac{b_n - u}{u - a_n} & \text{if } a_n \leq x < u, \\ k_n \frac{u - a_n}{b_n - u} & \text{if } u \leq x < b_n. \end{cases}$$

Now assume that there exists $c > 0$ such that for all pairs of integers (p, q) with $q \geq 1$, we have

$$\left| u - \frac{p}{q} \right| > \frac{c}{q^2}.$$

Such a number u is called a *badly approximable number* [4, p. 245]. It then follows that

$$\sup_{x \in [0, 1]} \varepsilon_n^2(x)/\sigma_n^2 \leq k_n \frac{k_n^{-1}}{ck_n^{-2}} = \frac{1}{c} k_n^2.$$

Since necessarily $k_n^2 = o(n)$ by the leverage condition, it follows that the indicator in the Lindeberg condition is zero for all sufficiently large n , and thus that the Lindeberg condition is fulfilled.

Example 3. (*Univariate polynomials.*) Suppose that $h_{n,j} = h_j$ is equal to the Legendre polynomial L_j of degree $j = 1, \dots, m_n$. The Legendre polynomials are orthogonal on $S = [-1, 1]$ with respect to the uniform distribution P . The Gram matrix $P(h_n h_n')$ is diagonal with entries $1/(2j+1)$ on the diagonal. Furthermore, the Legendre polynomials satisfy $|L_j(x)| \leq 1$ for $x \in [-1, 1]$ while $L_j(1) = 1$. Hence $q_n(x) = \sum_{j=1}^{m_n} (2j+1)L_j(x)^2$, with supremum $q_n(1) = \sum_{j=1}^{m_n} (2j+1) = m_n(m_n+2)$. Equation (16) is satisfied when $m_n = o(n^{1/3})$.

If f is $k+1$ times continuously differentiable for some integer $k \geq 1$, then the bounds on the Legendre coefficients in Theorem 2.1 in [32] imply that $\sigma_n^2 = O(m_n^{-2k-1})$. The convergence rate of the OLSMC estimator is thus $O(m_n^{-k-1/2} n^{-1/2})$. The smoother f , the faster the rate. Condition (23) is fulfilled as soon as f is twice continuously differentiable ($k \geq 1$). For such functions f , increasing the number of polynomial control variates reduces the integration error at a faster rate than increasing the number of Monte Carlo points can achieve. \triangle

For the Fourier basis on $S = [0, 1]$, it is shown in [28] that essentially the same conclusions hold as for the polynomial basis in Example 3.

Example 4. (*Multivariate polynomials.*) As in [1] and [20], suppose that $S = [-1, 1]^d$ (or more generally a Cartesian product of compact intervals) and that P is the uniform distribution on S . As control variates $h_{n,j} = h_j : S \rightarrow \mathbb{R}$, consider tensor products $h_j(x) = \prod_{\ell=1}^d \bar{L}_{a_j(\ell)}(x_\ell)$ for $x = (x_1, \dots, x_d) \in S$, where \bar{L}_a is the normalised Legendre polynomial of degree $a \in \mathbb{N} = \{0, 1, 2, \dots\}$. The sequence of degree vectors $a_j = (a_j(1), \dots, a_j(d)) \in \mathbb{N}^d \setminus \{(0, \dots, 0)\}$ is such that no polynomial of degree $a + 1$ appears in one of the coordinates as long as not all polynomials of degree up to a have appeared in all other coordinates.

As shown in [1, Example II] and the proof of Theorem 4 in [20], we then have $\sup_{x \in S} |h_j(x)| = O(j^{1/2})$ as $j \rightarrow \infty$ and the smallest eigenvalue of the $m \times m$ Gram matrix of (h_1, \dots, h_m) is bounded away from zero, uniformly for all m . By Remark 2, it then follows that $\sup_{x \in S} q_n(x) = O(\sum_{j=1}^{m_n} j) = O(m_n^2)$. As a consequence, the leverage condition is satisfied as soon as $m_n^2 = o(n/m_n)$, i.e., $m_n^3 = o(n)$ as $n \rightarrow \infty$.

Further, assume that the integrand $f = f_n$ is k times continuously differentiable on S , for some integer $k \geq 1$. In the proof of Theorem 4 in [20], Theorem 8 in [17] is cited according to which we have $\sup_{x \in S} |\varepsilon_n(x)| = O(m_n^{-k/d})$. But then also $\sigma_n = O(m_n^{-k/d})$. The convergence rate of the OLSMC estimator is then $O(m_n^{-k/d} n^{-1/2})$. In view of Equation (23), it is more efficient to increase the number of control variates than the Monte Carlo sample size as soon as $k > d$, i.e., the integrand f is sufficiently smooth.

6. Concluding remarks

The paper provides a new asymptotic theory for Monte Carlo integration with control variates. Our main result is that the $n^{-1/2}$ convergence rate of the basic Monte Carlo method can be improved when using a growing number, m , of control variates. The obtained convergence rate, $n^{-1/2} \sigma_m$, is then impacted by the value of σ_m , which reflects the approximation quality of the integrand in the space of control variates. The considered examples have shown that the practical benefits might be important depending, obviously, on σ_m and also on the computation time needed to invert the Gram matrix of the control variates. Attractive avenues for further research are now discussed.

Combination with other integration methods. Theorem 1 echoes other studies (based on different techniques than control variates) that establish acceleration of the standard Monte Carlo rate $n^{-1/2}$. This includes Quasi-Monte Carlo integration [6], Gaussian quadrature [3], which has been studied recently in a (repulsive) Monte Carlo sampling context [2], parametric [27] and nonparametric [33] adaptive importance sampling, and kernel smoothing methods [5]. Combining control variates with some of the previous methods, as has been done with Quasi-Monte Carlo in [23] and with parametric importance sampling in [25], might allow to design even more efficient algorithms.

Theoretical perspectives. Non-asymptotic bounds would offer a different type of guarantee than the one provided in the paper: for a pre-specified probability level, one would have an error bound depending on n and σ_m . In addition, the present work only considers the integration error for a single integrand whereas uniform bounds over some classes of integrands would be appropriate. Such results would apply to situations where many integrals are to be computed as for instance in likelihood-based inference for parametric models with latent variables.

Regularization. As illustrated by the *leverage condition*, the number of control variates at use needs to be limited but, in the mean time, the bound obtained, $n^{-1/2}\sigma_m$, is decreasing in the number of control variates. This advocates for selecting the most informative control variates before using them in the Monte Carlo procedure. Such an approach, based on the Lasso, has already been proposed in [30] and most recently, a pre-selection of the control variates, still by the Lasso, has been studied in [16]. The theoretical bounds obtained and the numerical illustration therein clearly advocate for pre-selecting the most effective control variates.

Un-normalized densities. Applications to Bayesian inference on models defined by un-normalized densities are not included in the present study. Two strategies might be conducted to handle such a situation. The first one consists in a normalized importance sampling approach. Suppose $h = (h_1, \dots, h_m)'$ is a vector of control variates with respect to Lebesgue measure λ . Let p denote the un-normalized target density and q the importance sampling density. Let (X_1, \dots, X_n) be an independent random sample from q . Let $\hat{\mu}_n^{\text{wOLS}}(f)$ denote the weighted OLS estimate defined as in (5) but replacing f by fp/q and h by h/q . Note that $\hat{\mu}_n^{\text{wOLS}}(f)$ is an unbiased estimate of $\int fp d\lambda$. Because p is un-normalized, the estimate cannot be computed and instead one needs to rely on the normalized version $\hat{\mu}_n^{\text{wOLS}}(f)/\hat{\mu}_n^{\text{wOLS}}(1)$. The second strategy follows from [22] and relies on a Markov chain Monte Carlo approach. The control variates are defined through the Stein identity, see Eq. (1) in the aforementioned paper. The sequence of integration points (X_1, \dots, X_n) is generated using the Metropolis–Hastings algorithm with target p . These two modifications allow to work with un-normalized densities. Non-trivial modifications of our proofs would be needed to analyse such procedures.

7. Proofs

Proof of (12). Put $\varepsilon^{(n)} = (\varepsilon(X_1), \dots, \varepsilon(X_n))'$. We have $f^{(n)} = \mu 1_n + \beta'_{\text{opt}} H^{(n)} + \varepsilon^{(n)}$. Since $I_n - \Pi_{n,m}$ is the projection matrix on the orthocomplement in \mathbb{R}^n of the column space of $H^{(n)}$, we have by (11) that

$$\begin{aligned} \hat{\sigma}_{n,\text{OLS}}^2 &= \frac{1}{n}(\mu 1_n + \varepsilon^{(n)} - \hat{\mu}_n^{\text{OLS}} 1_n)'(I_n - \Pi_{n,m})(\mu 1_n + \varepsilon^{(n)} - \hat{\mu}_n^{\text{OLS}} 1_n) \\ &= \frac{1}{n}(\varepsilon^{(n)})'(I_n - \Pi_{n,m})\varepsilon^{(n)} - \frac{1}{n}(\hat{\mu}_n^{\text{OLS}} - \mu)1_n'(I_n - \Pi_{n,m})\{2\varepsilon^{(n)} - (\hat{\mu}_n^{\text{OLS}} - \mu)1_n\}. \end{aligned}$$

Replace $\Pi_{n,m}$ by the right-hand side in (7) to find

$$\begin{aligned}\hat{\sigma}_{n,\text{OLS}}^2 &= P_n(\varepsilon^2) - P_n(\varepsilon h') P_n(h h')^{-1} P_n(h \varepsilon) \\ &\quad - (\hat{\mu}_n^{\text{OLS}} - \mu) \{2P_n(\varepsilon) - 2P_n(h') P_n(h h')^{-1} P_n(h \varepsilon)\} \\ &\quad + (\hat{\mu}_n^{\text{OLS}} - \mu)^2 \{1 - P_n(h') P_n(h h')^{-1} P_n(h)\}.\end{aligned}$$

Equation (10) and the identity $f = \mu + \beta'_{\text{opt}} h + \varepsilon$ imply that

$$\hat{\mu}_n^{\text{OLS}} - \mu = \frac{P_n(\varepsilon) - P_n(h') P_n(h h')^{-1} P_n(h \varepsilon)}{1 - P_n(h') P_n(h h')^{-1} P_n(h)}.$$

Use this identity to simplify the expression for $\hat{\sigma}_{n,\text{OLS}}^2$ and arrive at (12). \square

The Euclidean norm of a vector v is denoted by $|v| = (v'v)^{1/2}$. The corresponding matrix norm is $|A|_2 = \sup\{|Av|/|v| : v \neq 0\}$. The Frobenius norm of a rectangular matrix A is given by $|A|_F = (\sum_i \sum_j A_{ij}^2)^{1/2} = \{\text{tr}(A'A)\}^{1/2}$, with tr the trace operator. We have $|A|_2 \leq |A|_F$, since $|A|_2^2$ is equal to the largest eigenvalue of $A'A$, while $|A|_F^2$ is equal to the sum of all eigenvalues of $A'A$, all of which are nonnegative. Recall the cyclic property of the trace operator: for matrices A and B of dimensions $k \times \ell$ and $\ell \times k$, respectively, we have $\text{tr}(AB) = \text{tr}(BA)$.

Recall that the Gram matrix $P(h_n h_n')$ was assumed to be invertible. Let I_k denote the $k \times k$ identity matrix. Let B_n be an $m_n \times m_n$ matrix such that $B_n' B_n = P(h_n h_n')^{-1}$; use for instance the eigendecomposition of $P(h_n h_n')$ to construct B_n . Clearly, B_n is invertible. The OLS estimator based on the transformed vector of control functions

$$\tilde{h}_n = (\tilde{h}_{n,1}, \dots, \tilde{h}_{n,m_n})' = B_n h_n$$

is therefore identical to the one based on h_n . The transformed vector \tilde{h}_n has the advantage that its elements are orthonormal, i.e., its Gram matrix is equal to the identity matrix:

$$P(\tilde{h}_n \tilde{h}_n') = B_n P(h_n h_n') B_n' = B_n (B_n' B_n)^{-1} B_n' = I_{m_n}. \quad (25)$$

The function q_n defined in (14) is equal to $q_n = \tilde{h}_n' \tilde{h}_n$.

Lemma 1. *We have*

$$\mathbb{E}\{|P_n(h_n)|^2\} = n^{-1} P(h_n' h_n), \quad (26)$$

$$\mathbb{E}\{|P_n(\tilde{h}_n)|^2\} = m_n/n. \quad (27)$$

Proof. We have

$$|P_n(h_n)|^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h_n'(X_i) h_n(X_j).$$

The random variables X_1, \dots, X_n form an independent random sample from P . Furthermore, $P(h_n) = 0$. As a consequence,

$$\mathbb{E}\{|P_n(h_n)|^2\} = n^{-1} \mathbb{E}\{h_n'(X_1) h_n(X_1)\} = n^{-1} P(h_n' h_n),$$

yielding (26). Equation (27) follows from (26) and $P(\tilde{h}_n' \tilde{h}_n) = P(q_n) = m_n$, see (15). \square

Lemma 2.

$$\mathbb{E}\{|P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_F^2\} = n^{-1}\{P_n(q_n^2) - m_n\}. \quad (28)$$

Proof. We have $P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n} = n^{-1} \sum_{i=1}^n A_{n,i}$ with $A_{n,i} = \tilde{h}_n(X_i) \tilde{h}'_n(X_i) - I_{m_n}$. Since the matrix $\tilde{h}_n \tilde{h}'_n$ is symmetric and since the trace operator is linear,

$$\begin{aligned} \mathbb{E}\{|P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_F^2\} &= \mathbb{E}(\text{tr}[\{P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}\}^2]) \\ &= \text{tr}(\mathbb{E}[\{P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}\}^2]) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{tr}\{\mathbb{E}(A_{n,i} A_{n,j})\}. \end{aligned}$$

The triangular array of random matrices $(A_{n,i})_{n,i}$ is rowwise iid; the random matrices $A_{n,i}$ are square integrable and centered. If $i \neq j$, then $\mathbb{E}[A_{n,i} A_{n,j}] = 0$, the $m_n \times m_n$ null matrix. Hence

$$\mathbb{E}\{|P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_F^2\} = n^{-1} \text{tr}\{\mathbb{E}(A_{n,1}^2)\}.$$

By the cyclic property of the trace,

$$\text{tr}\{\mathbb{E}(A_{n,1}^2)\} = \text{tr}[P\{(\tilde{h}_n \tilde{h}'_n)^2\} - I_{m_n}] = P\{(\tilde{h}'_n \tilde{h}_n)^2\} - m_n.$$

Since $\tilde{h}'_n \tilde{h}_n = q_n$, the equality (28) follows. \square

Lemma 3.

$$\mathbb{P}\{P_n(h_n h'_n) \text{ is not invertible}\} \leq n^{-1} P(q_n^2) \quad (29)$$

Proof. Since $\tilde{h} = B_n h_n$ and since B_n is invertible, the matrix $P_n(h_n h'_n)$ is invertible if and only if the matrix $P_n(\tilde{h}_n \tilde{h}'_n)$ is so. Suppose $P_n(\tilde{h}_n \tilde{h}'_n)$ is not invertible. Then there exists a nonzero vector $v \in \mathbb{R}^{m_n}$ such that $P_n(\tilde{h}_n \tilde{h}'_n)v = 0$ and thus $\{P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}\}v = -v$. It then follows that

$$|P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_F \geq |P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_2 \geq 1.$$

But since $I_{m_n} = P(\tilde{h}_n \tilde{h}'_n)$ by (25), equation (28) yields

$$\begin{aligned} \mathbb{P}\{P_n(h_n h'_n) \text{ is not invertible}\} &\leq \mathbb{P}\{|P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_F \geq 1\} \\ &\leq \mathbb{E}\{|P_n(\tilde{h}_n \tilde{h}'_n) - I_{m_n}|_F^2\} \leq n^{-1} P(|\tilde{h}_n|^4). \end{aligned}$$

Finally, $|\tilde{h}_n|^4 = (\tilde{h}'_n \tilde{h}_n)^2 = q_n^2$. \square

Lemma 4. *If Condition 1 holds, then $P_n(h_n h'_n)$ and $P_n(\tilde{h}_n \tilde{h}'_n)$ are invertible with probability tending to one as $n \rightarrow \infty$ and*

$$|P_n(\tilde{h}_n \tilde{h}'_n)^{-1}|_2 \leq 1 + o_{\mathbb{P}}(1), \quad (30)$$

$$P_n(h'_n) P_n(h_n h'_n)^{-1} P_n(h_n) = O_{\mathbb{P}}(m_n/n). \quad (31)$$

Proof. In view of (29), the first part of Condition 1 implies that $P_n(h_n h'_n)$ and thus $P_n(\tilde{h}_n \tilde{h}'_n)$ are invertible with probability tending to one.

Write $J_n = P_n(\tilde{h}_n \tilde{h}'_n)$. On the event that $P_n(h_n h'_n)$ is invertible, J_n is invertible too, and $J_n^{-1} = I_{m_n} + J_n^{-1}(I_{m_n} - J_n)$ and thus $|J_n^{-1}|_2 \leq 1 + |J_n^{-1}|_2 |I_{m_n} - J_n|_2$ by multiplicativity of the matrix norm $|\cdot|_2$. It follows that, provided $|I_{m_n} - J_n|_2 < 1$, we have

$$|J_n^{-1}|_2 \leq \frac{1}{1 - |I_{m_n} - J_n|_2}.$$

Recall that $B'_n B_n = P(h_n h'_n)^{-1}$. By an application of (28) to the orthonormalized functions $\tilde{h}_n = B_n h_n$, we have

$$\mathbb{E}(|J_n - I_{m_n}|_F^2) \leq n^{-1} P(|\tilde{h}_n|^4) = n^{-1} P(q_n^2) = o(1)$$

as $n \rightarrow \infty$, in view of (17). Therefore, $|I_{m_n} - J_n|_2 \leq |I_{m_n} - J_n|_F = o_{\mathbb{P}}(1)$. We conclude that $|J_n^{-1}|_2 \leq 1 + o_{\mathbb{P}}(1)$.

Secondly, since

$$P_n(h'_n) P_n(h_n h'_n)^{-1} P_n(h_n) = P_n(\tilde{h}'_n) P_n(\tilde{h}_n \tilde{h}'_n)^{-1} P_n(\tilde{h}_n),$$

we have

$$|P_n(h'_n) P_n(h_n h'_n)^{-1} P_n(h_n)| \leq |P_n(\tilde{h}_n)|^2 |J_n^{-1}|_2.$$

We have just shown that $|J_n^{-1}|_2 = O_{\mathbb{P}}(1)$. Furthermore, $|P_n(\tilde{h}_n)|^2 = O_{\mathbb{P}}(m_n/n)$ by (27) and Markov's inequality. \square

Recall that $f_n = g_n + \varepsilon_n$, where g_n is the orthogonal projection of f_n on the linear subspace of $L^2(P)$ spanned by $\{1, h_{n,1}, \dots, h_{n,m_n}\}$.

Lemma 5. *We have*

$$\mathbb{E}\{|P_n(h_n \varepsilon_n)|^2\} = n^{-1} P(|h_n|^2 \varepsilon_n^2). \quad (32)$$

If Condition 1 holds, we have therefore

$$|P_n(\tilde{h}_n \varepsilon_n)| = o_{\mathbb{P}}(m_n^{-1/2} \sigma_n), \quad n \rightarrow \infty. \quad (33)$$

Proof. We have

$$|P_n(h_n \varepsilon_n)|^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h'_n(X_i) h_n(X_j) \varepsilon_n(X_i) \varepsilon_n(X_j).$$

Since $P(h_{n,k} \varepsilon_n) = 0$ for all $k = 1, \dots, m_n$ and since the variables X_1, \dots, X_n are iid P , we have $\mathbb{E}\{|P_n(h_n \varepsilon_n)|^2\} = n^{-1} \mathbb{E}\{h'_n(X_1) h_n(X_1) \varepsilon_n(X_1)^2\}$, yielding (32).

Apply (32) to \tilde{h}_n ; since $|\tilde{h}_n|^2 = \tilde{h}'_n \tilde{h}_n = q_n$, we find

$$\mathbb{E}\{|P_n(\tilde{h}_n \varepsilon_n)|^2\} = n^{-1} P(|\tilde{h}_n|^2 \varepsilon_n^2) = n^{-1} P(q_n \varepsilon_n^2) = o(m_n^{-1} \sigma_n^2)$$

as $n \rightarrow \infty$, by (21). \square

Proof of Theorem 1. On an event E_n with probability tending to one, $P_n(h_n h'_n)$ is invertible and $P_n(h'_n) P_n(h_n h'_n)^{-1} P_n(h_n)$ is less than 1 (Lemma 4). On E_n , the OLS estimator is given by (10). Substitute $f_n = \mu_n + \beta'_n h_n + \varepsilon_n$ to see that, on E_n , we have

$$\sqrt{n}(\hat{\mu}_n^{\text{OLS}} - \mu_n) = \sqrt{n} \frac{P_n(\varepsilon_n) - P_n(\varepsilon_n h'_n) P_n(h_n h'_n)^{-1} P_n(h_n)}{1 - P_n(h'_n) P_n(h_n h'_n)^{-1} P_n(h_n)}.$$

By (31), the denominator is $1 + o_{\mathbb{P}}(1)$ as $n \rightarrow \infty$. The second term in the numerator does not change if we replace h_n by \tilde{h}_n . Its absolute value is bounded by

$$|P_n(\tilde{h}_n \varepsilon_n)| |P_n(\tilde{h}_n h'_n)^{-1}|_2 |P_n(\tilde{h}_n)| = o_{\mathbb{P}}(m_n^{-1/2} \sigma_n) O_{\mathbb{P}}(1) O_{\mathbb{P}}\{(m_n/n)^{1/2}\} = o_{\mathbb{P}}(n^{-1/2} \sigma_n);$$

here we used (33), (30), and (27), respectively. We find

$$\sqrt{n}(\hat{\mu}_n^{\text{OLS}} - \mu_n) = \sqrt{n}\{1 + o_{\mathbb{P}}(1)\} P_n(\varepsilon_n) + o_{\mathbb{P}}(\sigma_n).$$

Since $\mathbb{E}\{P_n(\varepsilon_n)^2\} = n^{-1} \sigma_n^2$, we have $P_n(\varepsilon_n) = O_{\mathbb{P}}(n^{-1/2} \sigma_n)$. We conclude that

$$\sqrt{n}(\hat{\mu}_n^{\text{OLS}} - \mu_n) = \sqrt{n} P_n(\varepsilon_n) + o_{\mathbb{P}}(\sigma_n).$$

Divide both sides by σ_n to conclude the proof of Theorem 1. \square

Proof of Theorem 2. By the Lindeberg–Feller central limit theorem [15, Theorem 5.12] applied to the triangular array $\{\varepsilon_n(X_i) : i = 1, \dots, n\}$ of rowwise iid random variables, Condition 2 is necessary and sufficient for $(\sqrt{n}/\sigma_n) P_n(\varepsilon_n)$ to be asymptotically standard normal. In view of (18) and Slutsky's lemma, $(\sqrt{n}/\sigma_n) P_n(\varepsilon_n)$ is asymptotically standard normal if and only if $(\sqrt{n}/\sigma_n)(\hat{\mu}_n^{\text{OLS}} - \mu)$ is asymptotically standard normal.

We prove (20). As in the proof of Theorem 1, there is a sequence E_n of events with probability tending to one such that on E_n , the matrix $P_n(h_n h'_n)$ is invertible and such that $P_n(h'_n) P_n(h_n h'_n)^{-1} P_n(h_n) < 1$. On E_n , the OLS estimator of σ_n^2 is given by (12). Clearly, we can replace h_n by $\tilde{h}_n = B_n h_n$ and find

$$\begin{aligned} \hat{\sigma}_{n,\text{OLS}}^2 &= P_n(\varepsilon_n^2) - P_n(\varepsilon_n \tilde{h}_n') P_n(\tilde{h}_n \tilde{h}_n')^{-1} P_n(\tilde{h}_n \varepsilon_n) \\ &\quad - (\hat{\mu}_n^{\text{OLS}} - \mu)^2 \{1 - P_n(\tilde{h}_n') P_n(\tilde{h}_n \tilde{h}_n')^{-1} P_n(\tilde{h}_n)\}. \end{aligned}$$

The bounds established in the course of the proof of Theorem 1 together with the fact that $(\hat{\mu}_n^{\text{OLS}} - \mu)^2 = O_{\mathbb{P}}(n^{-1} \sigma_n^2)$ easily yield

$$\hat{\sigma}_{n,\text{OLS}}^2 = P_n(\varepsilon_n^2) + o_{\mathbb{P}}(m_n^{-1} \sigma_n^2).$$

It then suffices to show that $P_n(\varepsilon_n^2)/\sigma_n^2 = 1 + o_{\mathbb{P}}(1)$. But this is a consequence of Proposition 1 below applied to the triangular array $Y_{n,i} = \varepsilon_n^2(X_i)/\sigma_n^2$. The Lindeberg condition is exactly condition (34) in that Proposition. \square

Proposition 1. *Let $\{Y_{n,i} : 1 \leq i \leq n\}$ be a triangular array of nonnegative, rowwise iid random variables with unit expectation. If, as $n \rightarrow \infty$, for all $\delta > 0$, we have*

$$\mathbb{E}[Y_{n,1} \mathbf{1}\{Y_{n,1} > \delta n\}] = o(1), \tag{34}$$

then $n^{-1} \sum_{i=1}^n Y_{n,i} = 1 + o_{\mathbb{P}}(1)$.

Proof. We apply [7, Theorem 2.2.6] with $a_n = b_n = n$. We need to check two conditions: (i) $n\mathbb{P}(Y_{n,1} > n) \rightarrow 0$ and (ii) $n^{-1}\mathbb{E}[Y_{n,1}^2 \mathbf{1}\{Y_{n,1} \leq n\}] \rightarrow 0$ as $n \rightarrow \infty$.

Condition (i) follows at once from $n\mathbb{P}(Y_{n,1} > n) \leq \mathbb{E}[Y_{n,1} \mathbf{1}\{Y_{n,1} > n\}]$ and (34).

Regarding condition (ii), choose $\delta \in (0, 1]$ and note that, since $\mathbb{E}[Y_{n,1}] = 1$, we have

$$\begin{aligned} n^{-1}\mathbb{E}[Y_{n,1}^2 \mathbf{1}\{Y_{n,1} \leq n\}] &= n^{-1}\mathbb{E}[Y_{n,1}^2 \mathbf{1}\{Y_{n,1} \leq \delta n\}] + n^{-1}\mathbb{E}[Y_{n,1}^2 \mathbf{1}\{\delta n < Y_{n,1} \leq n\}] \\ &\leq \delta + \mathbb{E}[Y_{n,1} \mathbf{1}\{\delta n < Y_{n,1}\}]. \end{aligned}$$

The limsup as $n \rightarrow \infty$ is bounded by δ because of (34). Since δ was arbitrary, condition (ii) follows. \square

Acknowledgements

The authors are grateful to Chris Oates and to two anonymous reviewers for useful comments and additional references. The authors gratefully acknowledge support from the Fonds de la Recherche Scientifique (FNRS) A4/5 FC 2779/2014-2017 No. 22342320, from the contract “Projet d’Actions de Recherche Concertées” No. 12/17-045 of the “Communauté française de Belgique” and from the IAP research network Grant P7/06 of the Belgian government (Belgian Science Policy).

References

- [1] ANDREWS, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* **59**, 307–345.
- [2] BARDENET, R. AND HARDY, A. (2016). Monte Carlo with determinantal point processes. *ArXiv e-prints*. arXiv:1605.00361.
- [3] BRASS, H. AND PETRAS, K. (2011). *Quadrature theory* vol. 178 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI. The theory of numerical integration on a compact interval.
- [4] BUGEAUD, Y. (2012). *Distribution Modulo One and Diophantine Approximation*. Cambridge University Press, Cambridge.
- [5] DELYON, B. AND PORTIER, F. (2016). Integral approximation by kernel smoothing. *Bernoulli* **22**, 2177–2208.
- [6] DICK, J. AND PILlichshammer, F. (2010). *Digital nets and sequences*. Cambridge University Press, Cambridge. Discrepancy theory and quasi-Monte Carlo integration.
- [7] DURRETT, R. (2010). *Probability: Theory and Examples* fourth ed. Cambridge University Press, Cambridge, Cambridge.
- [8] GLASSERMAN, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer, New York.

- [9] GLASSERMAN, P. AND YU, B. (2005). Large sample properties of weighted Monte Carlo estimators. *Operations Research* **53**, 298–312.
- [10] GLYNN, P. W. AND SZECHTMAN, R. (2002). Some new perspectives on the method of control variates. In *Monte Carlo and quasi-Monte Carlo methods, 2000 (Hong Kong)*. Springer, Berlin pp. 27–49.
- [11] GOBET, E. AND LABART, C. (2010). Solving bsde with adaptive control variate. *SIAM Journal on Numerical Analysis* **48**, 257–277.
- [12] HESTERBERG, T. AND NELSON, B. (1998). Control variates for probability and quantile estimation. *Management Sci.* **44**, 1295–1312.
- [13] HUBER, P. J. (1981). *Robust Statistics*. John Wiley, New York.
- [14] JIE, T. AND ABBEEL, P. (2010). On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*. pp. 1000–1008.
- [15] KALLENBERG, O. (2002). *Foundations of Modern Probability* second ed. Springer, New York.
- [16] LELUC, R., PORTIER, F. AND SEGERS, J. (June 2019). Control variates selection for Monte Carlo integration. *ArXiv e-prints*. arXiv:1906.10920.
- [17] LORENTZ, G. G. (1986). *Approximation of Functions* second ed. Chelsea Publishing Co., New York.
- [18] MCCULLOCH, C. E. AND SEARLE, S. R. (2001). *Generalized, linear, and mixed models*. Wiley-Interscience [John Wiley & Sons], New York.
- [19] MCFADDEN, D. (2001). Economic choices. *The American Rconomic Review* **91**, 351–378.
- [20] NEWHEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79**, 147–168.
- [21] NOVAK, E. (2016). Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*. Springer pp. 161–183.
- [22] OATES, C. J., COCKAYNE, J., BRIOL, F.-X. AND GIROLAMI, M. (2018). Convergence rates for a class of estimators based on Stein’s method. *to appear in Bernoulli*.
- [23] OATES, C. J. AND GIROLAMI, M. (2016). Control functionals for quasi-Monte Carlo integration. *Nineteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Journal of Machine Learning Research W&CP* **51**, 56–65; arXiv:1501.03379v7.

- [24] OATES, C. J., GIROLAMI, M. AND CHOPIN, N. (2017). Control functionals for Monte Carlo integration. *J. R. Statist. Soc. B* **79**, 695–718.
- [25] OWEN, A. AND ZHOU, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95**, 135–143.
- [26] OWEN, A. B. (2013). Monte Carlo Theory, Methods and Examples. <http://statweb.stanford.edu/~owen/mc/>.
- [27] PORTIER, F. AND DELYON, B. (2018). Asymptotic optimality of adaptive importance sampling. In *Advances in Neural Information Processing Systems*. pp. 3134–3144.
- [28] PORTIER, F. AND SEGERS, J. (March 2018). Monte Carlo integration with a growing number of control variates. *ArXiv e-prints*. arXiv:1801.01797v3.
- [29] ROBERT, C. P. AND CASELLA, G. (2004). *Monte Carlo Statistical Methods* second ed. Springer Texts in Statistics. Springer-Verlag, New York.
- [30] SOUTH, L. F., OATES, C. J., MIRA, A. AND DROVANDI, C. (2018). Regularised zero-variance control variates. *arXiv preprint arXiv:1811.05073*.
- [31] VELLEMAN, P. F. AND WELSCH, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician* **35**, 234–242.
- [32] WANG, H. AND XIANG, S. (2012). On the convergence rates of Legendre approximation. *Math. Comp.* **81**, 861–877.
- [33] ZHANG, P. (1996). Nonparametric importance sampling. *J. Amer. Statist. Assoc.* **91**, 1245–1253.