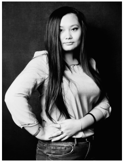
A Quadrature Rule combining Control Variates and Adaptive Importance Sampling

Rémi Leluc [1], François Portier [2], Johan Segers [3], Aigerim Zhuman [3]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] CREST, ENSAI, Rennes, France
[3] ISBA, UCLouvain, Louvain-la-Neuve, Belgium

Advances in Neural Information Processing Systems 2022

Background, Goal and Contributions

• Sequential simulation has emerged as a leading approach to compute multidimensional integrals where the **target density** may be known only up to a proportionality constant (e.g. Bayesian inference)

• While the design of algorithms with adaptive **policies** has been of major interest recently, only few studies have focused on **control variates** to reduce the variance.

GOAL: numerically calculate an integral using **importance sampling** and reduce the variance by including **control variates**.

Contributions:

(1) A simple weighted least squares approach is proposed to improve the procedure of sequential algorithms with control variates.

(2) The proposed approach significantly improves the accuracy of the initial algorithm, both theoretically and in practice.

(3) It takes the form of a **quadrature rule** with adapted quadrature weights that **do not depend on the integrand** and reflect the information brought in by the control variates.

(4) **Non-asymptotic bound** on the probabilistic error of the procedure.

Background, Goal and Contributions

• Sequential simulation has emerged as a leading approach to compute multidimensional integrals where the **target density** may be known only up to a proportionality constant (e.g. Bayesian inference)

• While the design of algorithms with adaptive **policies** has been of major interest recently, only few studies have focused on **control variates** to reduce the variance.

GOAL: numerically calculate an integral using **importance sampling** and reduce the variance by including **control variates**.

Contributions:

(1) A simple weighted least squares approach is proposed to improve the procedure of sequential algorithms with control variates.

(2) The proposed approach significantly improves the accuracy of the initial algorithm, **both theoretically and in practice**.

(3) It takes the form of a **quadrature rule** with adapted quadrature weights that **do not depend on the integrand** and reflect the information brought in by the control variates.

(4) **Non-asymptotic bound** on the probabilistic error of the procedure.

Preliminaries on Monte Carlo Integration

GOAL: Given an **integrand** $g : \mathbb{R}^d \to \mathbb{R}$ and a target **density function** f, the goal is to compute the integral

$$\alpha = \mathbb{E}_f[g] = \int_{\mathbb{R}^d} g(x) f(x) \, \mathrm{d}x$$

Can we sample from target distribution f ?

Preliminaries on Monte Carlo Integration

GOAL: Given an **integrand** $g : \mathbb{R}^d \to \mathbb{R}$ and a target **density function** f, the goal is to compute the integral

$$\alpha = \mathbb{E}_f[g] = \int_{\mathbb{R}^d} g(x) f(x) \, \mathrm{d}x$$

Can we sample from target distribution f ?

- YES, then use naive Monte Carlo estimate (later on control variates)

$$I_n^{(\mathrm{mc})}(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i), \quad X_1, \dots, X_n \sim f$$

- NO, then use **importance sampling** with sampling policy q

$$I_{\mathrm{norm}}^{(\mathrm{is})}(g) = \frac{\sum_{i=1}^{n} w_i g(X_i)}{\sum_{i=1}^{n} w_i}, \quad X_1, \dots, X_n \sim q,$$

where the sequence $(w_i)_{i=1,\dots,n}$ of **importance weights** is defined by

$$w_i = f(X_i)/q(X_i).$$

Generalization: Adaptive Importance Sampling (AIS)

GOAL:

$$\alpha = \mathbb{E}_f[g] = \int_{\mathbb{R}^d} g(x) f(x) \, \mathrm{d}x$$

• Use a **sampling policy** $(q_t)_{t \geq 0}$ = a sequence of densities which evolves adaptively depending on previous outcomes with $q_t \longrightarrow f$ when $t \to \infty$. and an **allocation policy** $(n_t)_{t \geq 0}$.

• At time t, draw n_t **particles** $X_{t,1}, \ldots, X_{t,n_t} \sim q_{t-1}$ with **importance weights** $w_{t,i} = f(X_{t,i}) / q_{t-1}(X_{t,i})$.

• The normalized AIS estimate [DP21] of α is given by

$$I_{\mathrm{norm}}^{(\mathrm{ais})}(g) = \frac{\sum_{t=1}^{T} \sum_{i=1}^{n_t} w_{t,i} g(X_{t,i})}{\sum_{t=1}^{T} \sum_{i=1}^{n_t} w_{t,i}}.$$

Control Variates: variance reduction with samples from f

GOAL:

$$\alpha = \mathbb{E}_f[g] = \int_{\mathbb{R}^d} g(x) f(x)\,\mathrm{d}x$$

• **Control variates** are functions $h_1, \ldots, h_m \in L_2(f)$ with known integrals. Let $h = (h_1, \ldots, h_m)^\top$, assume that $\mathbb{E}_f[h_j] = 0$ for all $j = 1, \ldots, m$. (Stein control variates)

• For any $\beta \in \mathbb{R}^m$, we have $\mathbb{E}_f[g - \beta^\top h] = \mathbb{E}_f[g]$ leading to the CV estimate of α, parameterized by β

$$I_n^{(\mathrm{cv})}(g, \beta) = \frac{1}{n} \sum_{i=1}^{n} [g(X_i) - \beta^\top h(X_i)], \quad X_1, \ldots, X_n \sim f.$$

• What optimal choice for $\beta^\star$? Look at variance and define

$$\beta^* = \underset{\beta \in \mathbb{R}^m}{\arg\min}\, \mathbb{E}_f\left[(g - \mathbb{E}_f[g] - \beta^\top h)^2\right]$$

Control Variates and Least-Squares

- Provided matrix $G = \mathbb{E}_f[hh^\top]$ is invertible, there is a unique $\beta^* \in \mathbb{R}^m$ for which the variance of $I_n^{(\mathrm{cv})}(g)$ is minimal: $\beta^* = \left(\mathbb{E}_f[hh^\top]\right)^{-1} \mathbb{E}_f[hg]$.
- Casting the problem in an **Ordinary Least Squares** framework leads to the control variate estimate

$$I_n^{(\mathrm{cv})}(g) = I_n^{(\mathrm{cv})}\left(g, \hat{\beta}_n^{(\mathrm{cv})}\right) = \hat{\alpha}_n^{(\mathrm{cv})} \quad \text{where } X_1, \ldots, X_n \sim f,$$

$$\left(\hat{\alpha}_n^{(\mathrm{cv})}, \hat{\beta}_n^{(\mathrm{cv})}\right) \in \underset{(a,b)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \{g(X_i) - a - b^\top h(X_i)\}^2$$

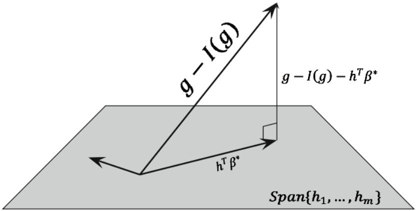


Figure: L^2 projection of g onto space of control variates $Span\{h_1, \ldots, h_m\}$.

Adaptive Importance Sampling with Control Variates

• AISCV estimate is the first coordinate of the solution to the **Weighted Least Squares** problem

$$(\hat{\alpha}_n, \hat{\beta}_n) = \underset{a\in\mathbb{R}, b\in\mathbb{R}^m}{\arg\min} \sum_{i=1}^{n} w_i \left[g(X_i) - a - b^\top h(X_i) \right]^2, w_i = f(X_i)/q_{i-1}(X_i).$$

• (a) (Exact integration) whenever g is of the form $\alpha + \beta^\top h$ for some $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^m$, the **error is zero**, i.e., $\hat{\alpha}_n = \alpha = \int gf \,\mathrm{d}\lambda$.

• (b) (Quadrature Rule) the estimate takes the form of a quadrature rule $\hat{\alpha}_n = \sum_{i=1}^n v_{n,i} g(X_i)$, for **quadrature weights** $v_{n,i}$ **that do not depend on the function** g and that can be computed by a single weighted least squares procedure.

• (c) (Bayesian) it can be computed even when f **is known only up to a multiplicative constant**.

• (d) (post-hoc scheme) CV can be brought into play in a ***post-hoc*** **scheme**, after generation of the particles and importance weights, and **this for any AIS algorithm**

Concentration inequality for the AISCV estimate

Theorem

Under assumptions, for any $\delta \in (0,1)$ and for all $n \geq C_1 c^2 B \log(10m/\delta)$, we have, with probability at least $1-\delta$, that

$$\left| I_{\mathrm{norm}}^{(\mathrm{aiscv})}(g) - \int_{\mathbb{R}^d} g(x) f(x)\,\mathrm{d}x \right| \leq C_2 \tau \sqrt{\frac{\log(10/\delta)}{n}} + C_3 c B \tau \frac{\log(10m/\delta)}{n},$$

where C_1, C_2, C_2 are some constants and $B = \sup_{x : f(x) > 0} \|\hbar(x)\|_2^2$, $\hbar = G^{-1/2} h$.

Control Variates in Practice and Bayesian Inference

• **Stein control variates** [OGC17] are built with operator $\mathcal{L}$ on functions $\varphi \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ to have $\mathbb{E}_f[\mathcal{L}\varphi] = 0$.

$$(\mathcal{L}\varphi)(x) = \Delta_x \varphi(x) + \nabla_x \varphi(x)^\top \nabla_x \log f(x).$$

• $\nabla_x \log f(x)$ can either be directly computed (Bayesian regression) or with autodiff (Tensorflow and PyTorch).

• Given data $\mathcal{D}$ and parameter of interest $\theta \in \mathbb{R}^d$, posterior integrals take the form $\int_{\mathbb{R}^d} g(\theta) p(\theta|\mathcal{D}) \,\mathrm{d}\theta$, where $p(\theta|\mathcal{D}) \propto \ell(\mathcal{D}|\theta)\pi(\theta)$ is the posterior distribution, proportional to prior $\pi(\cdot)$ and a likelihood $\ell(\mathcal{D}|\cdot)$.

Synthetic examples: Gaussian Mixtures

Integrand and Target: $g(x) = x, f_\Sigma(x) = 0.5\Phi_\Sigma(x-\mu) + 0.5\Phi_\Sigma(x+\mu)$ where $\mu = (1,\dots,1)^\top/2\sqrt{d}, \Sigma = I_d/d$ and Φ_Σ is pdf $\mathcal{N}(0,\Sigma)$.
Sampling policy: Multivariate Student
Control variates: Stein method with $\varphi =$ polynomial with bounded degree

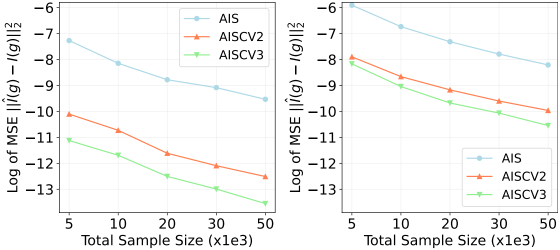
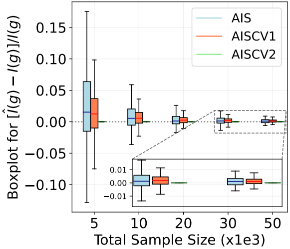


Figure: Gaussian mixture density: Logarithm of $\|\hat{I}(g) - I(g)\|_2^2$ for $g(x) = x$ with target isotropic f_Σ with $d = 4$ (left), $d = 8$ (right).

Bayesian Linear Regression on Real-world data

Data [DG19]: *housing* ($N = 506; d = 13; m \in \{12; 104\}$); *abalone* ($N = 4177; d = 8; m \in \{7; 44\}$).
Prior: $\pi(\theta) \sim \mathcal{N}(\mu_a, \Sigma_a)$, **Posterior**: $p(\theta|\mathcal{D}) \propto \ell(\mathcal{D}|\theta)\pi(\theta)$.
Integrand: $g(\theta) = \sum_{i=1}^{d} \theta_i^2$.
Control variates: Stein control variates with $\varphi_\alpha(\theta) = \theta_1^{\alpha_1} \cdots \theta_d^{\alpha_d}$, $\alpha_1 + \cdots + \alpha_d \leq Q$, $Q \in \{1; 2\}$.



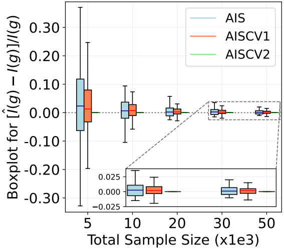


Figure: BLR: boxplots of $(\hat{I}(g) - I(g))/I(g)$ for $g(\theta) = \sum_{j=1}^{d} \theta_j^2$ with datasets Housing (left) and Abalone (right).

Conclusion and take-home message

- This paper provides a new method to incorporate **control variates** within standard **sequential algorithms**.

- The proposed approach significantly improves the accuracy of the initial algorithm, **both theoretically and in practice**.

- Control Variates can be brought into play in a ***post-hoc* scheme**, after generation of the particles and importance weights, and **this for any AIS algorithm**

Thank you and see you at the conference !

References I

[DG19] Dheeru Dua and Casey Graff, *Uci Machine Learning Repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california*, School of Information and Computer Science **25** (2019), 27.

[DP21] Bernard Delyon and François Portier, *Safe adaptive importance sampling: A mixture approach*, The Annals of Statistics **49** (2021), no. 2, 885–917.

[OGC17] Chris J Oates, Mark Girolami, and Nicolas Chopin, *Control functionals for Monte Carlo integration*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79** (2017), no. 3, 695–718.