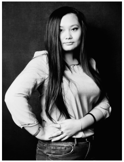
A Quadrature Rule combining Control Variates and Adaptive Importance Sampling

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

arXiv:2205.11890

Advances in Neural Information Processing Systems 2022

Background and Motivation

Underlying integration problem

Let $f : \mathbb{R}^d \to \mathbb{R}_+$ be a **target density** function and $g : \mathbb{R}^d \to \mathbb{R}$ integrable.

- **Goal:** Estimate

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

- **Constraints:**

Only based on evaluations $g(X_1), \ldots, g(X_n)$ where $X_1, \ldots, X_n$ are called *particles*; g may be black-box and sampling from f may be hard[1].

- **Central question:** Accuracy given number of particles

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

[1] In Bayesian statistics, f is the density *a posteriori*.

Background and Motivation: Monte Carlo integration

GOAL: 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 ?

Background and Motivation: Monte Carlo integration

GOAL: 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)

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

Books: Robert and Casella (1999); Evans and Swartz (2000); Glasserman (2004); Owen (2013)

Background and Motivation: Monte Carlo integration

GOAL: 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)

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

Books: Robert and Casella (1999); Evans and Swartz (2000); Glasserman (2004); Owen (2013)

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

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

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

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

Generalization: Adaptive Importance Sampling (AIS)

GOAL:

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

Sampling policy $(q_t)_{t \geq 0}$ = densities which evolve adaptively depending on previous outcomes with $q_t \to f$ when $t \to \infty$.

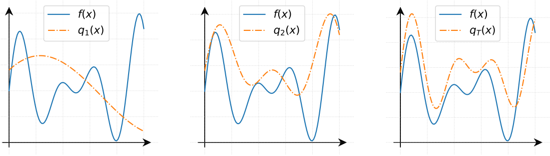


Figure: Evolution of sampling policy is AIS.

- At time t, draw n_t **particles** $X_{t,1}, \ldots, X_{t,n_t} \sim q_{t-1}$ with **importance weights** $w_{t,i} = f(X_{t,i}) / q_{t-1}(X_{t,i})$ and **allocation policy** $(n_t)_{t \geq 0}$.
- The normalized AIS estimate (Delyon and Portier, 2018) of α is given by

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

Monte Carlo jungle: (adaptive) importance sampling

Sequential simulation = leading approach to compute integrals

• Early works on sequential schemes include (Geweke, 1989; Kloek and Van Dijk, 1978; Oh and Berger, 1992) where the sampling policy $(q_t)_{t\geq 0}$ is chosen out of a **parametric family**.

• Extension of the parametric approach by the **Population Monte Carlo** framework (Cappé et al., 2008, 2004; Martino et al., 2017).

• Various **asymptotic results** in (Chopin, 2004; Douc and Moulines, 2008; Portier and Delyon, 2018).

• **non-parametric importance sampling** in (Dai et al., 2016; Delyon and Portier, 2021; Korba and Portier, 2022; Zhang, 1996)

Monte Carlo jungle: Control variates

Let $X_1, \ldots, X_n \sim f$, naive Monte Carlo estimator is

$$I_n^{(\mathrm{mc})}(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

- Unbiased, consistent, variance $\sigma^2(g)/n$ where $\sigma^2(g) = \mathbb{E}_f[(g - \mathbb{E}_f[g]))^2]$.

increasing n is prohibitive, how to reduce the variance ?

Monte Carlo jungle: Control variates

Let $X_1, \ldots, X_n \sim f$, naive Monte Carlo estimator is

$$I_n^{(\mathrm{mc})}(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

- Unbiased, consistent, variance $\sigma^2(g)/n$ where $\sigma^2(g) = \mathbb{E}_f[(g - \mathbb{E}_f[g]))^2]$.

increasing n is prohibitive, how to reduce the variance ?

Control variates technique

Use the knowledge of functions $h_1, \ldots, h_m$ with **known integrals** $\mathbb{E}_f[h_j]$.

- Benefits can be established theoretically in terms of:

error bounds (Oates et al., 2017); **weak convergence** (Portier and Segers, 2019); **excess risk** (Belomestny et al., 2022); **uniform error bounds** over classes of integrands (Plassier et al., 2020).

The existing control variate methods do not account for sequential changes in the particle distribution as is the case in AIS !

Goal and Contributions

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

<u>Contributions:</u>

(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.

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, Orthogonal Polynomial families)

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, Orthogonal Polynomial families) • 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

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

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

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

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

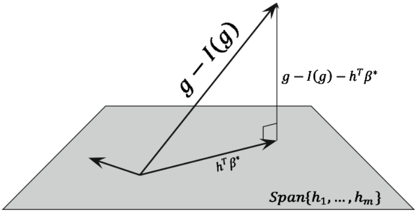


Figure: L^2 projection of g onto space of control variates $Span\{h_1, \ldots, h_m\}$.

Adaptive Importance Sampling with Control Variates

- AISCV estimate is the first coordinate of the solution to the **Weighted Least Squares** problem: $X_i \sim q_{i-1}$

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

Adaptive Importance Sampling with Control Variates

- AISCV estimate is the first coordinate of the solution to the **Weighted Least Squares** problem: $X_i \sim q_{i-1}$

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

- (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**

AISCV algorithm

Require: g, f, $T \in \mathbb{N}^*$, $(n_t)_{t=1}^{T}$, initial density q_0, update rule for q_i

for $t = 1, \ldots, T$ **do**
    Generate an independent random sample $X_{t,1}, \ldots, X_{t,n_t}$ from q_{t-1}
    Compute weights $(w_{t,i})_{i=1}^{n_t}$ where $w_{t,i} = f(X_{t,i})/q_{t-1}(X_{t,i})$
    Construct the matrix of control variates $H_t = \left(h_j(X_{t,i})\right)_{i=1,\ldots,n_t}^{j=1,\ldots,m}$
    Evaluate the integrand in the particles: $(g(X_{t,i}))_{i=1}^{n_t}$
    Update q_t based on the past $(X_{s,i} : s = 1, \ldots, t; i = 1, \ldots, n_s)$
end for
$(\hat{\alpha}_T, \hat{\beta}_T) = \underset{(a,b)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{n_t} w_{t,i} \left(g(X_{t,i}) - a - b^\top h(X_{t,i}) \right)^2 \right\}$
$I_n^{(\mathrm{aiscv})}(g) = \hat{\alpha}_T$.

Concentration inequality for the AISCV estimate

Assumptions

(A1): $\exists c \geq 1 : \forall x \in \mathbb{R}^d, \quad f(x) \leq c \cdot q_i(x)$.

(A2): $\sup_{x : f(x) > 0} |h_j(x)| < \infty$ and $G = \mathbb{E}_f[hh^\top]$ invertible.

(A3): $\exists \tau > 0 : \forall t > 0, i \geq 1, \ \mathbb{P}[|w_i \varepsilon(X_i)| > t \mathcal{F}_{i-1}] \leq 2 \exp(-t^2/(2\tau^2))$

Theorem

Under A1, A2, A3, 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

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

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

Control Variates in Practice and Bayesian Inference

- **Stein control variates** (Oates et al., 2017) are built with operator $\mathcal{L}$ (Stein, 1972; Gorham and Mackey, 2015) 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* in Tensorflow and PyTorch. (Abadi et al., 2016; Paszke et al., 2017)

- 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

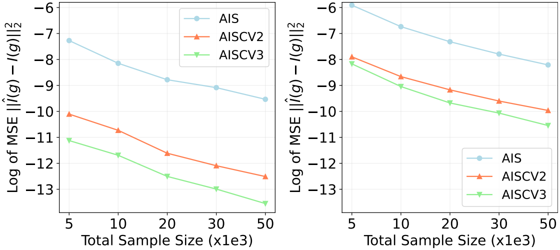


Figure: Gaussian mixture density: Logarithm of $\|\hat{I}(g) - I(g)\|_2^2$ for $g(x) = x$ with target isotropic f_Σ with $d = 4$ (left), $d = 8$ (right).

Bayesian Linear Regression on Real-world data

Data (Dua and Graff, 2019): *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 $\varphi_\alpha(\theta) = \theta_1^{\alpha_1} \cdots \theta_d^{\alpha_d}$, $\alpha_1 + \cdots + \alpha_d \leq Q$, $Q \in \{1; 2\}$.

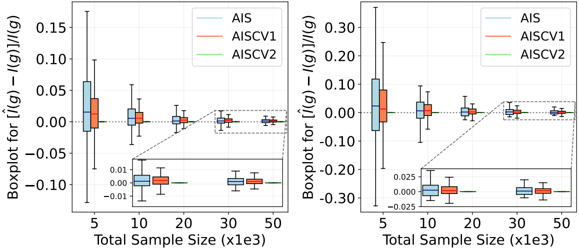


Figure: BLR: boxplots of $(\hat{I}(g) - I(g))/I(g)$ for $g(\theta) = \sum_{j=1}^{d} \theta_j^2$ with datasets Housing (left) and Abalone (right).

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 !

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Belomestny, D., Iosipoi, L., Paris, Q., and Zhivotovskiy, N. (2022). Empirical variance minimization with applications in variance reduction and optimal control. *Bernoulli*, 28(2):1382–1407.

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459.

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

References II

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411.

Dai, B., He, N., Dai, H., and Song, L. (2016). Provable Bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994. PMLR.

Delyon, B. and Portier, F. (2018). Asymptotic optimality of adaptive importance sampling. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3138–3148. Curran Associates Inc.

Delyon, B. and Portier, F. (2021). Safe adaptive importance sampling: A mixture approach. *The Annals of Statistics*, 49(2):885–917.

Douc, R. and Moulines, E. (2008). Limit theorems for weighted samples with applications to sequential monte carlo methods. *The Annals of Statistics*, pages 2344–2376.

References III

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

Evans, M. and Swartz, T. (2000). *Approximating integrals via Monte Carlo and deterministic methods*. Oxford Statistical Science Series. Oxford University Press, Oxford.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339.

Glasserman, P. (2004). *Monte Carlo methods in financial engineering*, volume 53. Springer, New York, NY, USA.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, 28.

References IV

Kloek, T. and Van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19.

Korba, A. and Portier, F. (2022). Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pages 11503–11527. PMLR.

Martino, L., Elvira, V., Luengo, D., and Corander, J. (2017). Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623.

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

Oh, M.-S. and Berger, J. O. (1992). Adaptive importance sampling in monte carlo integration. *Journal of Statistical Computation and Simulation*, 41(3-4):143–168.

Owen, A. B. (2013). Monte carlo theory, methods and examples.

References V

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Plassier, V., Portier, F., and Segers, J. (2020). Risk bounds when learning infinitely many response functions by ordinary linear regression. *arXiv preprint arXiv:2006.09223. To appear in Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*.

Portier, F. and Delyon, B. (2018). Asymptotic optimality of adaptive importance sampling. *Advances in Neural Information Processing Systems*, 31:3134–3144.

Portier, F. and Segers, J. (2019). Monte Carlo integration with a growing number of control variates. *Journal of Applied Probability*, 56(4):1168–1186.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2 of *Springer Texts in Statistics*. Springer, second edition.

References VI

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press.

Zhang, P. (1996). Nonparametric importance sampling. *J. Amer. Statist. Assoc.*, 91(435):1245–1253.