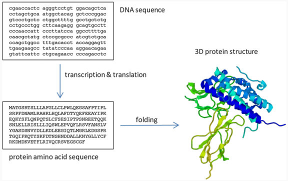
Control Variate Selection for Monte Carlo Integration

Rémi LELUC
Ecole Polytechnique, Institut Polytechnique de Paris, France

Joint work with François Portier and Johan Segers, paper
Published in *Statistics and Computing, 2021.*

Motivation: Machine Learning recent advances

AlphaGo (2016)



AlphaFold (2018)

GPT-3/4(2020/2023)

"Intelligence"

=

Data + Models + **Algorithms** + Computing Power

Motivation: need for integral estimators

Central Question: *Integration*

Computation of an *integral* through probabilistic objective $\mathcal{F}$

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi_\theta(x)}[f(x)] = \int_{\mathcal{X}} f(x)\pi_\theta(x)\mathrm{d}x. \quad (1)$$

Main issue: intractability and computational cost

- **(RL)** Trajectory $\tau = (s_0, a_0, \dots, s_{T-1}, a_{T-1})$ with policy π_θ and cumulative return $\mathcal{R}(\tau) = \sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$.

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi_\theta(\tau)}[\mathcal{R}(\tau)]$$

(2016) AlphaGo A.I. beats champion Lee Sedol in Go.

- **(VI)** $\mathcal{F}$ optimises the log-likelihood $\log p(x|z)$ under a regularization constraint which promotes closeness between the density q and the prior distribution $p(z)$

$$\text{ELBO} = \mathcal{F}(\theta) = \mathbb{E}_{q_\theta(z|x)}[\log p(x|z)] - \text{KL}(q_\theta(z|x)||p(z)).$$

Advantages of Random estimates

Easy and Practical

→ Requires only three steps: sampling, evaluating, averaging

Randomness as a Strength

→ Naturally escape local optima

→ Complete exploration of the search space

Large-Scale learning

→ simple, scalable, parallelizable

→ in supervised learning, deterministic gradient scales as $O(nd)$, stochastic version reduces to $O(d)$ operations

Theoretical justifications[1]

→ deterministic methods $O(n^{-s/d})$

→ optimal random procedure $O(n^{-1/2}n^{-s/d})$

[1] (Novak, 2016): Some results on the complexity of numerical integration

Integration $\mathcal{F}$

Monte Carlo Integration & Variance Reduction

Monte Carlo integration

Underlying integration problem

Let $(\mathcal{X}, \mathcal{A}, \pi)$ be a probability space, $f : \mathcal{X} \to \mathbb{R}$ with $f \in L_2(\pi)$.

- **Goal:**

$$\pi(f) := \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x) = \mathbb{E}_\pi[f(X)].$$

- **Constraints:** f is unknown (black-box) or no approximation is sufficiently accurate, sampling from π may be hard.

Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \pi$, naive Monte Carlo estimator $\hat{\alpha}_n^{\mathrm{mc}}(f)$ of $\pi(f)$ is

$$\hat{\alpha}_n^{\mathrm{mc}}(f) := \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{2}$$

Research Questions

- How to reduce the variance of Monte Carlo estimates?
- How to sample from π?
- How to achieve optimal convergence rates?

Ref: Metropolis and Ulam (1949); Robert and Casella (1999); Evans and Swartz (2000); Glasserman (2004); Owen (2013); Novak (2016); Chopin and Gerber (2024)

Variance Reduction with Control Variates

Definition: Control Variates

Functions $h_1, \ldots, h_m \in L_2(\pi)$ with known integrals:

$$\forall 1 \leq j \leq m, \quad \mathbb{E}_\pi[h_j] = 0$$

$\rightarrow$ Stein control variates, families of orthogonal polynomials

- Let $h = (h_1, \ldots, h_m)^\top$, for any $\beta \in \mathbb{R}^m$, we have $\mathbb{E}_\pi[f - \beta^\top h] = \mathbb{E}_\pi[f]$ leading to the CV estimate of α, parameterized by β

CV-Monte Carlo

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

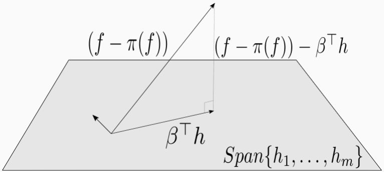
- What optimal choice for β^* ? Look at variance and define

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

Integration with Linear regression

From integration to linear regression

The integral $\pi(f)$ appears as the intercept of a linear regression model with response f and explanatory variables $h_1, \ldots, h_m$,



L_2-orthogonal projection.

- The integral and oracle coefficient satisfy

$$(\pi(f), \beta^\star(f)) \in \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \; \pi[(f - \alpha - \beta^\top h)^2] \tag{3}$$

- Replacing the distribution π by the sample measure $\hat{\pi}_n$ gives the **Ordinary Least Squares** (OLS) estimate, $X_1, \ldots, X_n \sim \pi$

$$\left(\hat{\alpha}_n^{(\mathrm{cv})}, \hat{\beta}_n^{(\mathrm{cv})}\right) \in \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \; \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \alpha - \beta^\top h(X_i)\right)^2 \tag{4}$$

Control Variates in the literature

Applications of control variates

- Finance (Gobet and Labart, 2010; Glasserman, 2004)
- Reinforcement Learning and policy-gradient methods (Jie and Abbeel, 2010; Liu et al., 2018)
- Inference in probabilistic models (Ranganath et al., 2014; Brosse et al., 2018; Belomestny et al., 2020)
- Gradient-based optimization (Wang et al., 2013; Gower et al., 2018)
- Time-series analysis (Davis et al., 2021) and semi-supervised inference (Zhang et al., 2019)

Theoretical results

- Stein method to build control functionals with non-parametric extension (Oates et al., 2017)
- Central Limit Theorem in the regime $m \to +\infty, n \to +\infty$ (Portier and Segers, 2019)
- Variance reduction via regularization (South et al., 2022)

From Ordinary Least Squares Monte Carlo...

Limitations of OLSMC.

- (*Overfitting*) Too many variables or/and few samples (case $m >> n$)
- (*Collinearity*) Dependence among variables $\rightarrow$ very large coefficients

How to avoid those problems ?

Bet on sparsity with **variable selection**!

Image generated by text-to-image A.I. midjourney with the command: "super-hero cowboy twirling his lasso in the air, comic-book style".

… to Lasso Monte-Carlo (LASSOMC/LSLASSO)

Control Variates estimates: OLS, LASSO, LSLASSO

$$\left(\hat{\alpha}_n^{\text{ols}}(f), \hat{\beta}_n^{\text{ols}}(f)\right) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \|f^{(n)} - \alpha\mathbb{1}_n - H\beta\|_2^2$$

$$\left(\hat{\alpha}_n^{\text{lasso}}(f), \hat{\beta}_n^{\text{lasso}}(f)\right) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \frac{1}{2n}\|f^{(n)} - \alpha\mathbb{1}_n - H\beta\|_2^2 + \lambda\|\beta\|_1$$

$$\left(\hat{\alpha}_n^{\text{lslasso}}(f), \hat{\beta}_n^{\text{lslasso}}(f)\right) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^{\hat{\ell}}}{\arg\min} \|f^{(n)} - \alpha\mathbb{1}_n - H_{\hat{S}}\beta\|_2^2$$

- **Active set** $S^\star = \{k : \beta_k^\star \neq 0\}$ and **sparsity level** $\ell^\star = Card(S^\star)$
- LSLASSOMC:

(1) $\hat{S} = \{k : \hat{\beta}_{N,k}^{\text{lasso}}(f) \neq 0\}$ estimated **active set** with **LASSO**

(2) Solve subproblem **OLS** with selected control variates

Non-asymptotic Error Analysis

Assumptions: **sub-gaussian** residuals $\varepsilon = f - \pi(f) - \beta^{\star\top} h$ with factor τ.

Concentration inequalities

For $\delta \in (0,1)$ with probability at least $1-\delta$, for OLS, LASSO, LSLASSO

$$|\hat{\alpha}_n^{\text{ols}}(f) - \pi(f)| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + C_1\sqrt{Bm\log(8m/\delta)}\frac{\tau}{n}$$

$$|\hat{\alpha}_n^{\text{lasso}}(f) - \pi(f)| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + C_2(U_h^2/\gamma^\star)\ell^\star\log(8m/\delta)\frac{\tau}{n}$$

$$|\hat{\alpha}_n^{\text{lslasso}}(f) - \pi(f)| \leq \sqrt{2\log(16/\delta)}\frac{\tau}{\sqrt{n}} + C_3\sqrt{B^\star\ell^\star\log(16\ell^\star/\delta)}\frac{\tau}{n}$$

$U_h = \max_{j=1,\ldots,m} \|h_j\|_\infty$

$G = \mathbb{E}_\pi[hh^\top], \gamma = \lambda_{\min}(G), \hbar = G^{-1/2}h; B = \sup_x \|\hbar(x)\|_2^2$

$G^\star, \gamma^\star, B^\star$ restricted on **active set**

Illustrative examples

Fourier On $\mathcal{X} = [0,1]$ equipped with the uniform distribution P, let $h_j(x)$ be equal to $\sqrt{2}\cos((j+1)\pi x)$ is j is odd and to $\sqrt{2}\sin(j\pi x)$ is j is even.

$G = I_m, \gamma = \gamma^\star = 1, U_h = U_h^\star = \sqrt{2}, \zeta_h = \zeta_h^\star = 2.$

$$|\hat{\alpha}_n^{\mathrm{lslasso}}(f) - P(f)| \le \sqrt{2\log(16/\delta)} + 83\ell^\star \sqrt{\log(16\ell^\star/\delta)\log(8/\delta)}\frac{\tau}{n}.$$

Polynomials Suppose that for all $j = 1, \ldots, m, h_j = L_j$ is the Legendre polynomial of degree j.

The Gram matrix $G = P(hh^T)$ is diagonal with entries $1/(2j+1)$ and $\gamma = 1/(2m+1)$.

$$|\hat{\alpha}_n^{\mathrm{lslasso}}(f) - P(f)| \le$$
$$\sqrt{2\log(16/\delta)} + 58\sqrt{(2\ell^\star+1)\ell^\star\log(16\ell^\star/\delta)\log(8/\delta)}\frac{\tau}{n}.$$

Numerical experiments

- $h_j(x) = L_j(2x-1)$ for $x \in [0,1]$, with L_j the univariate Legendre polynomial (Legendre function of the first kind) of degree j.
- For a multi-index $\ell = (\ell_1, \ldots, \ell_d)$ in $\{0, 1, \ldots, k\}^d \setminus \{(0, \ldots, 0)\}$, build

$$h_\ell(x_1, \ldots, x_d) = \prod_{j=1}^{d} h_{\ell_j}(x_j) = h_{\ell_1}(x_1) \times \ldots \times h_{\ell_d}(x_d)$$

- Sort in ascending order according to the total degree $\sum_{j=1}^{d} \ell_j$.

d	k	Degree threshold				
		1	3	5	10	12
3	12	3	19	55	285	454
5	10	5	55	251	3 001	6 157
8	3	8	164	1 214	20 993	36 813

Number of control variates by degrees

Numerical experiments

- λ is selected by imposing a lower bound and an upper bound on the number of activated random variables $\rightarrow$ **dichotomic search**.
- initialize $\lambda = \lambda_{max}$ and decrease it to have more and more control variates until their number lies in the range $[c_1\sqrt{n}, c_2\sqrt{n}]$.

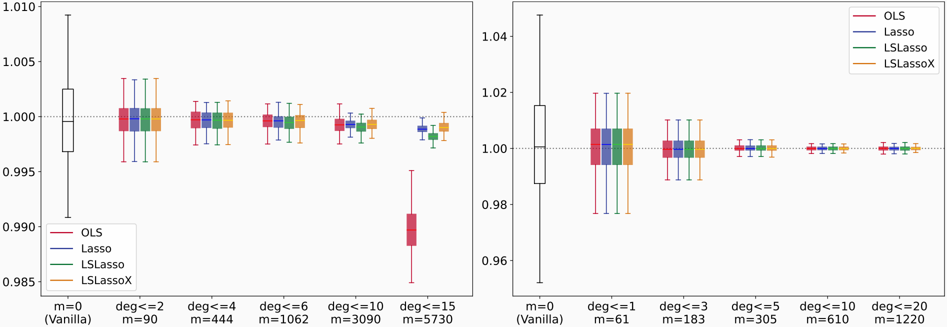
n	N	$\lfloor 3\sqrt{n} \rfloor$	$\lfloor 12\sqrt{n} \rfloor$
2 000	700	134	536
5 000	1 000	212	848
10 000	2 000	300	1 200

Parameters setting with range $(c_1\sqrt{n}, c_2\sqrt{n})$ of selected control variates.

Evidence Estimation in Bayesian Models

- Model likelihood $\ell(x|\theta)$ and prior distribution $\pi(\theta)$, compute evidence

$$Z = \int_{\Theta} \ell(x|\theta)\pi(\theta)\mathrm{d}\theta$$



Boxplots of Error Distribution for Capture ($d = 12$) and Sonar ($d = 61$) datasets[2] , $n = 5000; N = 1000$, obtained over 100 replications.

[2](Marzolin, 1988; Gorman and Sejnowski, 1988)

LASSOMC: Capture/Sonar experiments

$m =$	90	444	1 062	3 090	5 730
OLS	8.23	10.3	5.21	0.01	5e-3
LASSO	7.84	10.5	5.88	2.80	0.85
LSL	7.70	10.4	4.54	1.42	0.43
LSLX	7.59	9.77	7.58	2.73	1.04

Capture data: global efficiency ($n = 2000$)

$m =$	61	183	305	610	1220
OLS	0.27	0.33	3.87	4.68	1.47
LASSO	0.27	0.35	3.96	5.55	3.00
LSL	0.26	0.33	3.85	4.90	2.19
LSLX	0.26	0.35	3.80	4.81	3.17

Sonar data: global efficiency ($n = 2000$)

$m =$	90	444	1 062	3 090	5 730
OLS	5.21	9.56	8.31	1.28	3e-3
LASSO	5.16	9.69	8.59	4.87	1.72
LSL	5.16	9.59	7.88	2.49	0.59
LSLX	5.15	9.55	8.15	4.51	1.72

Capture data: global efficiency ($n = 5000$)

$m =$	61	183	305	610	1220
OLS	0.29	0.41	3.66	6.70	2.57
LASSO	0.28	0.41	3.73	6.85	3.10
LSL	0.28	0.41	3.56	6.66	2.68
LSLX	0.28	0.41	3.70	6.95	3.17

Sonar data: global efficiency ($n = 5000$)

Conclusion

- The use of high-dimensional control variates with the help of a LASSO-type procedure has been shown to be efficient in order to reduce the variance of the basic Monte Carlo estimate.
- The method, called LSLASSO(X), that first selects appropriate control variates by the LASSO, possibly on a smaller subsample, and then estimates the control variate coefficients by least squares performs excellently considering the modest computing time required.
- Future work on debiasing methods for LASSO-based procedures, sample splitting and construction of control variates in adaptive sampling framework.

References

Belomestny, D., L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov (2020). Variance reduction for Markov chains with application to MCMC. *Statistics and Computing 30*, 973–997.

Brosse, N., A. Durmus, S. Meyn, É. Moulines, and A. Radhakrishnan (2018). Diffusion approximations and control variates for MCMC. *arXiv preprint arXiv:1808.01665*.

Chopin, N. and M. Gerber (2024). Higher-order monte carlo through cubic stratification. *SIAM Journal on Numerical Analysis 62*(1), 229–247.

Davis, R., T. do Rego Sousa, and C. Klüppelberg (2021, 01). Indirect inference for time series using the empirical characteristic function and control variates. *Journal of Time Series Analysis 42*.

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

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

Bibliography ii

Gobet, E. and C. Labart (2010). Solving bsde with adaptive control variate. *SIAM Journal on Numerical Analysis 48*(1), 257–277.

Gorman, R. P. and T. J. Sejnowski (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks 1*(1), 75–89.

Gower, R., N. Le Roux, and F. Bach (2018). Tracking the gradients using the hessian: a new look at variance reducing stochastic methods. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Canary Islands, Spain, pp. 707–715. PMLR.

Jie, T. and P. Abbeel (2010). On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems*, Volume 23. Curran Associates, Inc.

Liu, H., Y. Feng, Y. Mao, D. Zhou, J. Peng, and Q. Liu (2018, February). Action-dependent control variates for policy optimization via stein identity. In *ICLR 2018 Conference* (ICLR 2018 Conference ed.).

Marzolin, G. (1988). Polygynie du Cincle plongeur (Cinclus cinclus) dans les côtes de Lorraine. *Oiseau et la Revue Francaise d'Ornithologie 58*(4), 277–286.

Metropolis, N. and S. Ulam (1949). The monte carlo method. *Journal of the American statistical association 44*(247), 335–341.

Novak, E. (2016). Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pp. 161–183. Springer.

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

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

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

Ranganath, R., S. Gerrish, and D. Blei (2014, 22–25 Apr). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Volume 33, Reykjavik, Iceland, pp. 814–822. PMLR.

Robert, C. P. and G. Casella (1999). *Monte Carlo statistical methods* (Second ed.), Volume 2 of *Springer Texts in Statistics*. Springer.

South, L., C. Oates, A. Mira, and C. Drovandi (2022). Regularized zero-variance control variates. *Bayesian Analysis 1*(1), 1–24.

Bibliography iv

Wang, C., X. Chen, A. Smola, and E. Xing (2013). Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

Zhang, A., L. D. Brown, and T. T. Cai (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics 47*(5), 2538–2566.