Monte Carlo Methods in Machine Learning

Rémi LELUC

Ecole Polytechnique, Institut Polytechnique de Paris, France

25th Anniversary Eurandom - April 2024

Personal Background

Currently: Postdoctoral Researcher at Ecole Polytechnique with Aymeric Dieuleveut, working on Federated Learning

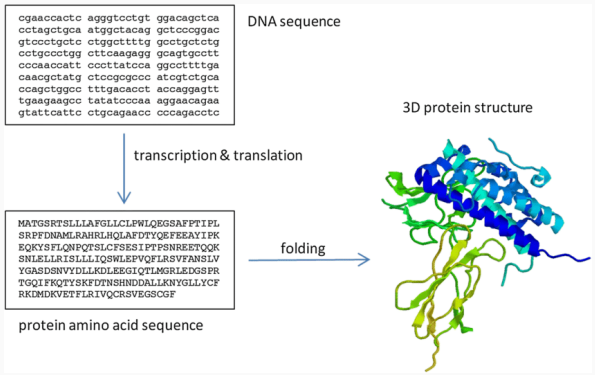
Research interests: Monte Carlo methods, Stochastic Optimization, Federated Learning, Reinforcement Learning

Education:

- PhD in Machine Learning, Télécom Paris, Oct.2019-Mar.2023
"Monte Carlo Methods and Stochastic Approximation" (François Portier)
- MSc in Machine Learning (Master MVA), ENS Paris-Saclay, 2018-2019
- MSc in Applied Maths and Computer Science, Télécom Paris, 2016-2019

Motivation: Machine Learning recent advances

AlphaGo (2016)



AlphaFold (2018)

GPT-3/4(2020/2023)

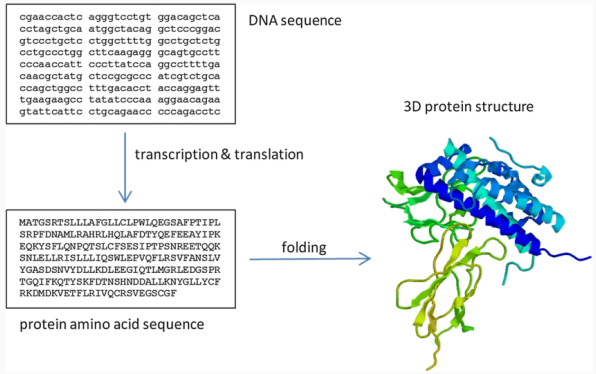
"Intelligence"

=

Data + Models + Algorithms + Computing Power

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 $\mathcal{F}$

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

Main issue: intractability and computational cost

Motivation: need for integral estimators

Central Question: *Integration*

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

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

Main issue: intractability and computational cost

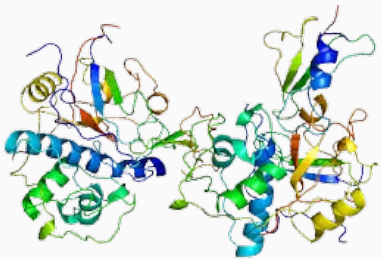
• **(RL)** Trajectory $\tau = (s_0, a_0, \ldots, s_{T-1}, a_{T-1})$ with policy π_θ and cumulative return $\mathcal{R}(\tau) = \sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$.

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

(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

$\rightarrow$ Requires only three steps: sampling, evaluating, averaging

Randomness as a Strength

$\rightarrow$ Naturally escape local optima

$\rightarrow$ Complete exploration of the search space

Large-Scale learning

$\rightarrow$ simple, scalable, parallelizable

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

Theoretical justifications[1]

$\rightarrow$ deterministic methods $O(n^{-s/d})$

$\rightarrow$ 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

1. **R. Leluc**, F. Portier and J. Segers. *Control Variate Selection for Monte Carlo Integration.* (Leluc et al., 2021)
In *Statistics and Computing 31, 50*, pages1-27, 2021.

2. **R. Leluc**, F. Portier, J. Segers and A. Zhuman. *A Quadrature Rule combining Control Variates and Adaptive Importance Sampling.* (Leluc et al., 2022)
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

3. **R. Leluc**, F. Portier, J. Segers and A. Zhuman. *Speeding up Monte Carlo Integration: Control Neighbors for Optimal Convergence*. (Leluc et al., 2023)
Bernoulli, 2024.

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 $h_1, \dots, h_m \in L_2(\pi)$ 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 $h = (h_1, \dots, h_m)^\top$, 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

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

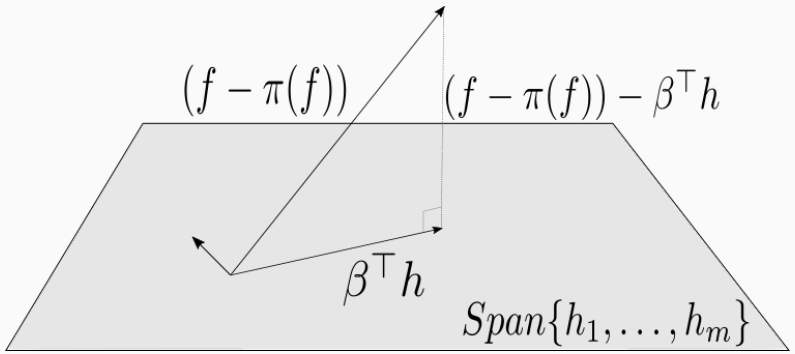
• What optimal choice for β^* ? Look at variance and define

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

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, $X_1, \ldots, X_n \sim \pi$

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

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 ?

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

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

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

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

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

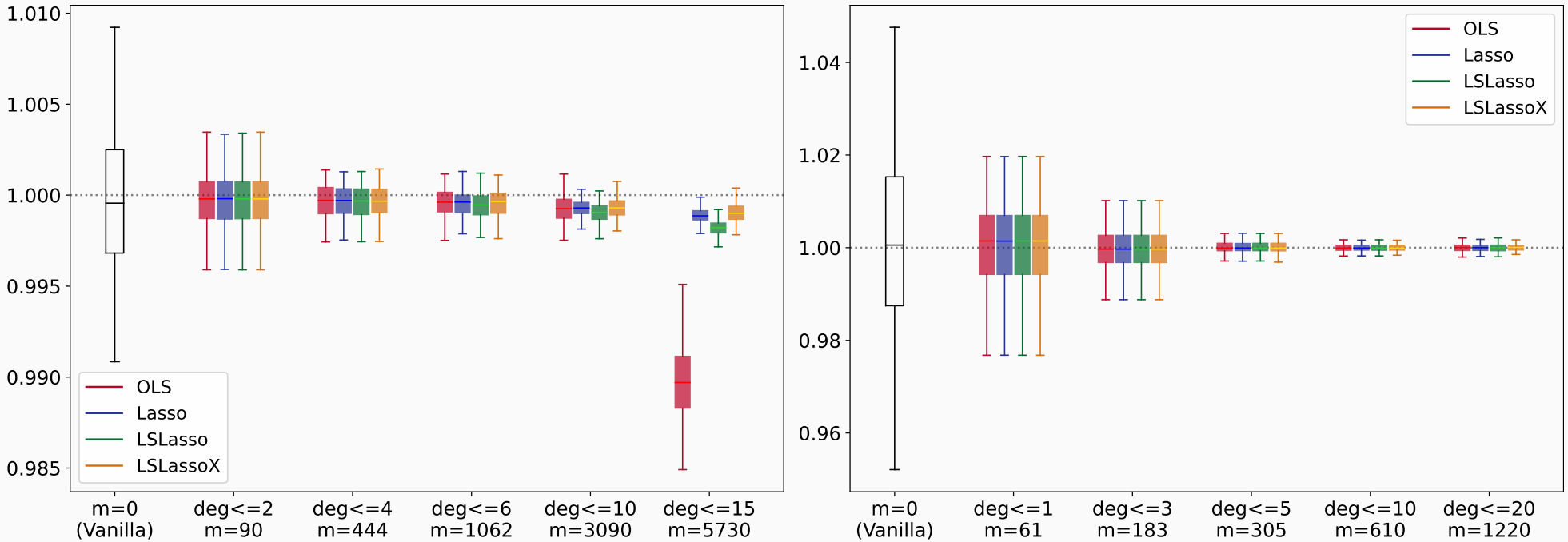
$G = \mathbb{E}_\pi[hh^\top], \gamma = \lambda_{\min}(G), \hbar = G^{-1/2}h; B = \sup_x \|\hbar(x)\|_2^2$

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

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)

Monte Carlo Integration and Importance Sampling

GOAL:

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x$$

Can we sample from target distribution π ?

Monte Carlo Integration and Importance Sampling

GOAL:

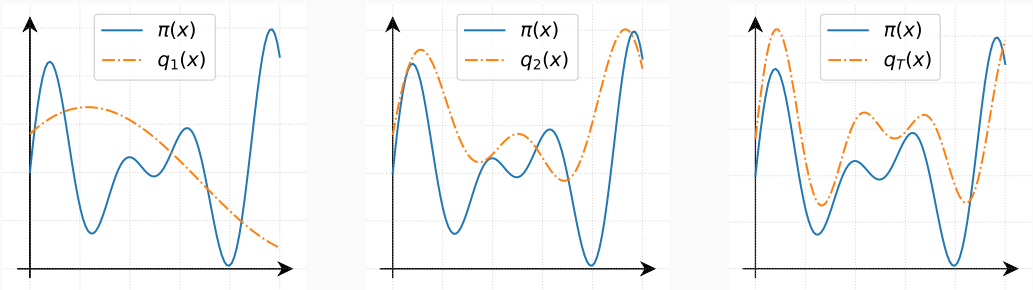
$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x$$

Can we sample from target distribution π ?

- YES, use naive Monte Carlo estimate (+ control variates)

$$\hat{\alpha}_n^{(\mathrm{mc})}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad X_1, \ldots, X_n \sim \pi$$

- NO, use **Adaptive Importance Sampling** with sampling policy $(q_i)_{i\geq 0}$



Evolution of sampling policy is AIS.

$$X_1 \sim q_0, \ldots, X_i \sim q_{i-1}$$

$$\hat{\alpha}_n^{(\mathrm{ais})}(f) = \frac{\sum_{i=1}^n w_i f(X_i)}{\sum_{i=1}^n w_i}$$

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

$$w_i = \pi(X_i)/q_{i-1}(X_i).$$

Adaptive Importance Sampling with Control Variates

AISCV estimate: Weighted Least Squares

Particles $X_i \sim q_{i-1}$ and weights $w_i = \pi(X_i)/q_{i-1}(X_i)$,

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

- (a) (Exact integration) whenever f 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., $\hat{\alpha}_n = \pi(f) = \int f\pi \,\mathrm{d}\lambda$.
- (b) (Quadrature Rule) $\hat{\alpha}_n = \sum_{i=1}^n v_{n,i} f(X_i)$, for **quadrature weights** $v_{n,i}$ **that do not depend on the function** f and that can be computed by a single weighted least squares procedure.
- (c) (Bayesian) it can be computed even when π **is known only up to a multiplicative constant**.
- (d) (post-hoc) 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**

Non-asymptotic error analysis

Residuals $\varepsilon = f - \alpha - \beta^\top h$ with $(\alpha, \beta) = \arg\min_{a,b} \int (f - a - b^\top h)^2 \pi \mathrm{d}\lambda$.

Assumptions

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

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

(A3)

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

Concentration inequality for AISCV estimate

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

$$\left|\hat{\alpha}_n^{(\mathrm{aiscv})}(f) - \pi(f)\right| \leq C_2 \sqrt{\log(10/\delta)} \frac{\tau}{\sqrt{n}} + C_3 cB \log(10m/\delta) \frac{\tau}{n},$$

where C_1, C_2, C_3 are some constants and $B = \sup_{x:\pi(x)>0} \|\hbar(x)\|_2^2$, $\hbar = G^{-1/2} h$.

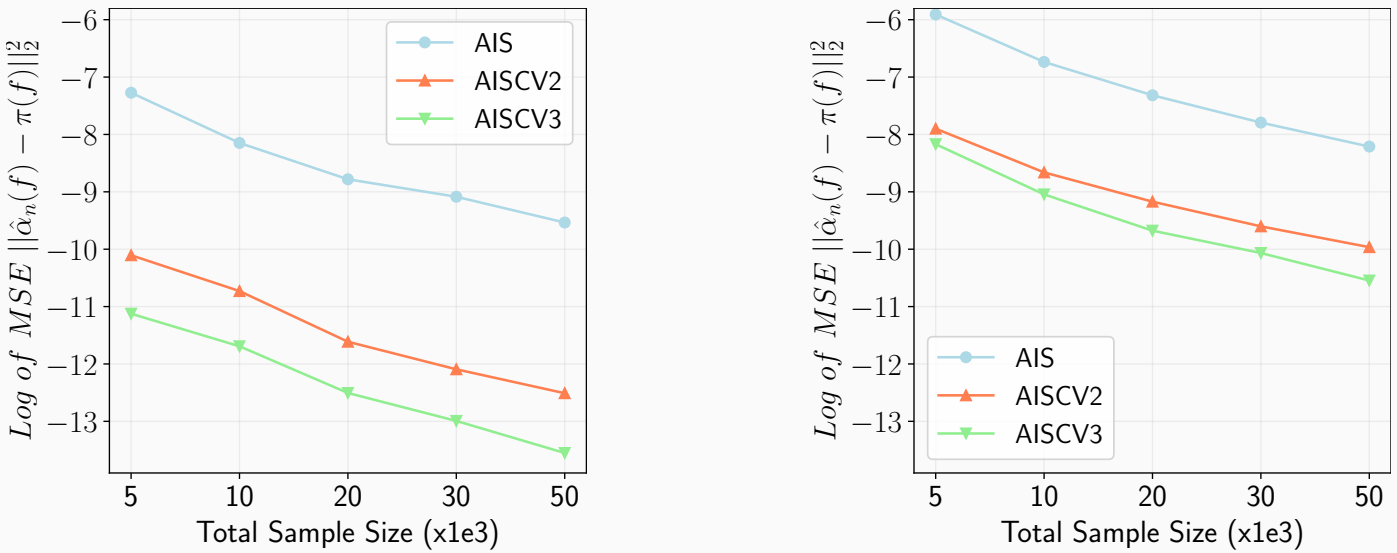
Synthetic examples: Gaussian Mixtures

Similar framework as Cappé et al. (2008).

Integrand and Target: $f(x) = x, \pi_\Sigma(x) = 0.5\Phi_\Sigma(x-\mu) + 0.5\Phi_\Sigma(x+\mu)$ where $\mu = (1, \ldots, 1)^\top / 2\sqrt{d}, \Sigma = I_d/d$ and Φ_Σ is pdf $\mathcal{N}(0, \Sigma)$.

Sampling policy: Multivariate Student

Control variates: Stein method with $\varphi =$ polynomial with bounded degree



Gaussian mixture density: Logarithm of $\|\hat{\alpha}_n(f) - \pi(f)\|_2^2$ for $f(x) = x$ with target isotropic π_Σ with $d = 4$ (left), $d = 8$ (right).

Complexity rates for integration error

Definition: Root Mean Squared Error (RMSE)

The error δ_n of a procedure $\hat{\alpha}_n(f)$ that approximates $\pi(f)$ is

$$\delta_n = \mathbb{E}\left[|\hat{\alpha}_n(f) - \pi(f)|^2\right]^{1/2}$$

$\rightarrow$ Lipschitz integrands[3], **optimal rate** in $O(n^{-1/2}n^{-1/d})$ (Novak, 2016)

OLS control variates (Portier and Segers, 2019)	$O(n^{-1/2}m^{-1/d})$
Determinantal sampling (Bardenet and Hardy, 2020)	$O(n^{-1/2}n^{-1/2d})$
Control Functionals (Oates et al., 2017)	$O(n^{-7/12})$
Cubic Stratification (Haber, 1966; Chopin and Gerber, 2024)	$O(n^{-1/2}n^{-1/d})$

[3]for integrand with s bounded derivatives, rate in $O(n^{-1/2}n^{-s/d})$

General view of Control Variates

Control Functionals

- Build surrogate function $\hat{f}$ with known integral $\pi(\hat{f})$
- Use centered variables $\hat{f}(X_i) - \pi(\hat{f})$ to derive the following enhanced Monte Carlo estimate with control variates

$$\hat{\alpha}_n^{(CV)}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i) - \left(\hat{f}(X_i) - \pi(\hat{f})\right)\right\}$$

Approximation in $L_2(\pi)$

Let $(X_1, \dots X_n) \sim \pi$. Suppose that $\hat{f}$ depends only on a surrogate sample $\tilde{X}_1, \dots, \tilde{X}_N$ which is independent from $(X_1, \dots X_n)$, then

$$\mathbb{E}\left[|\hat{\alpha}_n^{(CV)}(f) - \pi(f)|^2\right] \leq \frac{1}{n}\,\mathbb{E}\left[\int (f - \hat{f})^2 \mathrm{d}\pi\right].$$

Control Functionals examples

• **RKHS approximation:** (Oates, Girolami, and Chopin, 2017)

Ridge regression in Hilbert space $\mathcal{H}$

$$\hat{f} \in \arg\min_{\varphi \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} (f(\tilde{X}_i) - \varphi(\tilde{X}_i))^2 + \lambda \|\varphi\|_{\mathcal{H}}^2$$

• **Basis functions:** (Portier and Segers, 2019; Leluc et al., 2021)

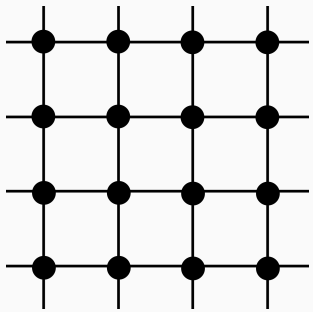
Use m basis functions $h_1, \ldots, h_m$ to fit OLS:

$$\hat{f} = \hat{\beta}_n^\top h, \qquad (\hat{\alpha}_n, \hat{\beta}_n) = \arg\min_{(\alpha,\beta) \in \mathbb{R} \times \mathbb{R}^m} \|f^{(n)} - \alpha \mathbb{1}_n - H\beta\|_2^2$$

• **Partitioning and Stratification:** (Chopin and Gerber, 2024)

$(\tilde{X}_1, \ldots, \tilde{X}_N)$ is the $(1/\ell)$-equidistant grid of $[0,1]^d$ with $N = \ell^d$, $\ell \geq 1$ and $(R_i)_{i=1,\ldots,N}$ is the partition of $[0,1]^d$ made of the rectangles.

$$\hat{f}(x) = \sum_{i=1}^{N} f(\tilde{X}_i) \mathbb{1}_{R_i}(x)$$



Nearest Neighbors

Control Neighbors

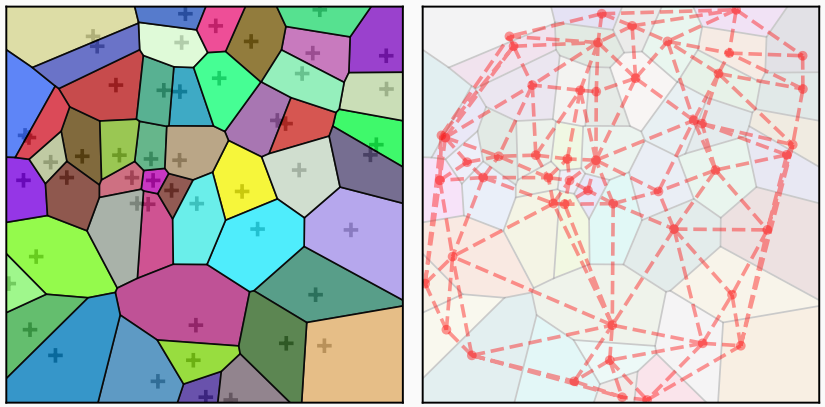
$$\hat{\alpha}_n^{(CVNN)}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i) - \left(\hat{f}_n^{(i)}(X_i) - \pi(\hat{f}_n)\right)\right\}$$

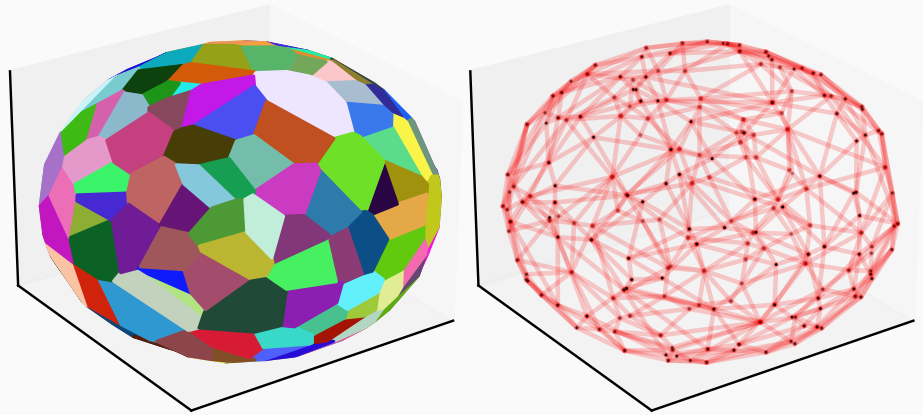
Leave-one-out Nearest Neighbors:

Take same sample $(X_1, \ldots, X_n)$ and define

$$\hat{f}_n(x) = \sum_{j=1}^{n} f(X_j)\mathbb{1}_{S_{n,j}}(x), \qquad \hat{f}_n^{(i)}(x) = \sum_{j\neq i} f(X_j)\mathbb{1}_{S_{n,j}^{(i)}}(x)$$

where $S_{n,j}$ are **Voronoï cells**





Control Neighbors properties

Control Neighbors

$$\hat{\alpha}_n^{(CVNN)}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i) - \left(\hat{f}_n^{(i)}(X_i) - \pi(\hat{f}_n)\right)\right\}$$

• (a) (Same framework as naive MC) does not require the existence of control variates with known integrals

• (b) (Quadrature Rule) $\hat{\alpha}_n = \sum_{i=1}^{n} w_{n,i} f(X_i)$, for **quadrature weights** $w_{n,i}$ **that do not depend on the function** f.

• (c) (Practical tool box) The weights $w_{n,i}$ are built using efficient nearest neighbors estimates (Bentley, 1975; Pedregosa et al., 2011)

Complexity rate for integration error of Control Neighbors

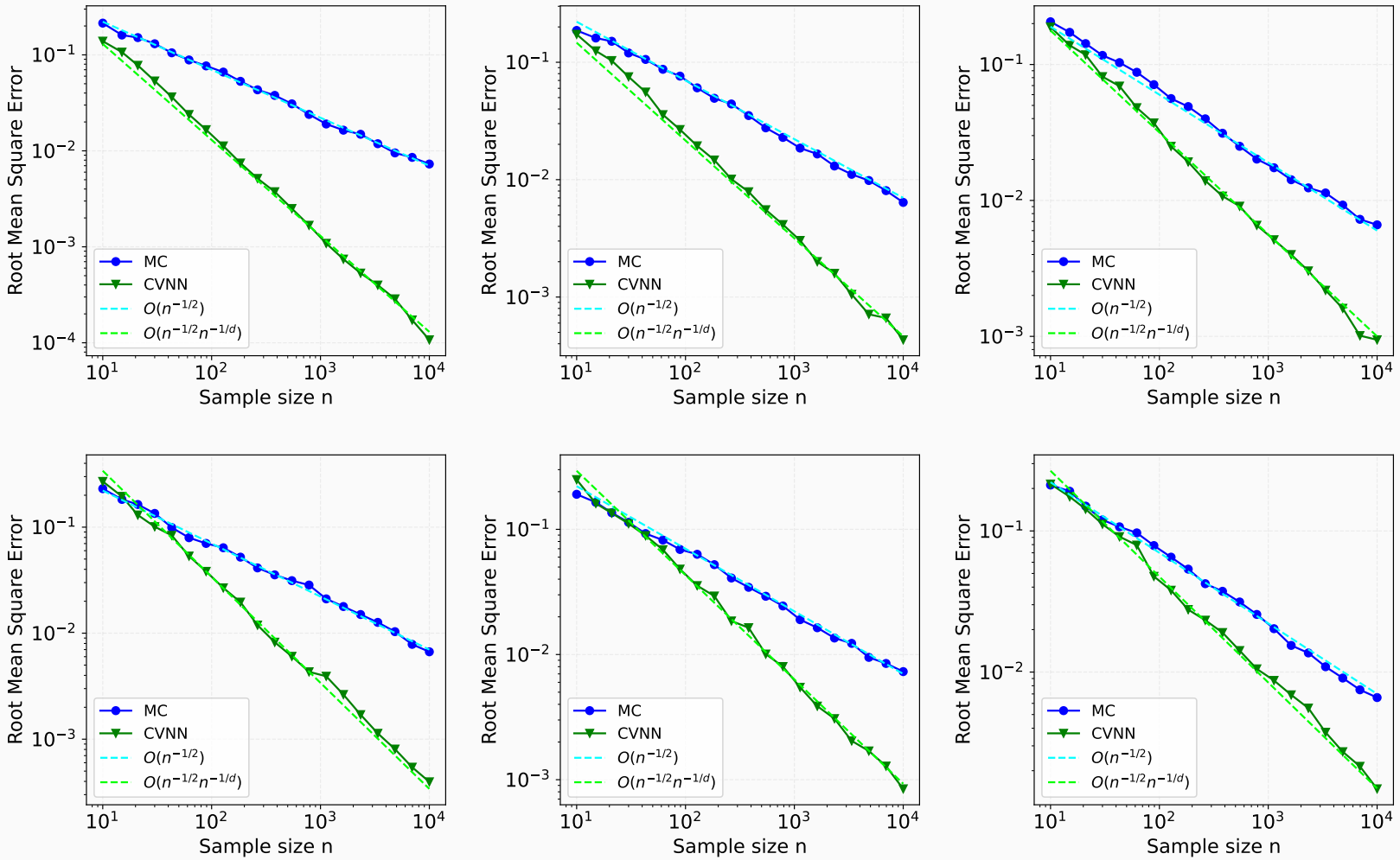
$$\mathbb{E}\left[|\hat{\alpha}_n^{(CVNN)}(f) - \pi(f)|^2\right]^{1/2} \lesssim n^{-1/2}n^{-s/d}$$

$$|\hat{\alpha}_n^{(CVNN)}(f) - \pi(f)| \lesssim \sqrt{\log(1/\varepsilon)}(\log n)^{1+s/d}n^{-1/2}n^{-s/d}$$

(with proba greater than $1-\varepsilon$)

Control Neighbors on synthetic integrands

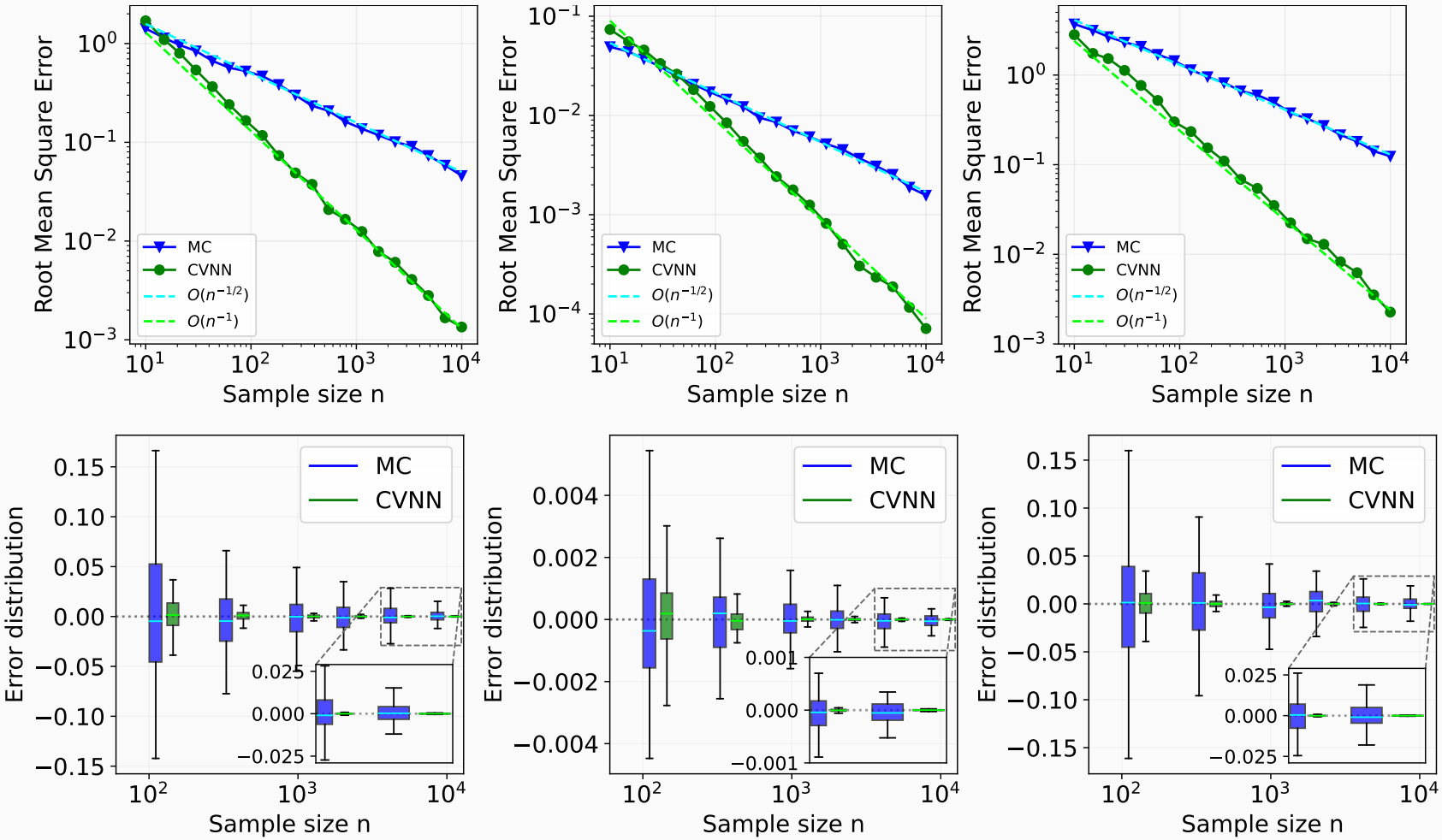
- $f_1(x_1, \ldots, x_d) = \sin(\pi(\frac{2}{d}\sum_{i=1}^{d} x_i - 1))$ with $\pi = \mathbb{1}_{[0,1]^d}$
- $f_2(x_1, \ldots, x_d) = \sin(\frac{\pi}{d}\sum_{i=1}^{d} x_i)$ with $\pi = \mathcal{N}_d(0, I_d)$



Error curves for f_1(top) and f_2(bottom) with $d \in \{2; 3; 4\}$

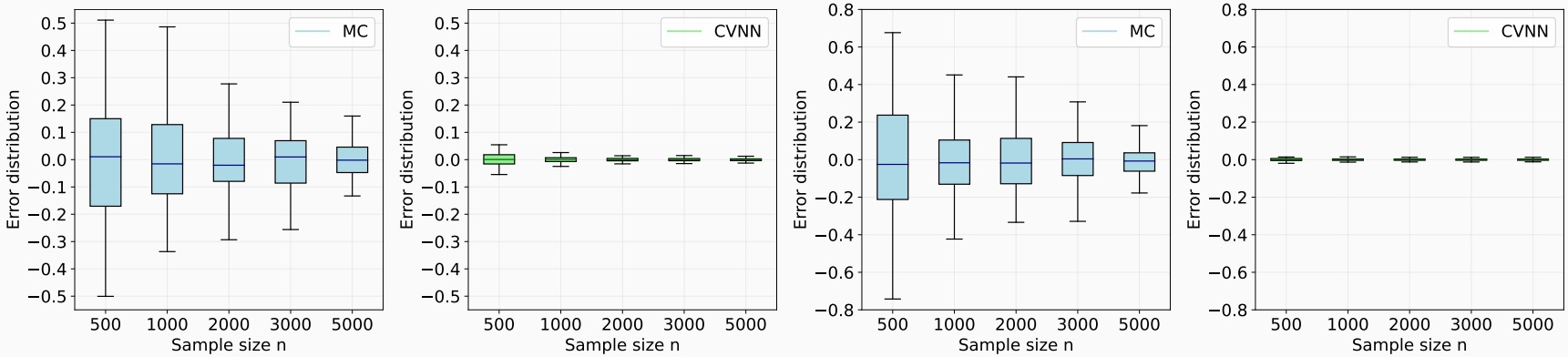
Control Neighbors on Sphere $\mathbb{S}^2$

$f_3(x,y,z) = \cos(x+y+z)$, $f_4(x,y,z) = \cos(x)\cos(y)\cos(z)$, $f_5(x,y,z) = \exp(x-y)$

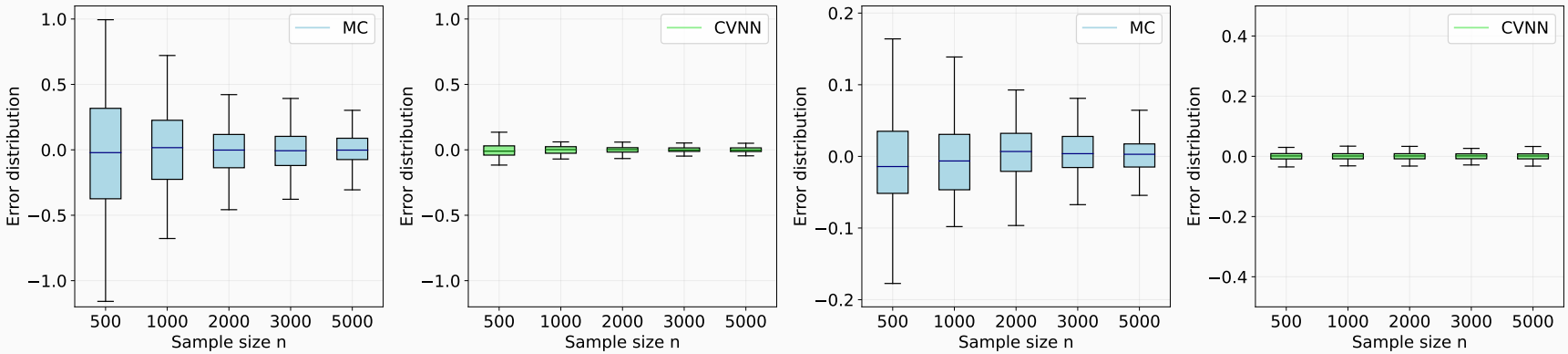


RMSE and boxplots for f_3(left), f_4(center) and f_5(right)

Control Neighbors for Option Pricing



Black-Scholes model with spot price $S_0 = 100$, strike $K = S_0$, maturity $T = 2$ months, risk-free rate $r = 0.1$, constant volatility $\sigma = 0.3$, barrier price $H = 130$.(left: "Up-In"/right: "Up-Out")



Heston Model with spot ptice $S_0 = 100$, strike $K = S_0$, barrier price $H = 130$, maturity $T = 2$ months, risk-free rate $r = 0.1$, initial volatility $v_0 = 0.1$, long-run average variance $\theta = 0.02$, rate of mean reversion $\kappa = 4$, instanteneous correlation $\rho = 0.8$ and volatility of volatility $\xi = 0.9$.(left: "Up-In"/right: "Up-Out")

Thank You and Happy 25th Birthday

References

Bardenet, R. and A. Hardy (2020). Monte carlo with determinantal point processes. *The Annals of Applied Probability 30*(1), 368–417.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM 18*(9), 509–517.

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

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

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

Bibliography ii

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.

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.

Haber, S. (1966). A modified monte-carlo quadrature. *Mathematics of Computation 20*(95), 361–368.

Leluc, R., F. Portier, and J. Segers (2021, 07). Control variate selection for Monte Carlo integration. *Statistics and Computing 31*.

Leluc, R., F. Portier, J. Segers, and A. Zhuman (2022). A Quadrature Rule combining Control Variates and Adaptive Importance Sampling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 11842–11853. Curran Associates, Inc.

Leluc, R., F. Portier, J. Segers, and A. Zhuman (2023). Speeding up monte carlo integration: Control neighbors for optimal convergence. *arXiv preprint arXiv:2305.06151*.

Bibliography iii

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.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

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

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

Appendix

Control Variate Selection

Assumptions.

- $\varepsilon \in \mathcal{G}(\tau^2) : \log \mathbb{E}[\exp(\lambda\varepsilon)] \leq \lambda^2\tau^2/2$ for all $\lambda \in \mathbb{R}$
- Uniformly bounded control variates
- Linearly independent control variates, $G := \pi(hh^\top)$ positive deinite
- (Restricted Eigenvalue) There exists $\gamma^\star > 0$ such that: $u^\top G u \geq \gamma^\star \|u\|_2^2$ for all $u \in \mathcal{C}(S^\star; 3)$

$$\mathcal{C}(S, \alpha) = \{u \in \mathbb{R}^m : \|u_{\bar{S}}\|_1 \leq \alpha \|u_S\|_1\}$$

- Orthogonality $\pi(h_j h_k) = 0$ for $k \in S^\star, j \in \{1, \ldots, m\} \setminus S^\star$

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 ($n = 5000$)

AISCV algorithm

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

1: **for** $t = 1, \ldots, T$ **do**

2: Generate an independent random sample $X_{t,1}, \ldots, X_{t,n_t}$ from q_{t-1}

3: Compute weights $(w_{t,i})_{i=1}^{n_t}$ where $w_{t,i} = \pi(X_{t,i})/q_{t-1}(X_{t,i})$

4: Construct the matrix of control variates $H_t = \left(h_j(X_{t,i})\right)_{i=1,\ldots,n_t}^{j=1,\ldots,m}$

5: Evaluate the integrand in the particles: $(f(X_{t,i}))_{i=1}^{n_t}$

6: Update q_t based on the past $(X_{s,i} : s = 1, \ldots, t; i = 1, \ldots, n_s)$

7: **end for**

8: $(\hat{\alpha}_T, \hat{\beta}_T) = \underset{(a,b)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \left\{ \sum_{t=1}^T \sum_{i=1}^{n_t} w_{t,i} \left(f(X_{t,i}) - a - b^\top h(X_{t,i}) \right)^2 \right\}$

9: $I_n^{(\mathrm{aiscv})}(f) = \hat{\alpha}_T$.

AISCV post-hoc algorithm

AISCV algorithm - *post-hoc*

Require: integrand f, $T \in \mathbb{N}^*$, allocation policy $(n_t)_{t=1}^T$, weights $(w_t)_{t=1}^T$ with $w_t = (w_{t,i})_{i=1}^{n_t}$, matrices $(H_t)_{t=1}^T$ with $H_t = \left(h_j(X_{t,i})\right)_{i=1,\dots,n_t}^{j=1,\dots,m}$, particles $(X_{t,i} : t = 1, \dots, T; i = 1, \dots, n_t)$

1: Compute $\hat{\beta}_n(\mathbb{1}_n) = \arg\min_{b \in \mathbb{R}^m} \sum_{t=1}^T \sum_{i=1}^{n_t} w_{t,i} \left(1 - b^\top h(X_{t,i})\right)^2$

2: Compute $u_t = diag(w_t)[\mathbb{1}_{n_t} - H_t \hat{\beta}_n(\mathbb{1}_n)]$ for $t = 1, \dots, T$

3: Compute $s = \sum_{t=1}^T \sum_{i=1}^{n_t} u_{t,i}$

4: Compute weights $v_{t,i} = u_{t,i}/s$ for $t = 1, \dots, T$ and $i = 1, \dots, n_t$

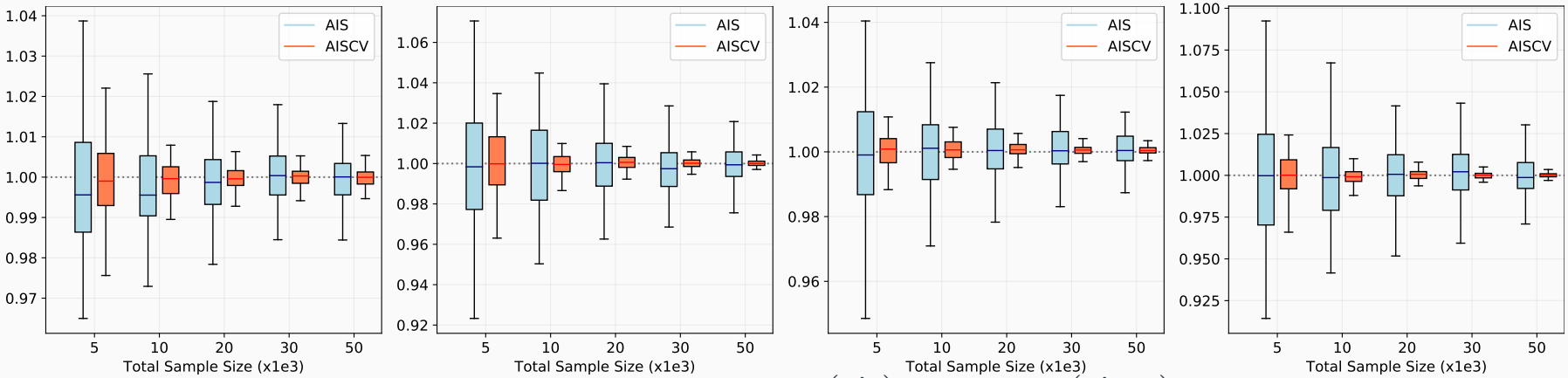
5: $I_T^{(\mathrm{aiscv})}(f) = \sum_{t=1}^T \sum_{i=1}^{n_t} v_{t,i} f(X_{t,i})$

AISCV: synthetic functions on $[0,1]^d$

- Uniform density $\pi(x) = 1$ for $x \in [0,1]^d$ in dimensions $d \in \{4; 8\}$.

$f_1(x) = \sin(\pi(\frac{2}{d}\sum_{i=1}^d x_i - 1)); f_2(x) = \prod_{i=1}^d (2/\pi)^{1/2} \frac{\mathrm{e}^{-\log(x_i)^2/2}}{x_i}; f_3(x) = \prod_{i=1}^d \log(2) 2^{1-x_i}$

- Legendre polynomials: $m = 240 (d = 4)$ and $m = 1056 (d = 8)$



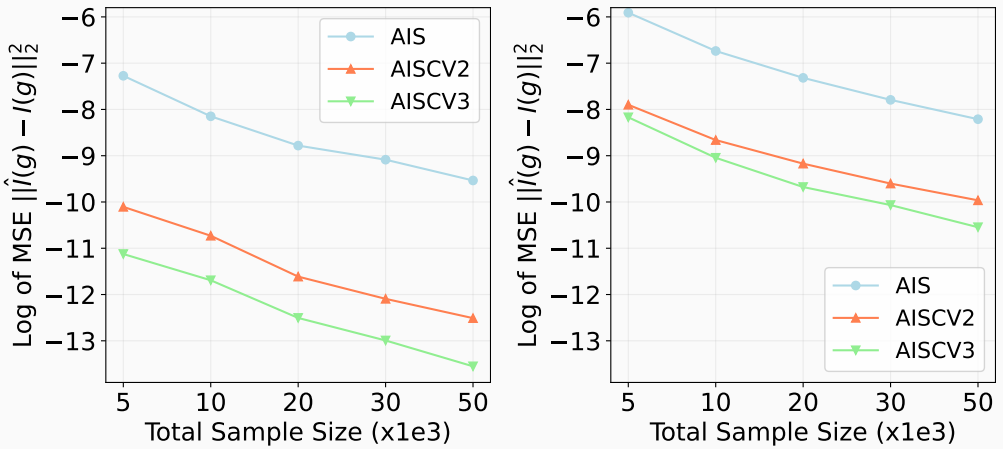
Integration on $[0,1]^d$: boxplots of estimates $\hat{\alpha}_n^{(\mathrm{ais})}(f)$ and $\hat{\alpha}_n^{(\mathrm{aiscv})}(f)$ with integrands $f_1(d=4); f_1(d=8); f_2(d=4); f_3(d=8)$ obtained over 100 replications.

AISCV: synthetic functions

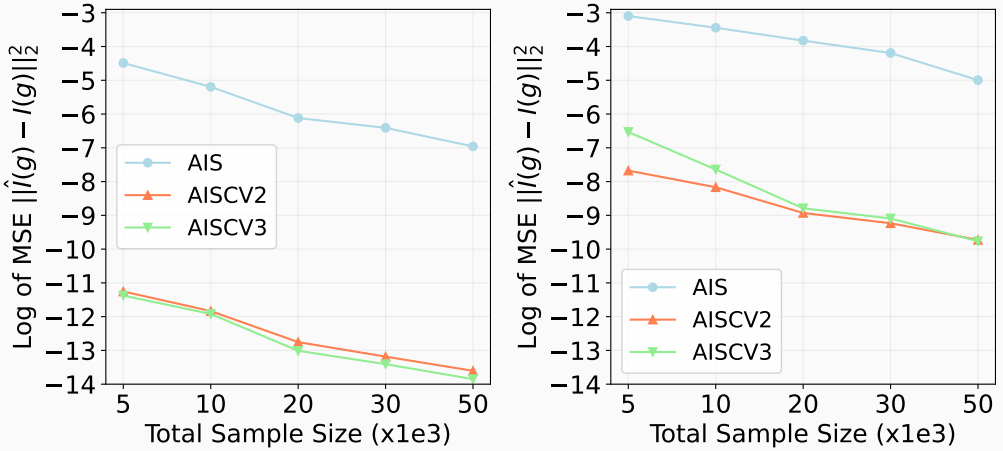
Integrand	Efficiency	Sample Size n $5,000$	$10,000$	$20,000$	$30,000$	$50,000$
f_1	standard	2.97	7.87	7.56	7.81	9.64
$(d = 4)$	global	0.76	1.88	1.63	1.53	1.47
f_1	standard	2.70	14.3	20.7	30.7	41.8
$(d = 8)$	global	0.12	0.63	0.96	1.65	2.10
f_2	standard	11.0	12.6	15.5	22.7	20.7
$(d = 4)$	global	9.90	10.7	12.6	18.0	15.9
f_3	standard	9.12	37.1	51.8	78.4	102
$(d = 8)$	global	2.52	10.6	14.3	21.3	26.2

Standard and global efficiencies for AISCV compared to AIS for f_1, f_2, f_3 in dimensions $d \in \{4; 8\}$ obtained over 100 replications.

AISCV: gaussian mixtures



Gaussian mixture density: Logarithm of $\|\hat{\alpha}(f) - \pi(f)\|_2^2$ for $f(x) = x$ with and target **isotropic** and $d = 4$ (left), $d = 8$ (right).



Gaussian mixture density: Logarithm of $\|\hat{\alpha}(f) - \pi(f)\|_2^2$ for $f(x) = x$ with and target **anisotropic** and $d = 4$ (left), $d = 8$ (right).

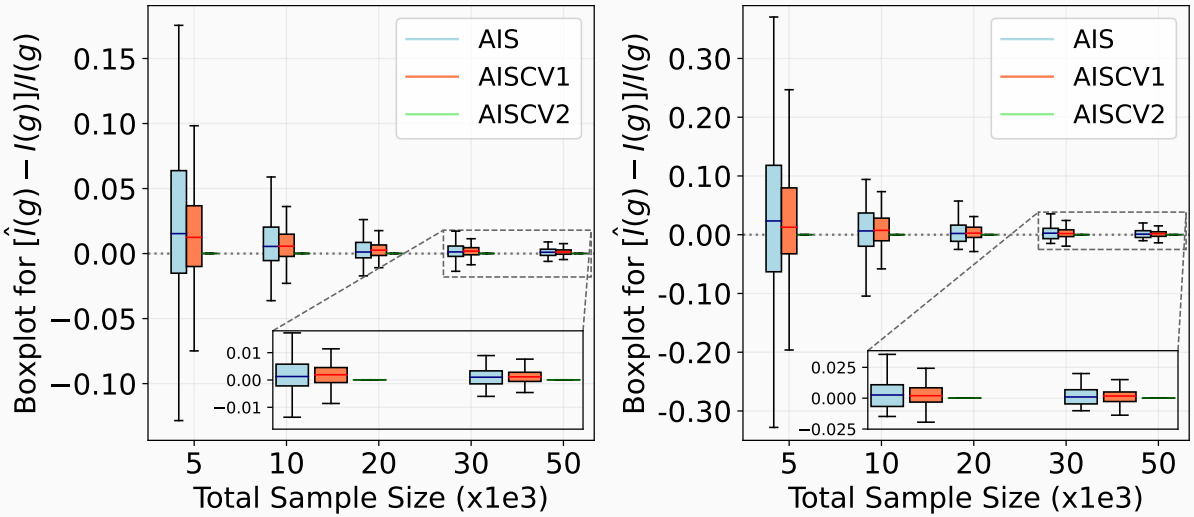
AISCV: Bayesian Linear Regression

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: $f(\theta) = \sum_{i=1}^{d} \theta_i^2$.

Control variates: Stein control variates with $\varphi_\alpha(\theta) = \theta_1^{\alpha_1} \cdots \theta_d^{\alpha_d}$, $\alpha_1 + \cdots + \alpha_d \leq Q$, $Q \in \{1; 2\}$.



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

AISCV: efficiency of Bayesian Linear Regression

Dataset	Efficiency	Sample Size n $5,000$	$10,000$	$20,000$	$30,000$	$50,000$
Housing	standard	7.60	6.77	19.3	17.2	53.0
	global	3.24	3.26	9.39	8.38	26.0
Abalone	standard	10.4	21.3	23.6	21.1	17.3
	global	5.63	12.2	13.5	12.0	9.85
Red Wine	standard	8.25	9.25	8.03	7.33	6.49
	global	3.84	4.66	4.01	3.66	3.24
White Wine	standard	1.60	1.74	2.06	2.03	1.96
	global	0.77	0.88	1.05	1.04	1.01

Standard and global efficiencies for AISCV1 compared to AIS for Bayesian Linear Regression on real-world datasets obtained over 100 replications.