Speeding-up Monte Carlo: Nearest Neighbors estimates as Control Variates

François Portier

ENSAI, CREST

September, 2022

Joint work with Rémi Leluc, Johan Segers and Aigerim Zhuman

CREST
CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

ENSAI
École nationale de la statistique et de l'analyse de l'information

Introduction : Why bother with random sampling?

A guided tour in Monte Carlo

- Sampling techniques
  - Importance sampling and MCMC
  - Quasi-Monte Carlo
  - Determinantal sampling
- Post-hoc scheme
  - Adaptive volume calculation
  - Control variates

Nearest neighbor as control functionals

- Control functional
- Nearest neighbor background
- Construction of the estimate

Numerical illustration

The underlying integration problem

Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be integrable.

- **Goal** : Estimate

$$\mu(\varphi) = \int \varphi \, \mathrm{d}\mu$$

- **Constraint:** only based on $\varphi(x_1), \ldots, \varphi(x_n)$, where $x_1, \ldots, x_n$ are called nodes. Here φ might be black-box function[1].

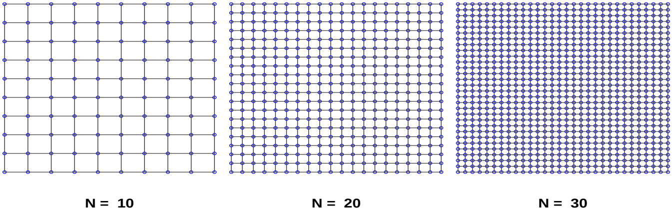
- **Central question:** number of nodes n necessary to obtain a given accuracy

[1] if φ has an explicit form, e.g., $\varphi(x) = \exp(-\|x\|^2)$, then some approximation techniques are probably more appropriate

Riemann's sums method for $\int_{[0,1]^d} \varphi(x)\,\mathrm{d}x$:

$$N^{-d} \sum_{x_i \in \text{Grid}} \varphi(x_i),$$

where $\text{Grid} = \{(i_1/N, \dots, i_d/N) \,:\, 1 \leq i_k \leq N,\ \forall k = 1, \dots, d\}$



Error bound

We have

$$\sup_{\varphi \in \Phi_d} \left| N^{-d} \sum_{x \in \text{Grid}} \varphi(x) - \int_{[0,1]^d} \varphi(x)\,\mathrm{d}x \right| \leq N^{-1}.$$

with $\Phi_d = \left\{ \varphi : [0,1]^d \mapsto \mathbb{R} \,:\, |\varphi(x) - \varphi(y)| \leq \max_{k=1,\dots,d} |x_k - y_k| \right\}$

Consider linear integration rules

$$\sum_{i=1}^{N^d} w_i \varphi(x_i).$$

The accuracy of the best algorithm over a class Φ_d is

$$e(N^d, \Phi_d) = \inf_{(w_i, x_i)_{i=1\ldots N^d}} \sup_{\varphi \in \Phi_d} \left| \sum_{i=1}^{N^d} w_i \varphi(x_i) - \int_{[0,1]^d} \varphi(x)\, \mathrm{d}x \right|$$

Complexity results (Novak, 2016)

$$e(N^d, \Phi_d) = \left(\frac{d}{2d+2} \right) N^{-1}$$

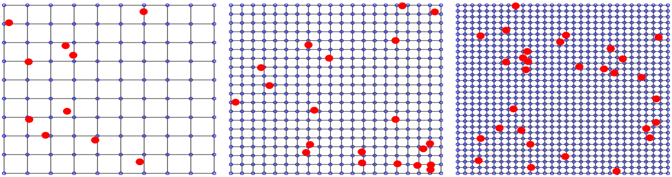
The midpoint rule is the optimal algorithm[2].

[2] If $\Phi_{k,d} = \{\varphi : [0,1]^d \to \mathbb{R}\,,\ \|D_\alpha \varphi\|_\infty \leq 1, \forall |\alpha| \leq k\}$, then $e(N^d, \Phi_{k,d}) \simeq N^{-k}$.

Monte Carlo method for $\int_{[0,1]^d} \varphi(x)\,\mathrm{d}x$:

Let $(X_1, \ldots, X_n) \stackrel{iid}{\sim} \mathcal{U}[0,1]^d$, compute

$$n^{-1} \sum_{i=1}^{n} \varphi(X_i)$$



n= 10 n= 20 n= 30

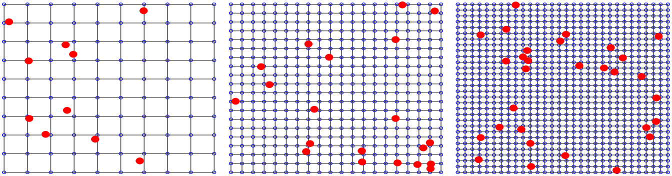
Uniform results (Talagrand, 1996; McDiarmid, 1998; Giné and Guillou, 2001)

with probability larger than $1-\delta$,

$$\sup_{\varphi\in\Phi} \left| n^{-1} \sum_{i=1}^{n} \varphi(X_i) - \int_{[0,1]^d} \varphi(x)\,\mathrm{d}x \right| \leq 2\mathbb{E}|R_n(\Phi)| + \sqrt{\frac{2\log(2/\delta)}{n}}$$

If for instance, Φ is of VC-type, $\mathbb{E}|R_n(\Phi)| \simeq n^{-1/2}$.

Summary



n= 10 n= 20 n= 30

	determisitic	random	Monte Carlo
$e(n, \Phi_d)$	$n^{-1/d}$	$n^{-1/d}$ $n^{-1/2}$	$n^{-1/2}$
$e(n, \Phi_d^k)$	$n^{-k/d}$	$n^{-k/d}$ $n^{-1/2}$	$n^{-1/2}$

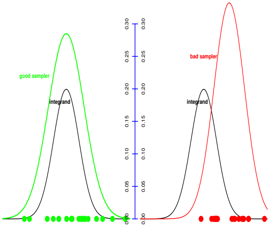
Monte Carlo

- Draw $X_1, \dots, X_n \overset{iid}{\sim} \mu$
- Compute $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$

MCMC

- Draw $X_1, \dots, X_n \rightsquigarrow \mu$
- Compute $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$

Importance sampling, stratified sampling...



Control variates

- Use the knowledge of $\int h_j d\mu = 0$ for functions $h_1, \dots, h_m$

Others

- Quasi-Monte Carlo
- DPP sampling

Books : Evans and Swartz (2000), Robert and Casella (2004), Glasserman (2003), Owen (2013a)

Sampling tool 1 (Importance sampling and MCMC)

A similar idea: sampling near target distribution

- (MCMC) $X_1, \ldots, X_n$ a Markov chain such that $(X_n) \rightsquigarrow \mu$

$$n^{-1} \sum_{i=1}^{n} g(X_i)$$

- (AIS) $X_i \sim q_{i-1}$ such that $q_i \to f$

$$n^{-1} \sum_{i=1}^{n} g(X_i)/q_{i-1}(X_i)$$

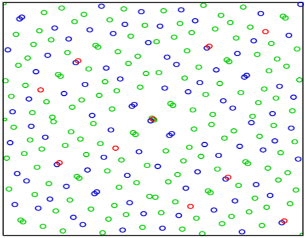
Highlights

- Relevant to Bayesian statistics
- the rate of convergence is not improved (only the asymptotic variance) (Robert and Casella, 2004; Evans and Swartz, 2000)

Sampling tool 2 (QMC)

Highlights

- Low-discrepancy sequences
- Using the Hardy-Krause variation of f
- Randomized version exists
- rate : $n^{-1}\log(n)^{d-1}$ (not under the same function class)



Issues

- Deterministic methods for the uniform measure $d\mu = \mathbb{I}_{[0,1]^d}$
- The bound decreases only when n is $\exp(d)$
- Hardy-Krauss variation is difficult to handle in practice

Sampling tool 3 (DPP)

A "random" quadrature rule

- Suppose that you have h_k such that $\int \varphi_k \varphi_j d\mu = \delta_{k,j}$ and define

$$K_n(x,y) = \sum_{i=1}^{n} \varphi_i(x)\varphi_i(y)$$

- $X_1, \dots, X_n$ follows a DPP with kernel K_N and reference measure μ. The estimate is

$$\sum_{i=1}^{n} \frac{g(X_i)}{K_N(X_i, X_i)}$$

Results (Bardenet and Hardy, 2020)

- unbiased
- rate : $n^{-1/2} n^{-1/2d}$

Issues

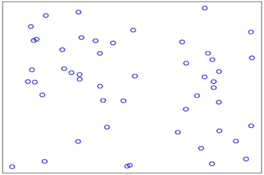
- Hard to sample from DPP (n^3 operations last time I checked)
- φ_k might not be known as it depends on μ

Post-hoc scheme 1: volume calculation

Integration problem

- $x_1, \ldots, x_n$ random points
- Observe $(x_1, g(x_1)), \ldots, (x_n, g(x_n))$
- Goal : Evaluate $\int g(x)dx$

$x_1, \ldots, x_n$ in $[0,1]^2$ with uniform law



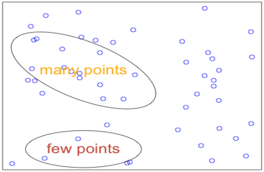
Monte-Carlo: $n^{-1} \sum_{i=1}^{n} g(x_i)$

Post-hoc scheme 1: volume calculation

Integration problem

- $x_1, \dots, x_n$ random points
- Observe $(x_1, g(x_1)), \dots, (x_n, g(x_n))$
- Goal : Evaluate $\int g(x)dx$

$x_1, \dots, x_n$ in $[0,1]^2$ with uniform law



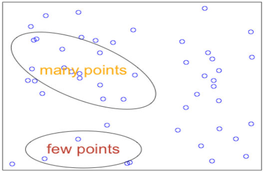
Monte-Carlo: $n^{-1} \sum_{i=1}^{n} g(x_i)$

Post-hoc scheme 1: volume calculation

Integration problem

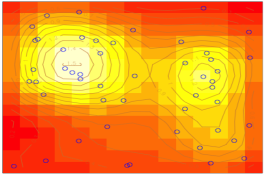
- $x_1, \ldots, x_n$ random points
- Observe $(x_1, g(x_1)), \ldots, (x_n, g(x_n))$
- Goal : Evaluate $\int g(x)dx$

$x_1, \ldots, x_n$ in $[0,1]^2$ with uniform law



Monte-Carlo: $n^{-1} \sum_{i=1}^{n} g(x_i)$

$\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} K(x - x_i)$



$n^{-1} \sum_{i=1}^{n} \frac{g(x_i)}{\widehat{f}(x_i)}$

(Delyon and Portier, 2016)

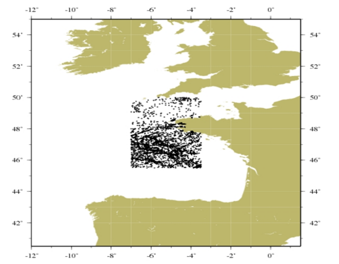
Advantages

- x_i's distribution is not used
- fast rates $n^{-1/2} n^{-(k-d)/2(k+d)}$
- robust to dependent x_i's

Difficulties

- computing time is n^2
- choice of the bandwidth
- dimension curse $k > d$

Evaluate average temperature of oceans (Azaïs et al., 2018)



Initial project was: use Voronoi cells volume to build the estimate (rate in $n^{-k/d}$)

Post-hoc scheme 2: Control variate

Idea Glasserman (2004); Owen (2013b)

- Use the knowledge of $h_1, \ldots, h_m$ such that

$$\int h_k d\mu = 0 \qquad k = 1, \ldots, m$$

- Let $X_1, \ldots, X_n$ be iid with common distribution μ

$$n^{-1} \sum_{i=1}^{n} \{g(X_i) - \sum_{k=1}^{m} \beta_k h_k(X_i)\}$$

First properties

- Unbiased property
- variance reduction up to $\min_{\beta \in \mathbb{R}^m} \mathbb{E}[(g(X_1) - \sum_{k=1}^m \beta_k h_k(X_1))^2]$

Issues

- Construction of h_k
- Computation of $\beta_1, \ldots, \beta_m$

Introduction : Why bother with random sampling?

A guided tour in Monte Carlo

- Sampling techniques
  - Importance sampling and MCMC
  - Quasi-Monte Carlo
  - Determinantal sampling
- Post-hoc scheme
  - Adaptive volume calculation
  - Control variates

Nearest neighbor as control functionals

- Control functional
- Nearest neighbor background
- Construction of the estimate

Numerical illustration

Control functional (Oates et al., 2017; Portier and Segers, 2019)

(i) building a function $\hat{g}$ of which the integral $\mu(\hat{g})$ is known
(ii) using this function to derive an enhanced Monte-Carlo estimate with the centered random variables $[\hat{g}(X_i) - \mu(\hat{g})]$ as

$$\hat{\mu}_n^{(CV)}(g) = \frac{1}{n} \sum_{i=1}^{n} \{g(X_i) - (\hat{g}(X_i) - \mu(\hat{g})\} .$$

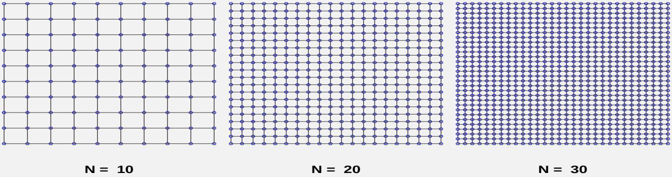
First property

Whenever the function $\hat{g}$ is constructed from another sample $\tilde{X}_1, \ldots \tilde{X}_n$ being either independent from $X_1, \ldots X_n$ or not random,

$$\mathbb{E}[(\hat{\mu}_n^{(CV)}(g) - \mu(g))^2] = \frac{1}{n} \int \mathbb{E}[(g - \hat{g})^2] d\mu$$

Example 1

Partitioning estimate



- μ is the uniform measure on $[0,1]^d$ and $\mathcal{G}$ is a regular grid with $n = N^d$ elements
- Define

$$\hat{g} = \text{pieceswise constant over elements of the grid}$$

Standard results give $\sqrt{\int (g - \hat{g})^2 \mathrm{d}\mu} = O(n^{-1/d})$

- implying an integration method with rate

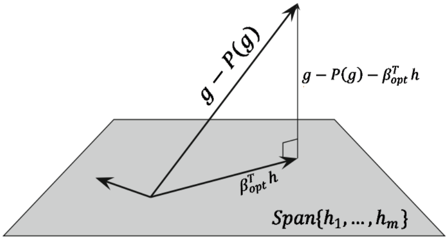
$$n^{-1/2} n^{-1/d}$$

(restrictive constraint $n = N^d$ plus $2n$ evaluations are needed)

Example 2

Ordinary least-squares

- Relying on two different samples, (Oates et al., 2017) propose to (a) build an RKHS control variate $\hat{g}$ and (b) compute the Monte Carlo average requires twice the number of request to g
- Using the same sample to $X_1, \ldots, X_n$ to approximate $\hat{g}$ has been investigated in Leluc et al. (2021) OLS is used to fit g with m basis functions. Theory says that the rates is in $n^{-1/2}m^{-1/d}$.
- Unfortunately, a constraint on m is needed (see for instance Remark 12 in Leluc et al. (2021)) which in general prevents from using $m = n$ control variates.



Definition (Nearest neighbor and distance)

Given a set of points $X_1, \ldots, X_n$ and any point $x \in \mathbb{R}^d$, define $\hat{N}_n(x)$ as the nearest neighbor of x among $X_1, \ldots, X_n$ and $\hat{\tau}_n(x)$ the associated distance, i.e.,

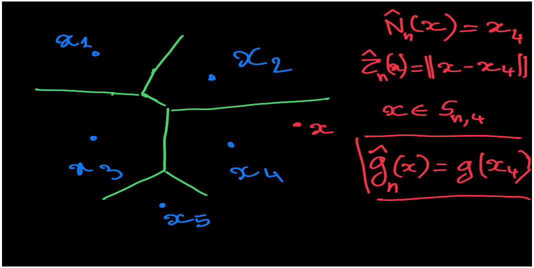
$$\hat{N}_n(x) \in \underset{Y \in \{X_1, \ldots, X_n\}}{\arg\min} \|x - Y\|, \qquad \hat{\tau}_n(x) = \|\hat{N}_n(x) - x\|.$$

Definition (Voronoï cells and volumes)

The Voronoï cells associated to $(X_i)_{i \geq 1}$ are given by

$$S_{n,i} = \{x \in \mathbb{R}^d : \hat{N}_n(x) = X_i\}.$$

Their volume with respect to μ is denoted by $V_{n,i} = \mu(S_{n,i})$.



Construction of the estimate

Definition (1-NN estimate of g)

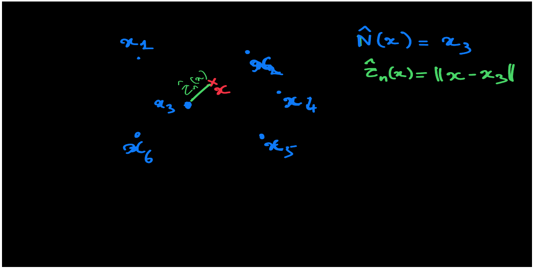
$$\forall x \in \mathbb{R}^d, \quad \hat{g}_n(x) = g(\hat{N}_n(x))$$

It is piece-wise constant on the Voronoï partition

$$\hat{g}_n(x) = \sum_{i=1}^{n} g(X_i) \mathbb{1}_{S_{n,i}}(x)$$

Main idea

- g is accessible without noise (no variance term) $\Rightarrow$ We take the 1-NN



PB 1

- The use of $\hat{g}_n$ as control functional leads to unsatisfactory strategy due to the over-fitting equation
$$\hat{g}_n(X_i) = g(X_i)$$

Solution 1

- Use the leave-one-out $\hat{g}_n^{(i)}(X_i)$ defined as the standard 1-NN except that the i-th observation has been removed

Following the previous idea we introduce

$$\hat{\mu}_n^{(\mathrm{NN-loo})}(g) = \frac{1}{n}\sum_{i=1}^{n}\{g(X_i) - (\hat{g}_n^{(i)}(X_i) - \mu(\hat{g}_n^{(i)}))\}, \tag{1}$$

Construction of the estimate

PB 2

- $\mu(\hat{g}_n^{(i)})$ implies to compute many integrals

Solution 2

- $\mu(\hat{g}_n^{(i)}) \simeq \mu(\hat{g}_n)$

The working estimate is then

$$\hat{\mu}_n^{(\mathrm{NN})}(g) = n^{-1} \sum_{i=1}^{n} \{g(X_i) - (\hat{g}_n^{(i)}(X_i) - \mu(\hat{g}_n))\}. \quad (2)$$

Degree and expected degree (isolation of point)

Denote by $S_{n,j}^{(i)}$ ($V_{n,j}^{(i)}$) (the volume of) the j-th Voronoï cell obtained from the sample $\mathcal{X}^{(i)} = \{X_1, \ldots, X_n\} \backslash X_i$.

Definition

The degree of point X_j is defined as

$$\hat{d}_j = \sum_{i \neq j} \mathbb{1}_{S_{n,j}^{(i)}}(X_i).$$

The expected degree is

$$\hat{c}_j = \sum_{i \neq j} V_{n,j}^{(i)}.$$

Proposition (Quadrature rules)

The estimate $\hat{\mu}_n^{(NN-loo)}(g)$ can be expressed as a linear estimates of the form

$$\hat{\mu}_n^{(\mathrm{NN-loo})}(g) = \sum_{i=1}^{n} w_{i,n}^{(\mathrm{NN-loo})} g(X_i)$$

where $w_{i,n}^{(\mathrm{NN-loo})} = (1 + \hat{c}_i - \hat{d}_i)/n$ (the weights does not depend on g)

Convergence rate

Proposition

Assume that $g : \mathbb{R}^d \to \mathbb{R}$ *is L-Lipschitz,* $\inf_{x \in \mathcal{X}} f(x) > b$ *and* $\sup_{x \in \mathcal{X}} f(x) < U$*. Then we have*

$$\mathbb{E}\Big[\big(\hat{\mu}_n^{(\mathrm{NN-loo})}(g) - \mu(g)\big)^2\Big] \leq C n^{-1} n^{-2/d}$$

Method implementation

The method CVNN returns the value of $\hat{\mu}_n^{(NN)}(g)$ for which the integral $\int \hat{g}_n d\mu$ is replaced by a Monte Carlo estimate that uses $M = n^2$ generation. That is

$$\int \hat{g}_n d\mu \simeq M^{-1} \sum_{i=1}^{M} \hat{g}_n(\tilde{X}_i),$$

where $\tilde{X}_i$ are i.i.d draws according to μ.

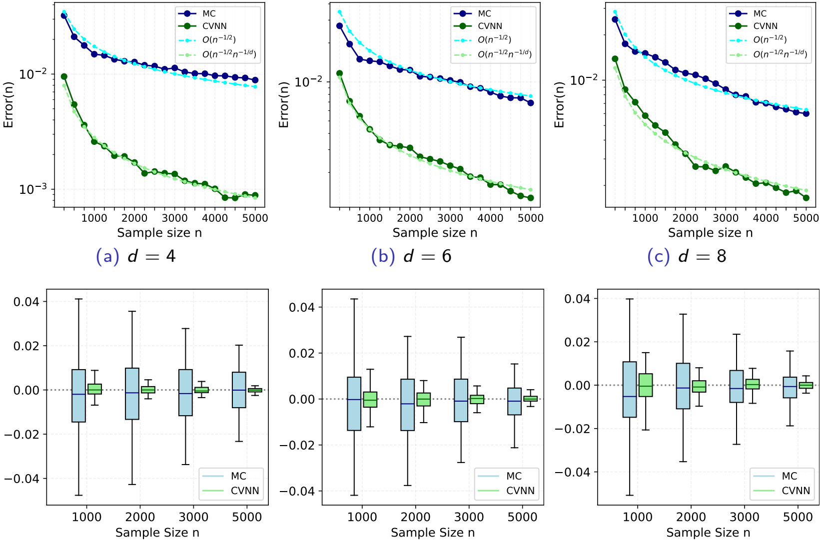
Integrand

$$g_1(x) = 1 + \sin(\pi(2d^{-1} \sum_{i=1}^{d} x_i - 1)) \qquad g_2(x) = \prod_{i=1}^{d} \log(2)2^{1-x_i},$$

(both integrate to 1 on $[0,1]^d$)

Parameters

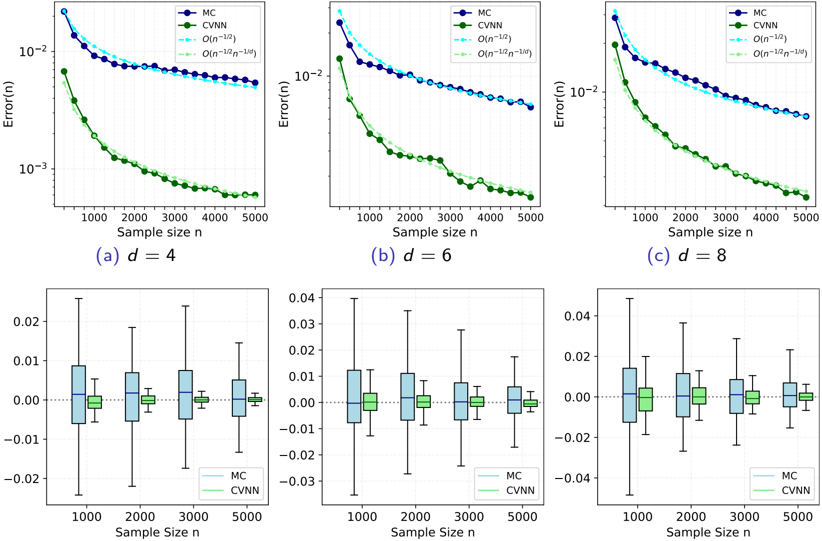
- dimensions $d \in \{4; 6; 8\}$,
- from $n = 250$ to $n = 5,000$
- performance measured with $\mathbb{E}[|\hat{\mu}_n^{(\mathrm{NN})}(g) - \mu(g)|^2]^{1/2}$



(a) $d = 4$ (b) $d = 6$ (c) $d = 8$

(d) $d = 4$ (e) $d = 6$ (f) $d = 8$

Figure: Boxplots obtained over 100 replications for function g_1 in dimension $d \in \{4; 6; 8\}$.



(a) $d = 4$ (b) $d = 6$ (c) $d = 8$

(d) $d = 4$ (e) $d = 6$ (f) $d = 8$

Figure: Boxplots obtained over 100 replications for function g_2 in dimension $d \in \{4; 6; 8\}$.

Integrand	Sample Size n / Method	500	$1,000$	$2,000$	$3,000$	$5,000$
g_1 $(d = 2)$	MC	9.4e-4	4.6e-4	1.9e-4	1.5e-4	1.1e-5
	QMC (Sobol)	4.1e-5	2.0e-5	3.6e-6	2.6e-6	1.0e-6
	NN-euclidean	6.7e-6	1.7e-6	3.7e-7	1.7e-7	5.3e-8
	NN-manhattan	7.0e-6	1.7e-6	4.4e-7	1.7e-7	5.7e-8
	NN-chebyshev	6.1e-6	1.7e-6	3.8e-7	2.3e-7	6.2e-8
g_2 $(d = 2)$	MC	5.0e-4	2.0e-4	1.1e-4	5.8e-5	3.1e-5
	QMC (Sobol)	6.3e-6	5.2e-6	1.7e-6	5.1e-7	2.1e-7
	NN-euclidean	9.2e-6	1.9e-6	5.0e-7	1.9e-7	1.0e-7
	NN-manhattan	8.5e-6	2.5e-6	5.3e-7	2.2e-7	9.7e-8
	NN-chebyshev	1.0e-5	1.9e-6	4.8e-7	1.8e-7	1.0e-7
g_3 $(d = 2)$	MC	1.8e-4	7.5e-5	3.4e-5	2.7e-5	2.1e-5
	QMC (Sobol)	5.7e-6	2.5e-6	5.4e-7	4.9e-7	1.7e-7
	NN-euclidean	7.9e-7	2.0e-7	3.9e-8	2.3e-8	7.7e-9
	NN-manhattan	8.2e-7	1.9e-7	4.0e-8	2.1e-8	6.7e-9
	NN-chebyshev	8.3e-7	2.2e-7	3.5e-8	2.4e-8	9.6e-9

Table: Mean Squared Error for integrands g_1, g_2, g_3 in dimension $d = 2$ obtained over 100 replications.

References I

The paper is not available yet but should be soon

Azaïs, R., B. Delyon, and F. Portier (2018). Integral estimation based on markovian design. *Advances in Applied Probability 50*(3), 833–857.

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

Delyon, B. and F. Portier (2016). Integral approximation by kernel smoothing. *Bernoulli 22*(4), 2177–2208.

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

Giné, E. and A. Guillou (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist. 37*(4), 503–522.

Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. New York: Springer.

Glasserman, P. (2004). *Monte Carlo methods in financial engineering*, Volume 53. Springer.

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

McDiarmid, C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, Volume 16 of *Algorithms Combin.*, pp. 195–248. Springer, Berlin.

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

References II

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. (2013a). Monte Carlo Theory, Methods and Examples. http://statweb.stanford.edu/~owen/mc/.

Owen, A. B. (2013b). 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*(4), 1168–1186.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. Springer-Verlag, New York.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae 126*(3), 505–563.