Mathematical Methods of Uncertainty Quantification

Oliver Ernst

Professur Numerische Mathematik

Sommersemester 2016

Mathematik!
TU Chemnitz

Contents

Contents

Contents

Monte Carlo Methods

Monte Carlo

- George Louis Leclerc, Comte de Buffon (1707–1788), French naturalist and mathematician, posed the following problem in 1777:

  Let a needle of length ℓ be thrown at random onto a horizontal plane ruled with parallel straight lines spaced by a distance $d > \ell$ from each other. What is the probability p that the needle will intersect one of these lines?

- Analog randomized experiment to approximate π, later used by Laplace.

Theorem 2.1

The probability of a needle falling in such a way that it intersects one of the lines as described above is

$$p = \frac{2}{\pi}\frac{\ell}{d}.$$

- Let $\{H_k\}_{k\in\mathbb{N}}$ denote a sequence of i.i.d. random variables whose value is

$$H_k(\omega) = \begin{cases} 1 & \text{if } k\text{-th needle intersects a line,} \\ 0 & \text{otherwise.} \end{cases}$$

- Their common distribution is that of a Bernoulli trial with success probability $p = 2\ell/\pi d$. In particular:

$$\mathbf{E}[H_k] = p \qquad \forall k.$$

- Then $S_N = H_1 + \cdots + H_N$ is the total number of hits after N throws.
- SLLN:

$$\frac{S_N}{N} \to p \qquad \text{a.s.}$$

- Monte Carlo simulation: compute realizations of H_k by randomly sampling $X_k \sim \mathrm{U}[0, d/2]$ (distance of needle center to closest line) and $\Theta_k \sim \mathrm{U}[0, \pi/2]$ (acute angle of needle with lines) using a random number generator.

- Setting $d = 2$, $\ell = 1$ gives $p = \frac{1}{\pi}$. For large N, we should have $N/S_N \approx \pi$.
- A MATLAB experiment (setting `rng('default')`) yields

N	S_N	N/S_N	rel. Error
10	3	3.33	6.1e-2
100	32	3.13	5.3e-3
1000	330	3.03	3.5e-2
10000	3188	3.14	1.5e-3

- The Italian mathematician Mario Lazzarini (1901) built a machine with which to carry out many repetitions of this random experiment. His needle was 2.5 cm long and the lines 3.0 cm apart. He claims to have observed 1808 intersections for 3408 throws, corresponding to

$$\pi \approx 2 \cdot \frac{2.5}{3} \cdot \frac{3408}{1808} = 3.141592920353983\ldots$$

  which corresponds to an (absolute) error of $2.67 \cdot 10^{-7}$.
- Is this too good to be true?

Contents

- Given a device for generating a sequence $\{X_k\}$ of i.i.d. realizations of a given random variable X, basic MC simulation uses the approximation

$$\mathbf{E}[X] \approx \frac{S_N}{N}, \qquad S_N = X_1 + \cdots + X_N.$$

- By the SLLN, $\frac{S_N}{N} \to \mathbf{E}[X]$ a.s.
- Similarly, for a measurable function f, $\mathbf{E}[f(X)] \approx \frac{1}{N}\sum_{k=1}^{N} f(X_k)$.
- For a RV $X \in L^2(\Omega; R)$ the standardized RV

$$X^* := \frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{Var}\, X}} \quad \text{has} \quad \mathbf{E}[X^*] = 0,\ \mathbf{Var}\, X^* = 1.$$

- If $\mu = \mathbf{E}[X]$, $\sigma^2 = \mathbf{Var}\, X$, then $\mathbf{E}[S_N] = N\mu$, $\mathbf{Var}\, S_N = N\sigma^2$ and, by the CLT,

$$S_N^* = \frac{S_N - N\mu}{\sqrt{N}\sigma} \to N(0,1).$$

- Since
$$\mathbf{E}\left[\left(\frac{S_N}{N}-\mu\right)^2\right] = \mathbf{Var}\,\frac{S_N}{N} = \frac{\sigma^2}{N} \to 0 \qquad (N\to\infty)$$
we have L^2-convergence of S_N/N to μ and, by Theorem A.25, for any $\epsilon > 0$,
$$\mathbf{P}\left\{\left|\frac{S_N}{N}-\mu\right| > N^{-1/2+\epsilon}\right\} \le \frac{\sigma^2}{N^{2\epsilon}}, \tag{2.1}$$
i.e., as the number N of samples increases, the probability of the error being larger than $O(N^{-1/2+\epsilon})$ converges to zero for any $\epsilon > 0$.
- If $\rho := \mathbf{E}\left[|X-\mu|^3\right] < \infty$, then the Berry-Esseen bound Theorem A.47 further gives
$$|\mathbf{P}(S_N^* \le x) - \Phi(x)| \le C\frac{\rho}{\sigma^3\sqrt{N}}, \tag{2.2}$$
where Φ denotes the cdf of $\mathrm{N}(0,1)$.

- For a RV $Z \sim \mathrm{N}(0,1)$ and $x \in \mathbb{R}$, this implies

$$\mathbf{P}(S_N^* \le x) = \mathbf{P}(Z \le x) + O(N^{-1/2})$$

and therefore

$$\begin{aligned}\mathbf{P}(|S_N^*| \le x) &= \mathbf{P}(S_N^* \le x) - \mathbf{P}(S_N^* < -x) \\ &= \mathbf{P}(Z \le x) - \mathbf{P}(Z < -x) + O(N^{-1/2}) \\ &= \mathbf{P}(|Z| \le x) + O(N^{-1/2}) \\ &= \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + O(N^{-1/2})\end{aligned}$$

where

$$\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) = 2\,\Phi(x) - 1.$$

- If the $O(N^{-1/2})$-term is assumed negligible, this can be used to construct (asymptotic) confidence intervals for S_N^*, i.e., the MC estimate S_N/N.

True confidence intervals are obtained if we carry along the bound in the Berry-Esseen estimate (2.2), denoted by B_N,

$$-B_N \le \mathbf{P}(S_N^* \le x) - \Phi(x) \le B_N$$

i.e., for $R \ge 0$ we have

$$\begin{aligned}\mathbf{P}(|S_N^*| \le R) &= \mathbf{P}(S_N^* \le R) - \mathbf{P}(S_N^* < -R) \\ &\ge \Phi(R) - B_N - (\Phi(-R) + B_N) \\ &= \underbrace{\Phi(R) - \Phi(-R)}_{=:\gamma_R} - 2B_N\end{aligned}$$

and, in the same manner, $\mathbf{P}(|S_N^*| \le R) \le \gamma_R + 2B_N$, i.e.,

$$\gamma_R - 2\,B_N \le \mathbf{P}\left(\mu \in \left[\frac{S_N}{N} - \frac{\sigma R}{\sqrt{N}}, \frac{S_N}{N} + \frac{\sigma R}{\sqrt{N}}\right]\right) \le \gamma_R + 2\,B_N.$$

In the Buffon needle problem, we have, with RV H denoting the outcome of each needle throw,

$$\mathbf{E}\,[H] = p, \quad \mathbf{Var}\,H = p(1-p), \quad \mathbf{E}\left[|H-p|^3\right] = p(1-p)[1-2p+2p^2]$$

and therefore

$$S_N^* = \frac{S_N/N - p}{\sqrt{\frac{p(1-p)}{N}}} \to N(0,1).$$

Choosing $R = 2$ gives $\gamma_2 = \mathrm{erf}(\sqrt{2}) \approx 0.9545$, so that an asymptotic confidence interval of level $\gamma_2 \approx 95\%$ is obtained as

$$\left[\frac{S_N}{N} - 2\sqrt{\frac{p(1-p)}{N}}, \frac{S_N}{N} + 2\sqrt{\frac{p(1-p)}{N}}\right].$$

In Lazzarini's experiment $\ell/d = 5/6$, $N = 3408$, giving $p = \frac{5}{3\pi} \approx 0.5305$, giving $\pi \approx \frac{5}{3} \cdot \frac{3408}{1808} = \frac{355}{113}$. This corresponds to an approximation error of

$$\left| \frac{S_N}{N} - p \right| = \left| \frac{1808}{3408} - \frac{5}{3\pi} \right| =: \epsilon_L \approx 4.5 \cdot 10^{-8}.$$

For the given values of p and N, we have

$$2\sqrt{\frac{p(1-p)}{3408}} \approx 0.0171,$$

giving a γ_2-asymptotic confidence interval around S_N/N of width

$$4\sqrt{\frac{p(1-p)}{3408}} \approx 0.0342.$$

The γ_2-asymptotic confidence interval has a width of $2\epsilon_L$ for $N > 4.9094 \cdot 10^{14}$.

To obtain true γ_R confidence intervals using the Berry-Esseen bound, note that here

$$B_N = C\frac{\rho}{\sigma^3\sqrt{N}} = C\frac{1-2p+2p^2}{\sqrt{p(1-p)N}} \le \frac{0.7096}{\sqrt{N}},$$

where we have used the value $C = 0.7056$ given in [Shevtsova, 2006].

The upper bound $\gamma_R + 2B_N$ for the probability that S_N/N is within ϵ_L of the true value p after $N = 3408$ throws, corresponds to

$$\frac{\sigma}{\sqrt{N}}R \le \epsilon_L, \qquad \text{i.e.,} \qquad R \le R_L := \frac{\sqrt{N}\epsilon_L}{\sigma} \approx 5.2695\cdot 10^{-6}, \quad \gamma_{R_L} \approx 4.2044\cdot 10^{-6}$$

giving

$$\mathbf{P}\left(\left|\frac{S_N}{N} - p\right| \le \epsilon_L\right) \le \gamma_{R_L} + 2B_{3408} \approx 0.0243.$$

Contents

- The MC method can be viewed as the equal-weight quadrature formula[3]

$$\mathbf{E}\left[f(X)\right] = \int_\Omega f(X(\omega))\,\mathrm{d}\mathbf{P}(\omega) = \int_0^1 f(x)\,\mathrm{d}x \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j),$$

  where the quadrature nodes x_j are drawn randomly from $\mathrm{U}[0,1]$.
- In quasi-Monte Carlo (QMC) methods, one chooses the nodes instead from certain deterministic number sequences possessing a property known as low discrepancy.
- Such sequences can be generated and analyzed using techniques from number theory, and can be used to obtain multivariate quadrature techniques which perform better than MC and are better suited for large dimension than classical methods of numerical integration.
- We consider the univariate case here to illustrate the basic principle.

[3] Here we assume $X \sim \mathrm{U}[0,1]$, otherwise assume X has a density and obtain a different integrand.

For $x \in \mathbb{R}$ denote by

$$[x] := \max\{k \in \mathbb{Z} : k \le x\}$$

the integral part of x and by

$$\{x\} := x - [x] \in [0, 1)$$

its fractional part.

Given a sequence $(x_n)_{n\in\mathbb{N}}$ of real numbers, denote for $N \in \mathbb{N}$ and $E \subset [0, 1)$ the counting function

$$A(E; N) := |\{x_1, \ldots, x_N\} \cap E|.$$

Definition 2.2

The sequence $(x_n)_{n\in\mathbb{N}}$ of real numbers is said to be uniformly distributed modulo 1 if, for every pair $a, b \in [0, 1)$, $a < b$, we have

$$\lim_{N\to\infty} \frac{A([a, b); N)}{N} = b - a. \tag{2.3}$$

Note: When proving an assertion about the discrepancy of a given sequence, no loss of generality will result if we assume the sequence to be contained in $[0,1)$.

Denoting by $\mathbb{1}_{[a,b)}$ the characteristic (indicator) function of $[a,b)$, we may write (2.3) equivalently as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{[a,b)}(\{x_j\}) = \int_0^1 \mathbb{1}_{[a,b)}(x)\,\mathrm{d}x. \tag{2.4}$$

Theorem 2.3

The sequence $(x_n)_{n\in\mathbb{N}}$ of real numbers is uniformly distributed modulo 1 if and only if for every real-valued function $f \in C[0,1]$ there holds

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} f(\{x_j\}) = \int_0^1 f(x)\,\mathrm{d}x. \tag{2.5}$$

Definition 2.4

For a finite set $\{x_1, \ldots, x_N\}$ of real numbers, the number

$$D_N = D_N(x_1, \ldots, x_N) := \sup_{0 \le \alpha < \beta \le 1} \left| \frac{A([\alpha, \beta); N)}{N} - (\beta - \alpha) \right| \tag{2.6}$$

is called the discrepancy of the set. For an (infinite) sequence $(x_n)_{n \in \mathbb{N}}$ we denote by $D_N((x_n))$ the discrepancy of the first N terms of the sequence.

Theorem 2.5

The sequence (x_n) is uniformly distributed modulo 1 if and only if $\lim_{N \to \infty} D_N((x_n)) = 0$.

Theorem 2.6

For any set of N real numbers we have $\frac{1}{N} \le D_N \le 1$.

Definition 2.7

For a finite set $\{x_1, \ldots, x_N\}$ of real numbers, the number

$$D_N^* = D_N^*(x_1, \ldots, x_N) := \sup_{0 \le \alpha \le 1} \left| \frac{A([0, \alpha); N)}{N} - \alpha \right| \tag{2.7}$$

is called the star-discrepancy of the set, with the same generalization to sequences.

Theorem 2.8

The discrepancies D_N and D_N^* are related by $D_N^* \le D_N \le 2D_N^*$.

Corollary 2.9

The sequence (x_n) is uniformly distributed modulo 1 if and only if $\lim_{N\to\infty} D_N^*((x_n)) = 0$.

Theorem 2.10

Let $x_1 \le x_2 \le \cdots \le x_N$ be N numbers in $[0,1)$. Then their discrepancy D_N^* is given by

$$D_N^* = \max_{j=1,\ldots,N} \max\left\{ \left| x_j - \frac{j}{N} \right|, \left| x_j - \frac{j-1}{N} \right| \right\} = \frac{1}{2N} + \max_{j=1,\ldots,N} \left| x_j - \frac{2j-1}{2N} \right|.$$

Definition 2.11

The total variation of a real-valued function f defined on $[0,1]$ is defined as the number

$$V(f) := \sup_{x_0,\ldots,x_N} \sum_{j=1}^{N} |f(x_j) - f(x_{j-1})|,$$

where the supremum is taken over all partitions of $[0,1]$. A function for which the variation is finite is called a function of bounded variation.

Note: If f is differentiable on $[0,1]$ there holds

$$V(f) = \int_0^1 |f'(t)|\,\mathrm{d}t.$$

Lemma 2.12

Let $x_1 \le x_2 \le \cdots \le x_N$ be N given points in $[0,1)$ and let f be a function of bounded variation on $[0,1]$. Then with $x_0 := 0$ and $x_{n+1} := 1$ there holds

$$\frac{1}{N}\sum_{j=1}^{N} f(x_j) - \int_0^1 f(t)\,\mathrm{d}t = \sum_{j=0}^{N} \int_{x_j}^{x_{j+1}} \left(t - \frac{j}{N}\right)\,\mathrm{d}f(t). \tag{2.8}$$

Theorem 2.13 (Koksma's inequality, 1942)

Let f be a function of bounded variation $V(f)$ on $[0,1]$ and assume we are given N points $x_1, \ldots, x_N$ with discrepancy D_N^*. Then

$$\left| \frac{1}{N}\sum_{j=1}^{N} f(x_j) - \int_0^1 f(t)\,\mathrm{d}t \right| \le V(f)\, D_N^*. \tag{2.9}$$

We now consider a classical example of a low-discrepancy sequence, the van der Corput sequence defined as follows: for an integer $b \geq 2$ set $Z_b := \{0, 1, 2, \ldots, b-1\}$ and denote by

$$n = \sum_{j=0}^{\infty} a_j(n) b^j, \qquad a_j(n) \in Z_b$$

the representation of $n \in \mathbb{N}_0$ in base b (only finitely many a_j are nonzero). In terms of this base-b representation, we define the radical-inverse function $\phi_b : \mathbb{N}_0 \to [0, 1)$ by

$$\phi_b(n) = \sum_{j=0}^{\infty} a_j(n) b^{-(j+1)},$$

i.e. as the symmetric reflection of the base-b digit sequence of n "at the decimal point".

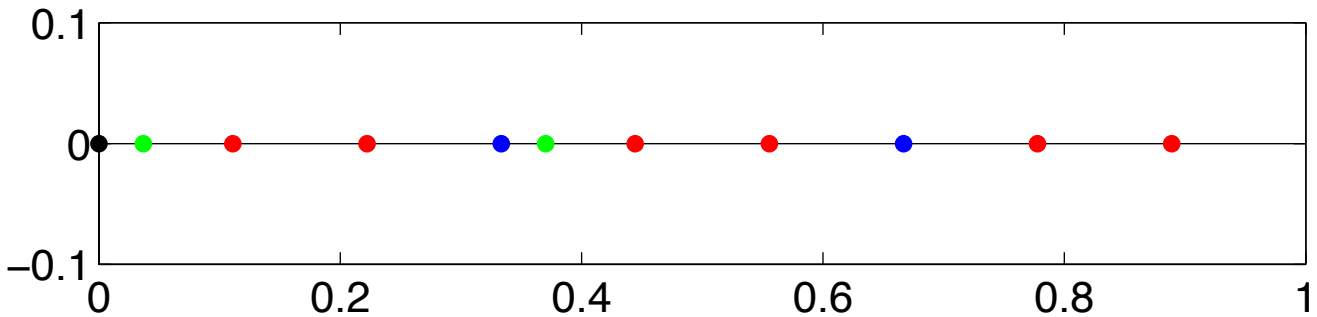
Definition 2.14

For an integer $b \geq 2$ the van der Corput sequence in base b is the sequence

$$(x_n)_{n\in\mathbb{N}_0} \quad \text{such that} \quad x_n = \phi_b(n), n \in \mathbb{N}_0.$$

Example: For $b = 3$ we obtain the first 11 terms in the sequence as

$$\{x_n\}_{n=0}^{10} = \left(0, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \frac{1}{27}, \frac{10}{27}, \dots\right).$$



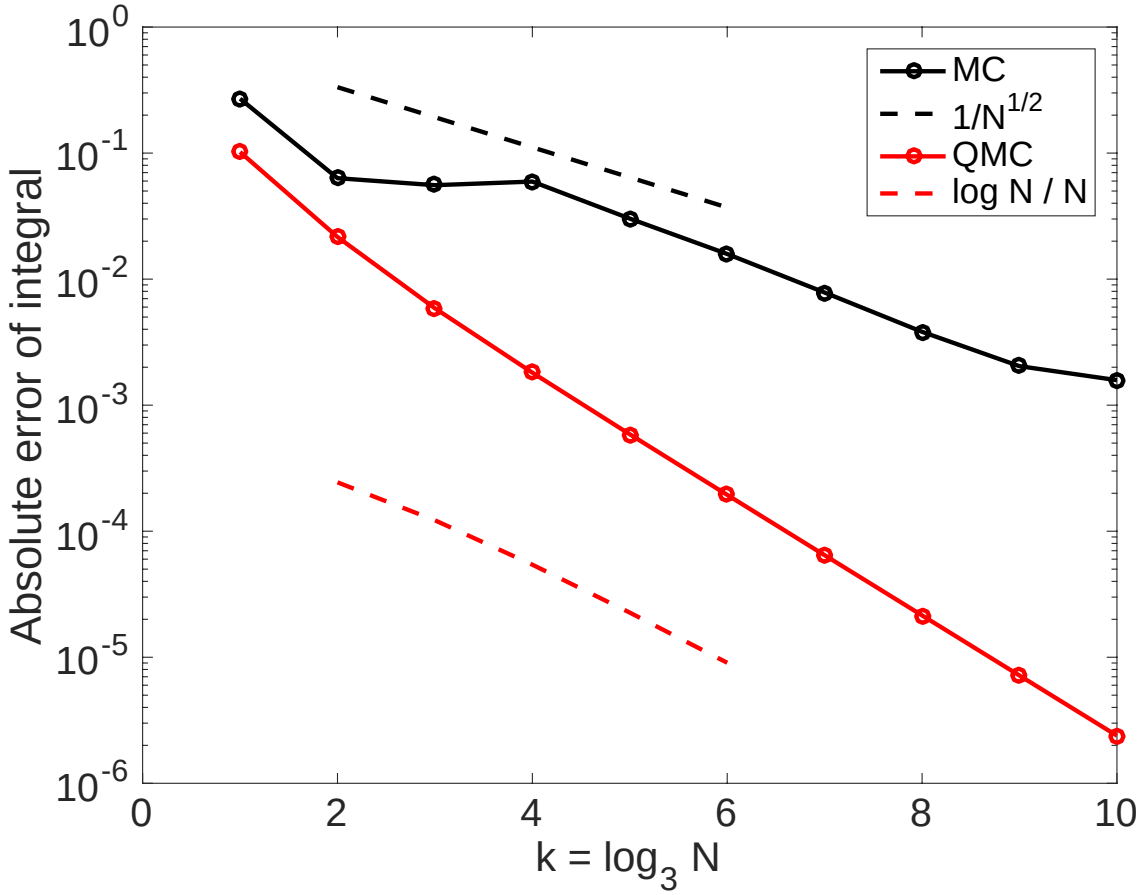
The colors indicate ordering in the sequence: black, blue, red, green.

Theorem 2.15

The discrepancy of the van der Corput sequence satisfies

$$D_N = O\left(\frac{\log N}{N}\right).$$

- A proof can be found in [Kuipers & Niederreiter, 1974].
- Theorem 2.15, in combination with Koksma's inequality, shows that the QMC approximation $\int_0^1 f(x)\,\mathrm{d}x \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j)$ using a van der Corput node sequence converges like $N^{-1}\log N$, a faster rate than that of $N^{-1/2}$ for MC.
- Further advantages of QMC over MC are the greater ease of generating deterministic node sequences rather than independent random samples following a particular distribution, smoother (deterministic) convergence of the QMC approximations vs. only probabilistic error bounds for MC.
- MC methods cannot take advantage of smoothness of the integrand.



$\int_0^1 \exp(x)\,\mathrm{d}x$ using the base 3 van der Corput sequence vs. 10 averaged MC runeswith uniformly distributed quandrature nodes.

- The concepts discrepancy and total variation can be generalized to multidimensional domains and multivariate integrands.
- The normalized problem is integration on the unit cube in $M > 1$ dimensions approximated by equal-weight quadrature

$$\int_{[0,1]^M} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{x}_j).$$

- This requires algorithms for constructing low-discrepancy point sets in several dimensions as well as convergence theory.

For an arbitrary subset $E \subset [0,1]^M$ and a finite sequence $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \subset [0,1]^M$ we define

$$A(E; N) := \sum_{j=1}^{N} \mathbb{1}_E(\boldsymbol{x}_j)$$

and the star discrepancy of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ by

$$D_N^* = \sup_E \left| \frac{A(E; N)}{N} - |E| \right|$$

with supremum taken over all subintervals $E \subset [0,1]^M$ of the form $\prod_{m=1}^{M} [0, b_M)$.

To generalize the concept of total variation of a function to several dimensions, consider

$$f : [0,1]^M \to \mathbb{R}, \quad M \geq 2, \qquad \boldsymbol{x} \mapsto f(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, \ldots, x_M).$$

For a subinterval $J \subset [0,1]^M$, let $\Delta(f; J)$ denote an alternating sum of the values of f at the vertices of J (function values at adjacent vertices have opposite signs).

The variation of f in the sense of Vitali on $[0,1]^M$ is defined as

$$V^{(M)}(f) = \sup_{\mathfrak{P}} \sum_{J \in \mathfrak{P}} |\Delta(f; J)|$$

where the supremum is taken over all partitions $\mathfrak{P}$ of $[0,1]^M$ into subintervals.

There also holds

$$V^{(M)}(f) = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^M f}{\partial x_1 \cdots \partial x_M} \right| \mathrm{d}x_1 \cdots \mathrm{d}x_M$$

when this mixed partial derivative of f exists and is continuous.

For $1 \le m \le M$ and $1 \le i_1 < i_2 < \cdots < i_m \le M$ let $V^{(m)}(f; i_1, \ldots, i_m)$ denote the Vitali variation of f restricted to the m-dimensional face

$$\left\{(x_1, \ldots, x_M) \in [0,1]^M : x_j = 1 \text{ for } j \neq i_1, \ldots, i_m\right\}.$$

Then

$$V(f) := \sum_{m=1}^{M} \sum_{1 \le i_1 < \cdots < i_m \le M} V^{(m)}(f; i_1, \ldots, i_m)$$

is called the variation of f in the sense of Hardy and Krause on $[0,1]^M$.

The multivariate generalization of Koksma's inequality is the Koksma-Hlawka inequality is due to [Hlawka, 1961]: (cf. [Niederreiter, 1992])

Theorem 2.16 (Koksma-Hlawka inequality)

If f has bounded variation $V(f)$ on $[0,1]^M$ in the sense of Hardy and Krause, then, for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in [0,1)^M$, we have

$$\left| \frac{1}{N} \sum_{j=1}^{N} f(\boldsymbol{x}_j) - \int_{[0,1]^M} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| \leq V(f) \, D_N^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N).$$

Theorem 2.17

For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in [0,1)^M$ and any $\epsilon > 0$ there exists a function $f \in C^\infty([0,1]^M)$ with $V(f) = 1$ and

$$\left| \frac{1}{N} \sum_{j=1}^{N} f(\boldsymbol{x}_j) - \int_{[0,1]^M} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| > D_N^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) - \epsilon.$$

There is a vast number of construction principles for sequences of low discrepancy in higher dimensions. An immediate generalization of the van der Corput sequence is the Halton sequence:

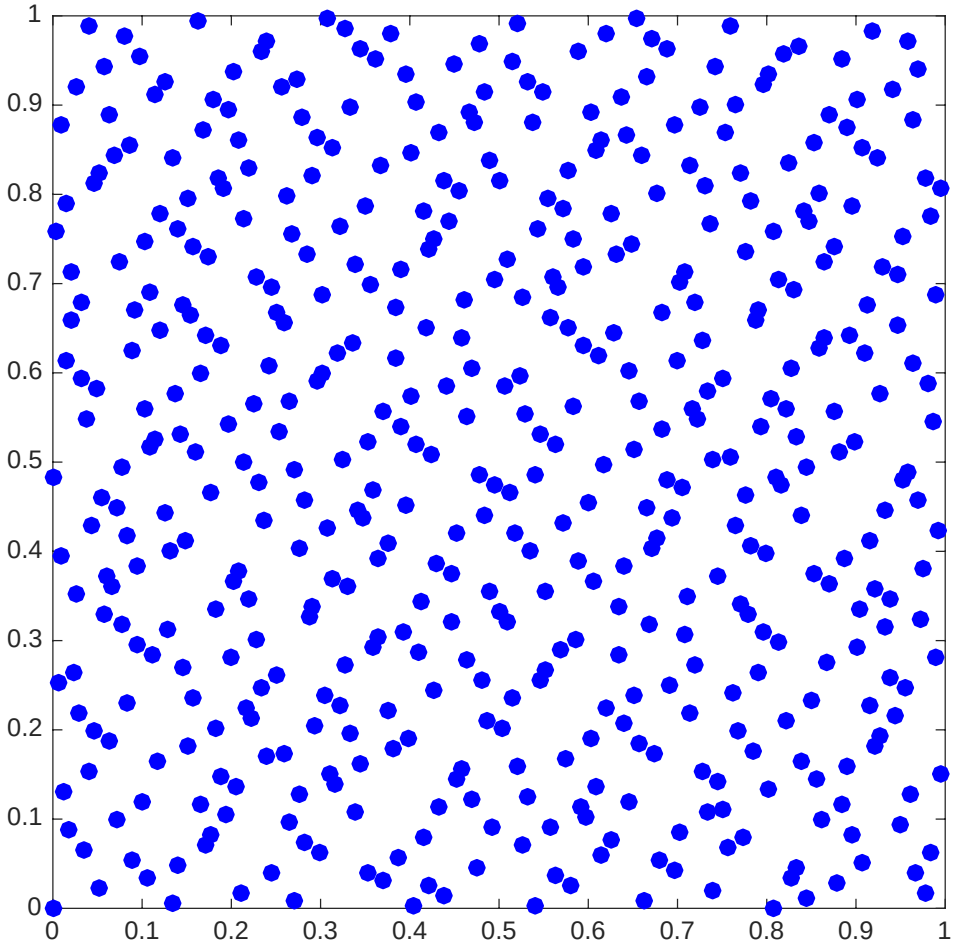
Let $p_1, p_2, \ldots, p_M$ be the first M prime numbers. The Halton sequence $(\boldsymbol{x}_j)$ in M dimensions is given by

$$\boldsymbol{x}_j = (\phi_{p_1}(j), \phi_{p_2}(j), \ldots, \phi_{p_M}(j)) \qquad j = 0, 1, 2, \ldots \ .$$

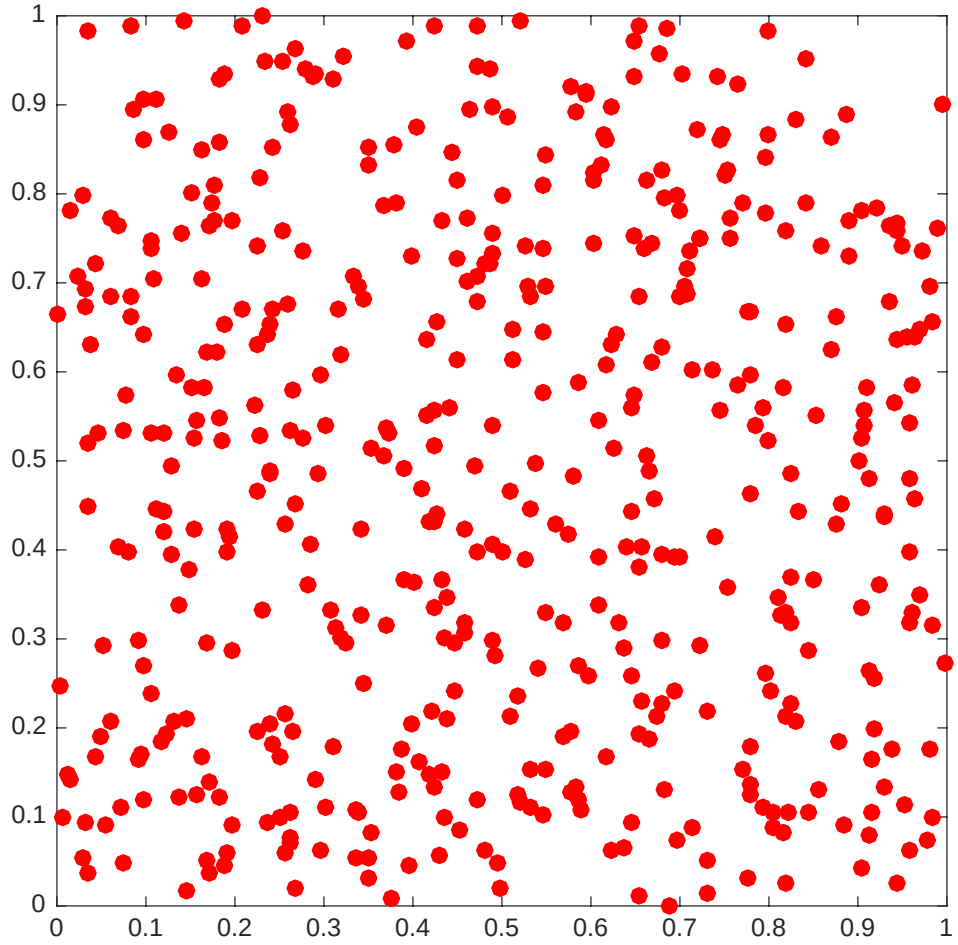
The star discrepancy of the Halton sequence satisfies

$$D_N^* = O\left(\frac{(\log N)^M}{N}\right)$$

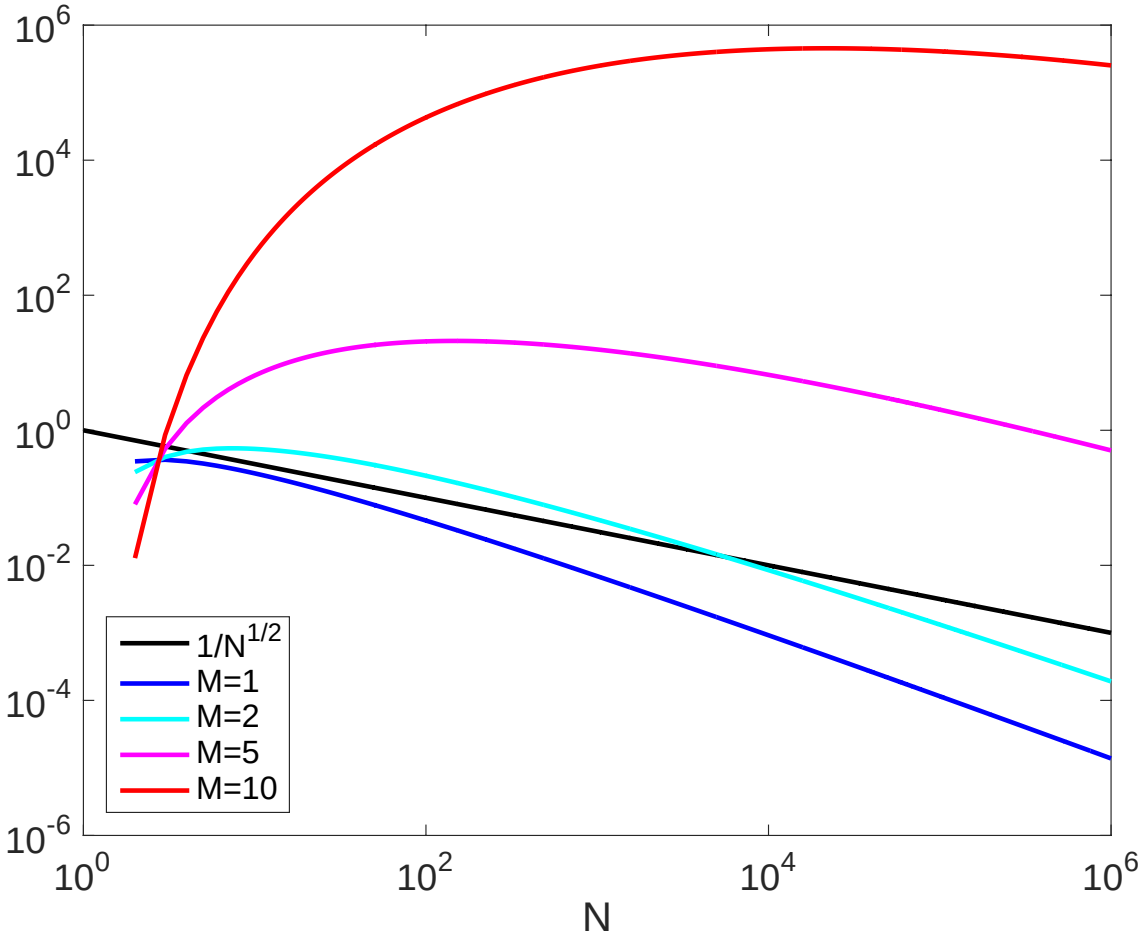
with an implied constant depending on M.



500 Halton points



500 random (uniform) points



Asymptotic behavior of $N^{-1/2}$ versus $(\log N)^M/N$ for different values of M.

- Initial investigations of QMC integration focussed on convergence for fixed dimension M.
- Until the mid 1990s it was believed that QMC quadrature is effective up to at most $M = 30$.
- [Paskov & Traub, 1995] gave an example where QMC clearly outperformed MC for a financial mathematics problem with dimension $M = 360$.
- Attempts to explain this observation led to the identification of smoothness properties of the integrand which lead to complexities essentially independent of the dimension M, i.e., $O(1/N)$ or even better.
- Analysis tools: weighted Sobolov spaces, reproducing kernel Hilbert spaces, further multivariate low-discrepancy sequences (lattice rules, digital nets). Practical tools: fast generation algorithms for these ("component by component").
- [Dick, Kuo & Sloan, 2013].

- The constant in the MC convergence rate appearing in (2.1) is the variance of the RV from which MC samples are being drawn.
- By designing an equivalent MC approximation with lower variance, we can expect to obtain faster convergence.
- Techniques for achieving this are collectively known as variance reduction methods.
- Define the efficiency of an estimator $\hat{\mu}$ for μ by

$$\mathrm{Eff}(\hat{\mu}) := \frac{1}{\mathrm{MSE}(\hat{\mu}) \cdot \mathscr{C}(\hat{\mu})}$$

  where $\mathrm{MSE}(\hat{\mu}) = (\mathrm{Bias}(\hat{\mu}))^2 + \mathbf{Var}(\hat{\mu})$ and $\mathscr{C}(\hat{\mu})$ is the expected cost of $\hat{\mu}$.
- For an (unbiased) MC estimator $\hat{\mu}$, we have $\mathbf{Var}(\hat{\mu}) = \frac{\sigma^2}{N}$ and $\mathscr{C}(\hat{\mu}) = CN$ for some constant $C > 0$. Thus $\mathrm{Eff}(\hat{\mu}) = (\sigma^2 C)^{-1}$ is independent of N.
- For increased efficiency reduction in variance must not increase cost proportionally.

- To approximate $\mathbf{E}[X]$ by standard MC, we draw independent samples $\{X_k\}_{k=1}^N$ of X and compute the sample average S_N/N.
- Now assume a second set of samples $\tilde{X}_k$ of X is given with sample average $\tilde{S}_N/N$.
- Since both sample averages converge to $\mathbf{E}[X]$, so does $\frac{1}{2}(S_N/N + \tilde{S}_N/N)$.
- When X_k and $\tilde{X}_k$ are negatively correlated they are called antithetic samples, and the approximation $\frac{1}{2N}(S_N + \tilde{S}_N)$ is a more reliable approximation of $\mathbf{E}[X]$ than $\frac{1}{2N}S_{2N}$.

Theorem 2.18

Let thew two sequences $\{X_k\}$ and $\{\tilde{X}_k\}$ of random variables be identically distributed with

$$\mathbf{Cov}(X_j, X_k) = \mathbf{Cov}(\tilde{X}_j, \tilde{X}_k) = 0 \quad \text{for } j \neq k.$$

Then the sample averages S_N/N and $\tilde{S}_N/N$ satisfy

$$\mathbf{Var}\,\frac{S_N + \tilde{S}_N}{2N} = \mathbf{Var}\,\frac{S_{2N}}{2N} + \frac{1}{2}\,\mathbf{Cov}\left(\frac{S_N}{N}, \frac{\tilde{S}_N}{N}\right) \leq \mathbf{Var}\,\frac{S_N}{N}. \tag{2.10}$$

- In general: Variance of average of N samples and N antithetic samples less than variance of N independent samples.
- Better: negatively correlated S_N/N and $\tilde{S}_N/N$, therefore variance of N samples and N antithetic samples less than variance of $2N$ indepependent samples.

Consider the popular model of the dynamics of two interacting populations

$$\dot{\boldsymbol{u}} = \begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \end{bmatrix} = \begin{bmatrix} u_1(1-u_2) \\ u_2(u_1-1) \end{bmatrix}, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0.$$

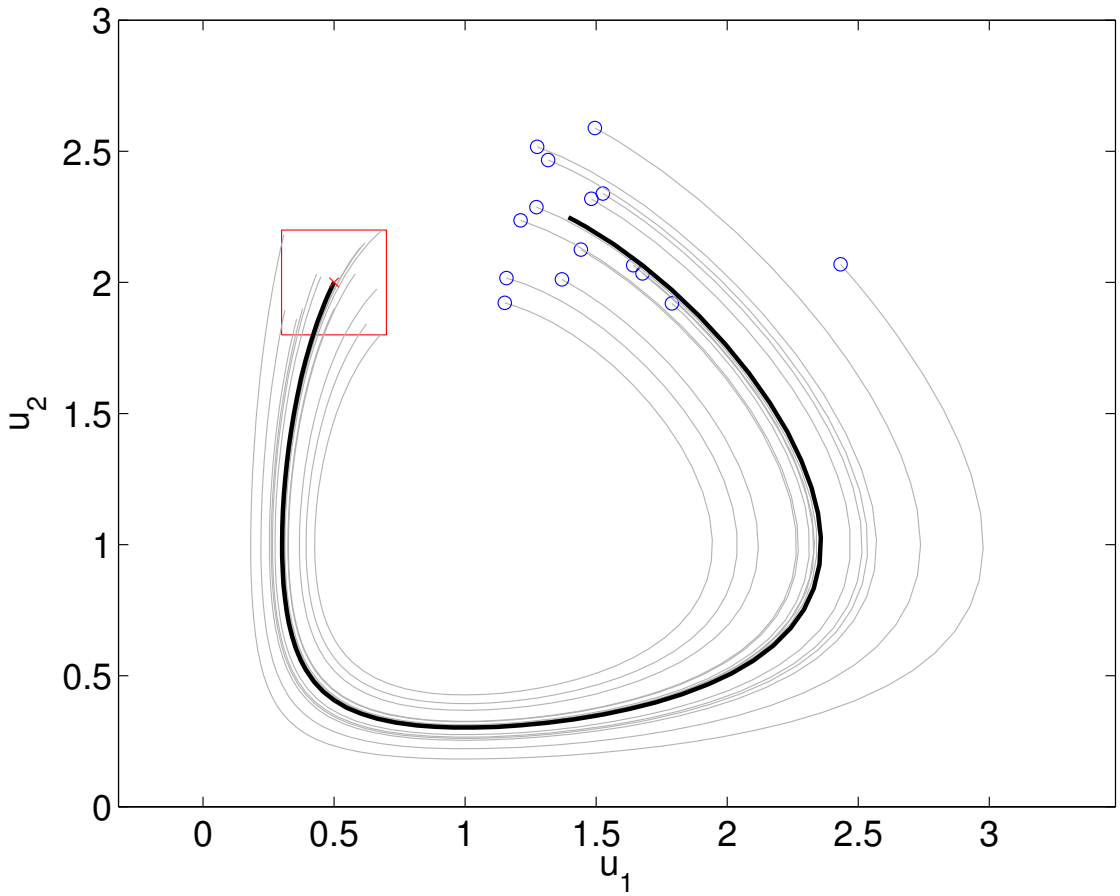
Assume the vector of initial conditions $\boldsymbol{u}_0$ is uncertain and that it is modeled as a random vector $\boldsymbol{u}_0 \sim \mathrm{U}(\Gamma)$, where Γ denotes the square

$$\Gamma = \overline{\boldsymbol{u}}_0 + [-\epsilon, \epsilon]^2.$$

- **Goal:** estimate $\mathbf{E}[u_1(T)]$ at time $T > 0$.
- Denote by $\boldsymbol{u}_n = \boldsymbol{u}_n(\omega)$ the explicit Euler approximation after n time steps of length Δt starting with initial data $\boldsymbol{u}_0 = \boldsymbol{u}_0(\omega)$.
- Define $\phi(\boldsymbol{u}) = u_1$ for $\boldsymbol{u} = [u_1, u_2]^{\mathsf{T}} \in \mathbb{R}^2$, estimate $\mathbf{E}[\phi(\boldsymbol{u}_{n_T})]$ for $n_T \Delta t = T$, using the MC method.
- Denote by $\overline{S}_N := S_N/N$ the average over N samples of $u_1(T)$.
- Expect better approximations for N large and Δt small.

Monte Carlo Methods

Example: Predator-prey dynamical system



Population dynamics problem integrated over $[0, T = 6]$ with $\boldsymbol{u}_0 = [0.5, 2] + \mathrm{U}[-\epsilon, \epsilon]$ for $\epsilon = 0.2$. Unperturbed trajectory (black) along with 15 perturbed trajectories. For the unperturbed trajectory $u_1(T) = 1.3942$.

- **Notation:**

forward map:	$G : \Gamma \to C([0,T];\mathbb{R}^2)$
discretized forward map:	$G_{\Delta t} : \Gamma \to C([0,T];\mathbb{R}^2)$
quantity of interest (QoI):	$Q : C([0,T];\mathbb{R}^2) \to \mathbb{R}, \quad \boldsymbol{u} \mapsto u_1(T) = \phi(\boldsymbol{u}(T))$
approximation of QoI:	$Q_{\Delta t} := \phi(\boldsymbol{u}_{n_T}) = \phi(G_{\Delta t}(\boldsymbol{u}_0)\vert_{t=T})$
MC estimate, N samples:	$\widehat{Q}_{\Delta t} := \widehat{Q}_{\Delta t,N} \approx \mathbf{E}[Q_{\Delta t}] \approx \mathbf{E}[Q]\,.$

- Error with N samples and $n_T = T/\Delta t$ time steps:

$$e_{N,\Delta t} = |\mathbf{E}[Q] - \widehat{Q}_{\Delta t}| \le \underbrace{|\mathbf{E}[Q] - \mathbf{E}[Q_{\Delta t}]|}_{\text{explicit Euler error}} + \underbrace{|\mathbf{E}[Q_{\Delta_t}] - \widehat{Q}_{\Delta t}|}_{\text{Monte Carlo error}}$$

- Explicit Euler error:
$$\|\boldsymbol{u}(T) - \boldsymbol{u}^{\Delta t}(T)\| \leq K\Delta t.$$
- ϕ Lipschitz-continuous with constant $L = 1$:
$$|\phi(\boldsymbol{u}(T)) - \phi(\boldsymbol{u}^{\Delta t}(T))| \leq K\,L\,\Delta t.$$
- Therefore
$$|\mathbf{E}[Q] - \mathbf{E}[Q_{\Delta t}]| = |\mathbf{E}[Q - Q_{\Delta t}]| \leq K\,L\,\Delta t. \tag{2.11}$$

- For MC error, apply CLT, confidence intervals: if $\mathbf{Var}\, Q_{\Delta t} = \sigma^2$,

$$\mathbf{P}\left(\left|\mathbf{E}\left[Q_{\Delta t}\right] - \widehat{Q}_{\Delta t,N}\right| \leq \frac{2\sigma}{\sqrt{N}}\right) > \gamma_2 + O(N^{-1/2})$$

- Combined with (2.11):

$$\mathbf{P}\left(e_{N,\Delta t} \leq K\, L\, \Delta t + \frac{2\sigma}{\sqrt{N}}\right) > \gamma_2 + O(N^{-1/2}).$$

- Balance discretization and MC errors: split total error δ equally

$$K\, L\, \Delta t \approx \frac{\delta}{2}, \qquad \frac{2\sigma}{\sqrt{N}} \approx \frac{\delta}{2},$$

leads to

$$\Delta t \approx \frac{\delta}{2KL} \quad \text{and} \quad N \approx \frac{16\sigma^2}{\delta^2}.$$

We introduce antithetic sampling to this problem by noting that, if $\boldsymbol{u}_0 \sim \mathrm{U}(\Gamma)$, then the same holds for the random vector

$$\tilde{\boldsymbol{u}}_0 := 2\overline{\boldsymbol{u}}_0 - \boldsymbol{u}_0.$$

Thus, the trajectories generated by the random initial data $\tilde{\boldsymbol{u}}_0$ have the same distribution as those generated by $\boldsymbol{u}_0$.

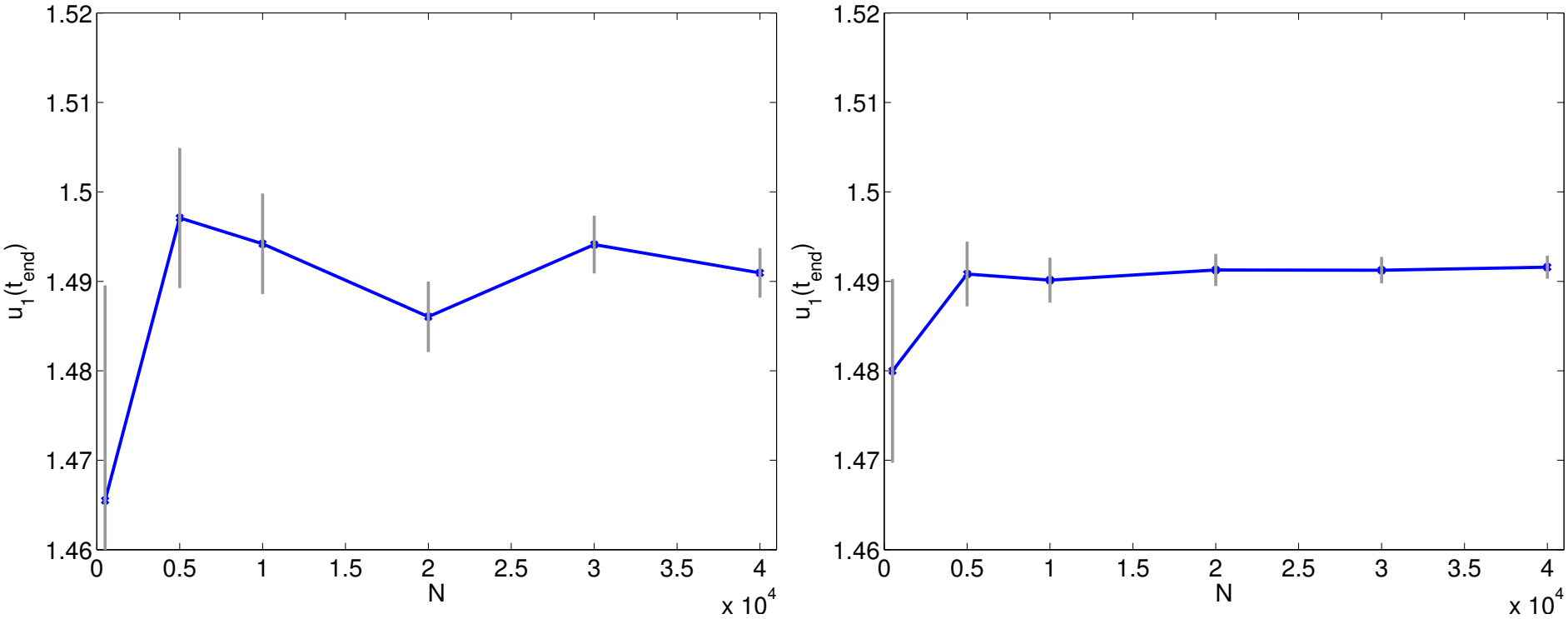
- Denote by $X_k = \phi(\boldsymbol{u}^{\Delta t}(T))$ the basic samples, by $\tilde{X}_k$ the antithetic counterparts. Note that all pairs of samples are independent except each sample and its antithetic counterpart.
- We estimate $\overline{S}_{2N}$ using the sample variance.
- To estimate $\frac{1}{2}\left(\overline{S}_N + \overline{\tilde{S}}_N\right)$ by (2.10), note that

$$\mathbf{Cov}(\overline{S}_N, \overline{\tilde{S}}_N) = \frac{1}{N^2}\,\mathbf{Cov}(S_N, \tilde{S}_N) = \frac{1}{N^2}\sum_{k=1}^{N}\mathbf{Cov}(X_k, \tilde{X}_k) = \frac{1}{N}\,\mathbf{Cov}(X, \tilde{X})$$

The last quantity can be estimated using the sample covariance.

Monte Carlo Methods

Example: Predator-prey dynamical system



MC estimation of $\mathbf{E}[u_1(T)]$ using standard MC with N samples (left) versus MC with antithetic sampling using $N/2$ samples (right) of the initial data. Both curves show the estimate along with a 95% (asymptotic) confidence interval.

- When estimating $\mathbf{E}[X]$ for a RV X using MC, let Y be another RV for which $\mathbf{E}[Y]$ is either known or easier to compute than $\mathbf{E}[X]$.
- For the modified RV

$$\tilde{X} := X - \beta(Y - \mathbf{E}[Y]), \qquad \beta \text{ a constant},$$

we clearly have $\mathbf{E}\left[\tilde{X}\right] = \mathbf{E}[X]$ as well as

$$\mathbf{Var}\,\tilde{X} = \mathbf{Var}\,X + \beta^2\,\mathbf{Var}\,Y - 2\beta\,\mathbf{Cov}(X,Y).$$

- Similarly, for the MC estimator $\tilde{\mu}_{\mathrm{MC}} = \frac{1}{N}\sum_{j=1}^{N}\tilde{X}_j$ with independent samples $\tilde{X}_j$ of $\tilde{X}$, we have $\mathbf{E}[\tilde{\mu}_{\mathrm{MC}}] = \mathbf{E}[X]$ as well as

$$\mathbf{Var}\,\tilde{\mu}_{\mathrm{MC}} = \frac{\mathbf{Var}\,X + \beta^2\,\mathbf{Var}\,Y - 2\beta\,\mathbf{Cov}(X,Y)}{N}$$

which is minimized for the value $\beta = \dfrac{\mathbf{Cov}(X,Y)}{\mathbf{Var}\,Y}$.

- When the estimation of a quantity of interest depends on observing an event of low-probability (rare event) sufficiently often, MC simulation can require an inordinately high number of samples, or result in unreliable estimates with large relative error.
- Examples: expected waiting time of bank customers waiting longer than 15 minutes; the probability of losing information packets in a communication link.
- Idea: modify probability distribution to make rare event more likely, adjust estimator to keep it unbiased.

Contents

The following summary of basic MLMC techniques and analysis closely follows [Teckentrup, 2013], see also [Cliffe et al., 2011]

To estimate the expectation $\mathbf{E}[Q]$ of a (random) quantity of interest (QoI) Q, assume only approximations $Q_h \approx Q$ are computable, where $h > 0$ denotes a discretization parameter for which

$$\lim_{h\to 0} \mathbf{E}[Q_h] = \mathbf{E}[Q].$$

More precisely, we shall assume the error in mean to converge at a rate α, i.e.,

$$|\mathbf{E}[Q_h - Q]| \lesssim h^\alpha \quad \text{as } h \to 0, \qquad \alpha > 0.$$

- Given an unbiased estimator $\widehat{Q}_h$ for $\mathbf{E}[Q_h]$, the associated mean-square error (MSE) may always be decomposed as

$$\begin{aligned}\mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}[Q]\right)^2\right] &= \mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}\left[\widehat{Q}_h\right] + \mathbf{E}\left[\widehat{Q}_h\right] - \mathbf{E}[Q]\right)^2\right] \\ &= \mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}\left[\widehat{Q}_h\right]\right)^2\right] + \left(\mathbf{E}\left[\widehat{Q}_h\right] - \mathbf{E}[Q]\right)^2 \\ &= \mathbf{Var}\,\widehat{Q}_h + \left(\mathbf{E}[Q_h] - \mathbf{E}[Q]\right)^2 \\ &= \mathbf{Var}\,\widehat{Q}_h + \mathbf{E}[Q_h - Q]^2\end{aligned}$$

  consisting of the variance of the estimator and the squared expectation of the discretization error (systematic error, bias).
- We shall sometimes refer to the root mean-square error (RMSE), which is simply the square root of the MSE, i.e., the L^2-norm of the estimation error

$$\sqrt{\mathbf{E}\left[\left(\widehat{Q}_h - \mathbf{E}[Q]\right)^2\right]}$$

- If the standard Monte Carlo estimator $\widehat{Q}_h = \widehat{Q}_{h,N}^{\mathrm{MC}}$ with N samples is used, and $Q_h^{(i)}$ denote i.i.d. RV with the same distribution as Q_h, then

$$\mathbf{Var}\,\widehat{Q}_{h,N}^{\mathrm{MC}} = \mathbf{Var}\left(\frac{1}{N}\sum_{i=1}^{N} Q_h^{(i)}\right) = \frac{1}{N^2} N\,\mathbf{Var}\,Q_h = \frac{\mathbf{Var}\,Q_h}{N},$$

giving

$$\mathbf{E}\left[\left(\widehat{Q}_{h,N}^{\mathrm{MC}} - \mathbf{E}\left[Q\right]\right)^2\right] = \frac{\mathbf{Var}\,Q_h}{N} + \mathbf{E}\left[Q_h - Q\right]^2.$$

- We denote by $\mathscr{C}(\widehat{Q})$ the cost, in terms of the number of floating-point operations required for its evaluation, associated with an estimator $\widehat{Q}$. The cost will often depend on the type of discretization, typically inversely proportional to h or, more generally, satisfying a relation of the form

$$\mathscr{C}(Q_h^{(i)}) \lesssim h^{-\gamma}, \qquad \gamma > 0. \tag{2.12}$$

so that $\mathscr{C}(\widehat{Q}_{h,N}^{\mathrm{MC}}) \lesssim Nh^{-\gamma}$.

- To balance the two error components, assume each ist bounded by $\frac{\epsilon^2}{2}$, resulting in a total bound of ϵ for the RMSE.
- Assuming $\mathbf{Var}\, Q_h$ is independent of h (reasonable, at least for h sufficiently small, since $Q_h \to Q$ as $h \to 0$), this error balance requires

$$N \geq 2\frac{\mathbf{Var}\, Q_h}{\epsilon^2} \qquad \text{and} \qquad h \lesssim \epsilon^{1/\alpha}.$$

- In view of the cost-per-sample assumption (2.12), which for MC sampling implies

$$\mathscr{C}(\widehat{Q}_{h,N}^{\mathrm{MC}}) \lesssim Nh^{-\gamma},$$

we arrive at a scaling relation for the total cost of achieving a RMSE of $O(\epsilon)$ using a standard MC estimator of

$$\mathscr{C}_\epsilon(\widehat{Q}_{h,N}^{\mathrm{MC}}) \lesssim \epsilon^{-2-\gamma/\alpha},$$

to which we refer as the ϵ-cost.

- The idea underlying multilevel estimators is to use realizations of Q_h on different levels, i.e., for different values $h_0, \ldots, h_L$ of the discretization parameter, and decompose $\mathbf{E}[Q_h]$ as

$$\mathbf{E}[Q_h] = \mathbf{E}[Q_{h_0}] + \sum_{\ell=1}^{L} \mathbf{E}\left[Q_{h_\ell} - Q_{h_{\ell-1}}\right] =: \sum_{\ell=0}^{L} \mathbf{E}[Y_\ell],$$

  where $Y_0 := Q_{h_0}$ and

$$h_\ell = s^{-1} h_{\ell-1}, \quad \ell = 1, \ldots, L, \quad s \in \mathbb{N} \setminus \{1\}, \quad h_0 > 0. \tag{2.13}$$

- Given (unbiased) estimators $\{\widehat{Y}_\ell\}_{\ell=0}^{L}$ for $\mathbf{E}[Y_\ell]$, we refer to

$$\widehat{Q}_h^{\mathrm{ML}} := \sum_{\ell=0}^{L} \widehat{Y}_\ell$$

  as a multilevel estimator for Q.
- Since all expectations $\mathbf{E}[Y_\ell]$ are sampled independently, we have

$$\mathbf{Var}\, \widehat{Q}_h^{\mathrm{ML}} = \sum_{\ell=0}^{L} \mathbf{Var}\, \widehat{Y}_\ell.$$

- If each $\widehat{Y}_\ell$ is itself a standard Monte Carlo estimator, i.e.,

$$\widehat{Y}_0 = \widehat{Y}^{\mathrm{MC}}_{0,N_0} := \widehat{Q}^{\mathrm{MC}}_{h_0,N_0}$$

and

$$\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell} := \frac{1}{N_\ell}\sum_{i=0}^{N_\ell}\left(Q^{(i)}_{h_\ell} - Q^{(i)}_{h_{\ell-1}}\right), \qquad \ell = 1,\ldots,L,$$

one obtains a multilevel Monte Carlo estimator, denoted $\widehat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}}$.

Note: The superscript in $Q^{(i)}_{h_\ell} - Q^{(i)}_{h_{\ell-1}}$ indicates that the QoI is obtained for the same realization on both levels.

- The associated MSE then has the standard decomposition

$$\mathbf{E}\left[\left(\widehat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}} - \mathbf{E}\left[Q\right]\right)^2\right] = \sum_{\ell=0}^{L}\frac{\mathbf{Var}\, Y_\ell}{N_\ell} + \mathbf{E}\left[Q_h - Q\right]^2 \tag{2.14}$$

into estimation variance and bias.

- To achieve a balanced RMSE of ϵ, note that the the bias term in (2.14) is the same as for the standard MC estimator, leading again to a choice of $h = h_L$ satisfying $h \lesssim \epsilon^{1/\alpha}$.
- Achieving a bound of $\epsilon^2/2$ for the variance term in the MSE is typically possible at lower cost (or, equivalently, variance reduction for the same cost) than for standard MC for the following two reasons:

(1) If $Q_h \to Q$ not just in mean but also in mean square, then

$$\mathbf{Var}\, Y_\ell = \mathbf{Var}(Q_{h_\ell} - Q_{h_{\ell-1}}) \to 0 \quad \text{as } \ell \to \infty,$$

allowing for smaller and smaller sample sizes N_ℓ on finer and finer levels.

(2) By assumption (2.12), the cost per sample decreases by a factor of s^γ when moving to a coarser level. If this number is sufficiently large, sampling on coarse levels becomes very inexpensive.

- The cost of the MLMC estimator is

$$\mathscr{C}(\widehat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}}) = \sum_{\ell=0}^{L} N_\ell \mathscr{C}_\ell, \qquad \mathscr{C}_\ell := \mathscr{C}(Y_\ell^{(i)}).$$

- Treating the N_ℓ as continuous variables, we can minimize the cost of the MLMC estimator subject to the constraint that its variance is fixed at the value $\epsilon^2/2$. The solution to this constrained minimization problem is obtained as

$$N_\ell \simeq \sqrt{\frac{\mathbf{Var}\, Y_\ell}{\mathscr{C}_\ell}} \qquad \ell = 0, \ldots, L \tag{2.15}$$

with the implied constant given by

$$\frac{2}{\epsilon^2} \sum_{\ell=0}^{M} \sqrt{\mathscr{C}_\ell \,\mathbf{Var}\, Y_\ell}.$$

- This results in a total cost on level ℓ proportional to $\sqrt{\mathscr{C}_\ell \, \mathbf{Var}\, Y_\ell}$ or, more specifically,

$$\mathscr{C}(\widehat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}}) \leq \frac{2}{\epsilon^2}\left(\sum_{\ell=0}^{L}\sqrt{\mathscr{C}_\ell \, \mathbf{Var}\, Y_\ell}\right)^2$$

  For comparison: the cost for standard MC is $\mathscr{C}(\widehat{Q}^{\mathrm{MC}}_{h_L,N}) = \frac{2}{\epsilon^2}\mathscr{C}_L \, \mathbf{Var}\, Q_{h_L}$.
- If $\mathbf{Var}\, Y_\ell$ decays faster than $\mathscr{C}_\ell$ increases, the cost on level $\ell = 0$ dominates, and, since $\mathbf{Var}\, Q_{h_0} \approx \mathbf{Var}\, Q_{h_L}$, the cost ratio of MLMC to ML estimation is approximately

$$\frac{\mathscr{C}_0}{\mathscr{C}_L} \eqsim \left(\frac{h_L}{h_0}\right)^\gamma = s^{-\gamma L}.$$

- If $\mathscr{C}_\ell$ increases faster than $\mathbf{Var}\, Y_\ell$ decays, then the cost on level $\ell = L$ dominates, and then the cost ratio is approximately

$$\frac{\mathbf{Var}\, Y_L}{\mathbf{Var}\, Q_{h_L}} \eqsim \epsilon^2,$$

  (provided $\mathbf{E}\left[(Q - Q_{h_L})^2\right] \eqsim \mathbf{E}\left[Q - Q_{h_L}\right]^2$, which is problem dependent).

Theorem 2.19

Let $\{h_\ell\}_{\ell=0}^L$ satisfy (2.13), $\epsilon < \exp(-1)$, and assume there exist constants α, β, γ, δ, c_{M1}, c_{M2}, $c_{\mathrm{M4}} > 0$ such that $\alpha \geq \frac{1}{2}\min\{\beta, \gamma/\delta\}$ and $\delta \in (\frac{1}{2}, 1]$. Assume further that

(M1) $|\mathbf{E}[Q_{h_\ell}] - \mathbf{E}[Q]| \leq c_{\mathrm{M1}} h_\ell^\alpha$.

(M2) $\mathbf{Var}\, \widehat{Y}_\ell \leq c_{\mathrm{M2}} N_\ell^{-1/\delta} h_\ell^\beta$.

(M3) $$\mathbf{E}\left[\widehat{Y}_\ell\right] = \begin{cases} \mathbf{E}[Q_{h_0}], & \ell = 0, \\ \mathbf{E}\left[Q_{h_\ell} - Q_{h_{\ell-1}}\right], & \ell = 1, \ldots, L. \end{cases}$$

(M4) $\mathscr{C}(\widehat{Y}_\ell) \leq c_{\mathrm{M4}} N_\ell h_\ell^{-\gamma}$.

Then there exist $\{N_\ell\}_{\ell=0}^L$ such that $\mathbf{E}\left[\left(\widehat{Q}_h^{\mathrm{ML}} - \mathbf{E}[Q]\right)^2\right] \leq \epsilon^2$ where $h = h_L$ and

$$\mathscr{C}(\widehat{Q}_h^{\mathrm{ML}}) \leq c \begin{cases} \epsilon^{-2\delta}, & \text{if } \delta\beta > \gamma, \\ \epsilon^{-2\delta} |\log \epsilon|^{1+\delta}, & \text{if } \delta\beta = \gamma, \\ \epsilon^{-2\delta - (\gamma - \delta\beta)/\alpha}, & \text{if } \delta\beta < \gamma, \end{cases}$$

where the constant c depends on $c_{\mathrm{M1}}, c_{\mathrm{M2}}$ and c_{M4}.

- The following MLMC algorithm computes the optimal values of N_ℓ 'on the fly' using (unbiased) sample averages and sample variances of Y_ℓ.
- We assume there exists an $h^\star > 0$ such that the error decay in $|\mathbf{E}[Q_h - Q]|$ is monotonic for $h \leq h^\star$ and satisfies $|\mathbf{E}[Q_h - Q]| \eqsim h^\alpha$.
- This ensures that $|\mathbf{E}[Y_L]| \eqsim h^\alpha$ since $s > 1$ and thus $|\widehat{Y}_L| \eqsim h^\alpha$ for N_L sufficiently large.
- This gives a computable error estimator to determine whether h is sufficiently small or whether L needs to be increased.

Algorithm 1: MLMC algorithm

1 $L \leftarrow 0$.
2 Estimate **Var** Y_L by the sample variance of an initial number of samples.
3 Calculate optimal $\{N_\ell\}_{\ell=1}^{L}$ using (2.15).
4 Evaluate extra samples at each level as needed for the new N_ℓ.
5 **if** $L \geq 1$ **then**
6 test for convergence using $\widehat{Y}_L \approx h^\alpha$.
7 **if** *not converged or* $L = 0$ **then**
8 $L \leftarrow L + 1$ and go back to 2.

- Step 3 aims to make the variance of the MLMC estimator less than $\epsilon^2/2$.
- Step 5 ensures that the remaining bias is less than $\epsilon/\sqrt{2}$.

Contents

Monte Carlo Finite Element Method

We return to our model elliptic boundary value problem with random data

$$-\nabla\cdot(a\nabla u) = f \quad \text{on } D \subset \mathbb{R}^2, \qquad u_{|\partial D} = 0, \tag{2.16}$$

where a and f are random fields defined on D with respect to a probability space $(\Omega, \mathfrak{A}, \mathbf{P})$.

- If f is random, we assume $f(\cdot,\omega) \in L^2(D)$ for (almost) all $\omega \in \Omega$.
- Our goal is to use the MC method to estimate a quantity of interest which depends on the (random) solution u.
  We focus, for now, on the mean $\mathbf{E}\,[u(\boldsymbol{x},\cdot)]$ and variance $\mathbf{Var}\,u(\boldsymbol{x},\cdot)$.
- With each of N i.i.d. realizations $a^{(j)} = a(\cdot,\omega_j)$ and $f^{(j)} = f(\cdot,\omega_j)$ we associate the unique solution $u^{(j)}$, approximate $u_h^{(j)} \approx u^{(j)}$ using the finite element method and compute the ($H_0^1(D)$-valued) estimates

$$\mu_{N,h} := \frac{1}{N}\sum_{j=1}^{N} u_h^{(j)}, \qquad \sigma^2_{N,h} := \frac{1}{N-1}\sum_{j=1}^{N}\left(u_h^{(j)} - \mu_{N,h}\right)^2$$

To ensure a unique solution $u^{(j)}$ for each realization, we could require the coefficient a to satisfy Assumption B.3. However, this proves too restrictive in many applications, and for many cases it is sufficient to require merely realization-wise bounds:

Assumption 2.20

For almost all $\omega \in \Omega$, realizations $a(\cdot,\omega)$ of the coefficient function $a = a(\boldsymbol{x})$ lie in $L^\infty(D)$ and satisfy

$$0 < a_{\min}(\omega) \le a(\boldsymbol{x},\omega) \le a_{\max}(\omega) < \infty \qquad \text{a.e. in } D, \tag{2.17}$$

where

$$a_{\min}(\omega) := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} a(\boldsymbol{x},\omega), \qquad a_{\max}(\omega) := \operatorname*{ess\,sup}_{\boldsymbol{x}\in D} a(\boldsymbol{x},\omega). \tag{2.18}$$

For any realization ω for which Assumption 2.20 holds and $f(\omega) \in L^2(D)$, we may apply the Lax-Milgram lemma and obtain a unique solution of (2.16).

Theorem 2.21

Let Assumption 2.20 hold and $f(\cdot,\omega) \in L^2(D)$ **P**-a.s. Then (2.16) has a unique solution $u(\cdot,\omega) \in H_0^1(D)$ **P**-a.s.

The following theorem provides sufficient conditions for the realization-wise solutions u to have finite p-th moments, i.e., to lie in $L^p(\Omega; H_0^1(D))$.

Theorem 2.22

Under Assumption 2.20, assume the mappings $a : \Omega \to L^\infty(D)$ and $f : \Omega \to L^2(D)$ are measurable, let $V^h \subset H_0^1(D)$ denote a closed subspace and $u_h : \Omega \to V^h$ satisfy **P**-a.s.

$$\int_D a(\boldsymbol{x},\omega)\nabla u_h(\boldsymbol{x},\omega)\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x},\omega)v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall v \in V^h.$$

Then, with C_D the Poincaré-Friedrichs constant from Lemma B.7:

(a) If $f \in L^2(D)$ is deterministic, then $1/a_{\min} \in L^p(\Omega;\mathbb{R})$ with $p \geq 1$ implies

$$\|u_h\|_{L^p(\Omega;H_0^1(D))} \leq C_D \|a_{\min}^{-1}\|_{L^p(\Omega;\mathbb{R})} \|f\|_{L^2(D)}.$$

(b) If $1/a_{\min} \in L^q(\Omega;\mathbb{R})$ and $f \in L^r(\Omega;L^2(D))$ with $q, r \geq 1$, $1/q + 1/r = 1/p \leq 1$, then

$$\|u_h\|_{L^p(\Omega;H_0^1(D))} \leq C_D \|a_{\min}^{-1}\|_{L^q(\Omega;\mathbb{R})} \|f\|_{L^r(\Omega;L^2(D))}.$$

If, in addition, a and f are independent, the above bound holds with $q = r = p$.

We now assume the solution possesses H^2-regularity combined with a moment bound:

Assumption 2.23

There exists a constant $K_2 > 0$ such that, for every $f \in L^2(\Omega; L^2(D))$, we have $u \in L^4(\Omega; H^2(D))$ and

$$|u|_{L^4(\Omega;H^2(D))} := \mathbf{E}\left[|u|^4_{H^2(D)}\right]^{1/4} \leq K_2 \|f\|_{L^2(\Omega;L^2(D)}.$$

Theorem 2.24

Under the conditions of Theorem 2.21 together with Assumption 2.23 and assuming that $a_{\min}^{-1/2} a_{\max}^{1/2} \in L^4(\Omega; \mathbb{R})$, the piecewise linear finite element approximation u_h with respect to a shape-regular triangulation $\mathscr{T}_h$ satisfies

$$\|u - u_h\|_{L^2(\Omega;H^1_0(D))} \leq Kh \|a_{\min}^{-1/2} a_{\max}^{1/2}\|_{L^4(\Omega;\mathbb{R})} \|f\|_{L^2(\Omega)}.$$

We split the error in approximating $\mathbf{E}[u]$ by the MC estimate $\mu_{N,h}$ in the $H_0^1(D)$-norm as

$$\|\mathbf{E}[u] - \mu_{N,h}\|_{H_0^1(D)} \le \underbrace{\|\mathbf{E}[u] - \mathbf{E}[u_h]\|_{H_0^1(D)}}_{\text{discretization error}} + \underbrace{\|\mathbf{E}[u_h] - \mu_{N,h}\|_{H_0^1(D)}}_{\text{MC error}}.$$

For the discretization error we obtain, using Jensen's inequality noting that norms are convex functions,

$$\|\mathbf{E}[u - u_h]\|_{H_0^1(D)} \le \mathbf{E}\left[\|u - u_h\|_{H_0^1(D)}\right] = \left(\mathbf{E}\left[\|u - u_h\|_{H_0^1(D)}\right]^2\right)^{1/2}$$

and again for the convex function $\phi(x) = x^2$ to obtain

$$\|\mathbf{E}[u - u_h]\|_{H_0^1(D)} \le \mathbf{E}\left[\|u - u_h\|^2_{H_0^1(D)}\right]^{1/2} = \|u - u_h\|_{L^2(\Omega;H_0^1(D)},$$

which is $O(h)$ by Theorem 2.24.

Alternatively, one can split the Monte Carlo Finite Element error using the analogous MSE splitting into variance and squared bias, where the distance between the ($H_0^1(D)$-valued) quantities $\mathbf{E}[u]$ and $\mu_{N,h}$ is measured w.r.t. the norm in this space:

$$\mathbf{E}\left[\|\mathbf{E}[u]-\mu_{N,h}\|^2_{H_0^1(D)}\right] = \underbrace{\mathbf{E}\left[\|\mu_{N,h}-\mathbf{E}[\mu_{N,h}]\|^2_{H_0^1(D)}\right]}_{=\mathbf{Var}\,\mu_{N,h}} + \underbrace{\|\mathbf{E}[u-u_h]\|^2_{H_0^1(D)}}_{=\mathbf{bias^2}}$$

This follows by inserting $\mathbf{E}[u]-\mu_{N,h} = \mathbf{E}[u]-\mathbf{E}[\mu_{N,h}]+\mathbf{E}[\mu_{N,h}]-\mu_{N,h}$, writing the squared norm in terms of the inner product, using unbiasedness $\mathbf{E}[\mu_{N,h}] = \mathbf{E}[u_h]$ and noting that the cross terms cancel out due to

$$\mathbf{E}[(X,y)] = (\mathbf{E}[X],y)$$

for any Hilbert space-valued random variable X and fixed element y of the same space.

Theorem 2.25

Under the conditions of Theorem 2.22 there holds

$$\mathbf{E}\left[\|\mathbf{E}[u_h] - \mu_{N,h}\|^2_{H^1_0(D)}\right] \leq \frac{K}{N}$$

with a constant K independent of h.

Corollary 2.26

Under the conditions of Theorem 2.22 there holds for any $\epsilon > 0$

$$\mathbf{P}\left(\|\mathbf{E}[u_h] - \mu_{N,h}\|_{H^1_0(D)} \geq N^{-1/2+\epsilon}\right) \leq LN^{-2\epsilon}$$

for a constant $L > 0$ independent of h.

- **Result:** The total error of estimating the mean $\mathbf{E}[u]$ of the solution of (2.16) using a piecewise linear FE discretization with mesh size h and a MC sample size of N decays at the rate

$$\|\mathbf{E}[u] - \mu_{N,h}\|_{H_0^1(D)} = O(h) + O(N^{-1/2}), \qquad h \to 0,\ N \to \infty.$$

- This is already very slow convergence, and, particularly for low-regularity solutions as arise, e.g., in groundwater flow applications, more advanced techniques such as MLMC methods are attractive.

- Recalling Theorem 2.19, we note that for rough problems we are typically in the regime $\beta < \gamma$. For standard MC estimators on each level ($\delta = 1$) and, as is typical, $\beta = 2\alpha$, we obtain a cost on the order of $\epsilon^{-\gamma/\alpha}$, which is asymptotically the cost of computing one sample on a mesh sufficiently fine to approximate one realization with sufficient spatial accuracy.