Mathematische Methoden der Unsicherheitsquantifizierung

Oliver Ernst

Professur Numerische Mathematik

Sommersemester 2014

TECHNISCHE UNIVERSITÄT
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

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

- 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:

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

- 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	3.3	6.1e-2
100	32	3.12	5.2e-3
1000	330	3.0	3.5e-2
10000	3188	3.13	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 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

$$\mathbf{E}\left[\left(\frac{S_N}{N}-\mu\right)^2\right] = \mathbf{Var}\,\frac{S_N}{N} = \frac{\sigma^2}{N} \to 0,$$

we have L^2-convergence of S_N/N to μ and, by Theorem A.25, for any $\epsilon > 0$,

$$\mathbf{P}\left\{\left|\frac{S_N}{N}-\mu\right| > N^{-1/2+\epsilon}\right\} \leqslant \frac{\sigma^2}{N^{2\epsilon}}, \tag{2.1}$$

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

$$|\mathbf{P}\{S_N^* \leqslant x\} - \Phi(x)| \leqslant C\frac{\rho}{\sigma^3\sqrt{N}}, \tag{2.2}$$

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

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

and therefore

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

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,

$$-B_N \leqslant \mathbf{P}(|S_N^*| \leqslant c) - \Phi(x) \leqslant B_N$$

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

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

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

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

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

$$\left|\frac{S_N}{N} - p\right| = \left|\frac{1808}{3404} - \frac{5}{3\pi}\right| =: \epsilon_L \approx 4.5 \cdot 10^{-8}.$$

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 ϵ_L for $N > 4.9094 \cdot 10^{14}$.

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

$$B_N = C\frac{\rho}{\sigma^3\sqrt{N}} = C\frac{1-2p+2p^2}{\sqrt{p(1-p)N}} \leqslant \frac{0.3116}{\sqrt{N}},$$

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

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

giving

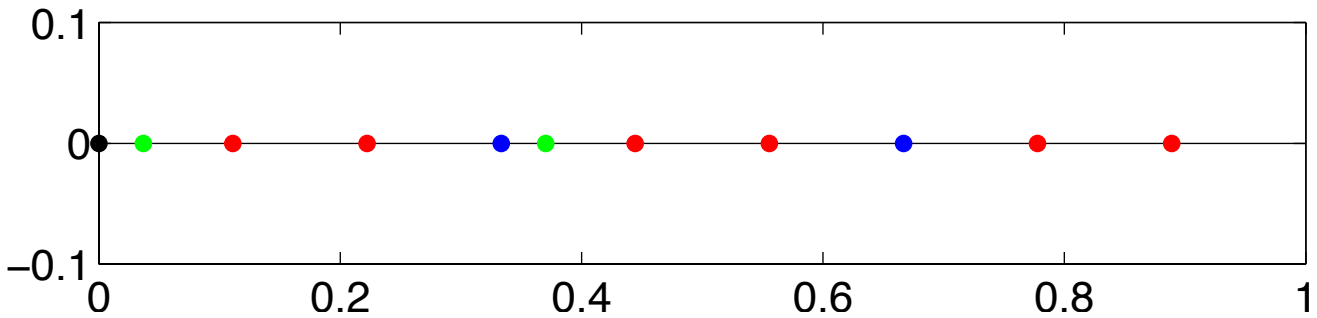
$$\mathbf{P}\left(\left|\frac{S_N}{N} - p\right| \leqslant \epsilon_L\right) \leqslant \gamma_{R_L} + 2B_{3408} \approx 0.0107.$$

Contents

In quasi-Monte Carlo methods, the samples are not chosen randomly, but special (deterministic) number sequences, known as low-discrepancy sequences, are used instead. Discrepancy is a measure of equidistribution of a number sequence.

Example: The van der Corput sequence to base 3 is such a low-discrepancy sequence for the unit interval. It is given by $x_n = \frac{k}{3^j}$, where j increases monotonically and, for each j, k runs through all nonnegative integers such that $k/3^j$ is an irreducible fraction. The ordering in k is obtained by representing k in base 3 and reversing the digits. The first 11 numbers are

$$\{x_n\}_{n=1}^{11} = 0, \frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \frac{1}{27}, \frac{10}{27}.$$



- Replacing i.i.d. random numbers sampled from $\mathrm{U}[0,1]$ in a standard Monte Carlo approximation of $\mathbf{E}\left[f(X)\right]$ for some $f \in C^\infty(0,1)$ and $X \sim \mathrm{U}[0,1]$, by the van der Corput sequence of length N, yields a quasi-Monte Carlo method.
- The convergence rate is improved from $O(N^{-1/2})$ to $O(N^{-2})$.
- Although this improvement is impressive, the method does not generalise easily and the rate of convergence depends on the problem.
- In particular, the rate of convergence for a quasi-Monte Carlo method generally does depend on the dimension.

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.

- 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 2.2

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

$$\mathbf{Var}\,\frac{S_N + \tilde{S}_N}{2N} = \mathbf{Var}\,\frac{S_{2N}}{2N} + \frac{1}{2}\,\mathbf{Cov}\left(\frac{S_N}{N}, \frac{\tilde{S}_N}{N}\right) \leqslant \mathbf{Var}\,\frac{S_N}{N}. \tag{2.3}$$

- Worst case: Variance of average of N samples and N antithetic samples less than variance of N independent samples.
- Best case: 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.

Monte Carlo Methods

Example: Predator-prey dynamical system

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 $\mathbf{E}\left[u_1(T)\right]$ 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 $\mathbf{E}\left[\phi(\boldsymbol{u}_{n_T})\right]$ 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.

- **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:	$\widehat{Q}_{\Delta t} := \widehat{Q}_{\Delta t,N} \approx \mathbf{E}\left[Q_{\Delta t}\right] \approx \mathbf{E}\left[Q\right].$

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

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

- Explicit Euler error:
$$\|\boldsymbol{u}(T) - \boldsymbol{u}^{\Delta t}(T)\| \leqslant K\Delta t.$$
- ϕ Lipschitz-continuous with constant $L = 1$:
$$|\phi(\boldsymbol{u}(T)) - \phi(\boldsymbol{u}^{\Delta t}(T))| \leqslant K\,L\,\Delta t.$$
- Therefore
$$|\mathbf{E}\left[Q\right] - \mathbf{E}\left[Q_{\Delta t}\right]| = |\mathbf{E}\left[Q - Q_{\Delta t}\right]| \leqslant K\,L\,\Delta t. \tag{2.4}$$

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

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

- Combined with (2.4):

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

- Balance discretization and MC errors:

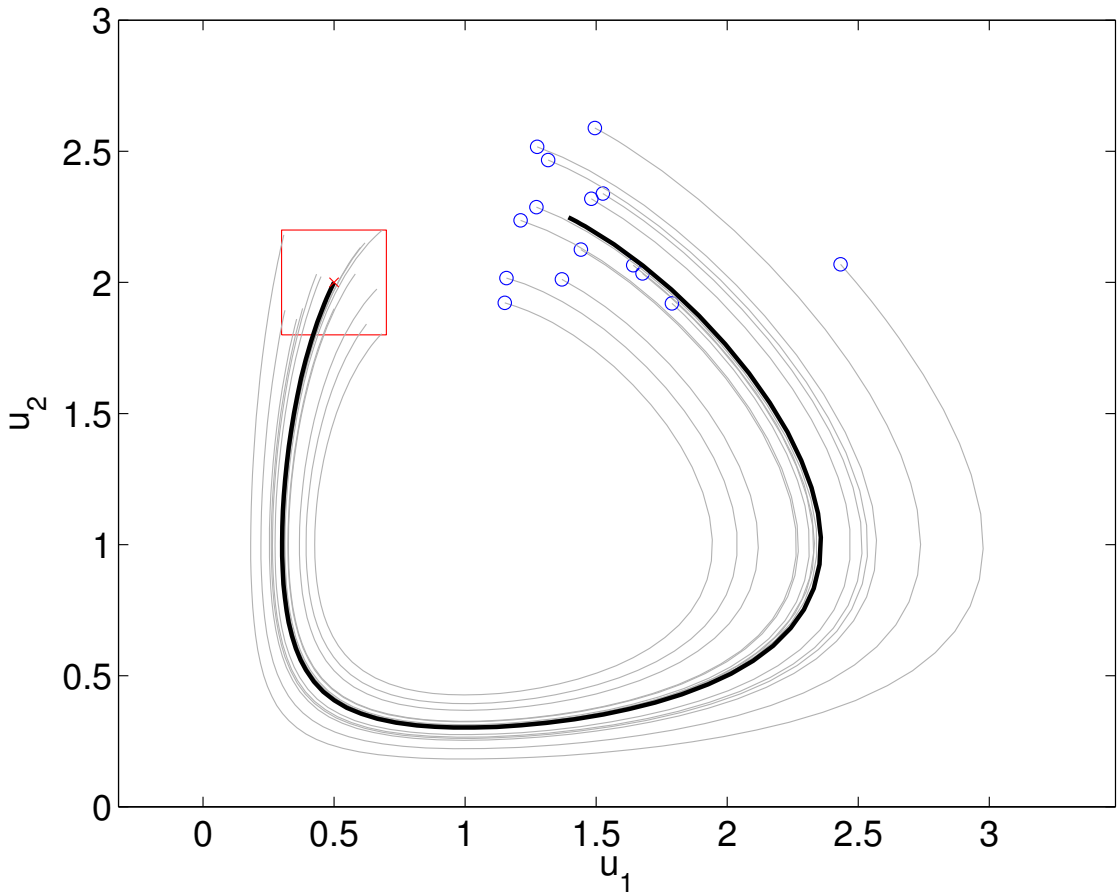
$$KL\Delta t \approx \frac{\delta}{2}, \quad \frac{2\sigma}{\sqrt{N}} \approx \frac{\delta}{2},$$

leads to

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

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 $u_1(T) = 1.3942$.

We may 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 $\boldsymbol{u}_0$.

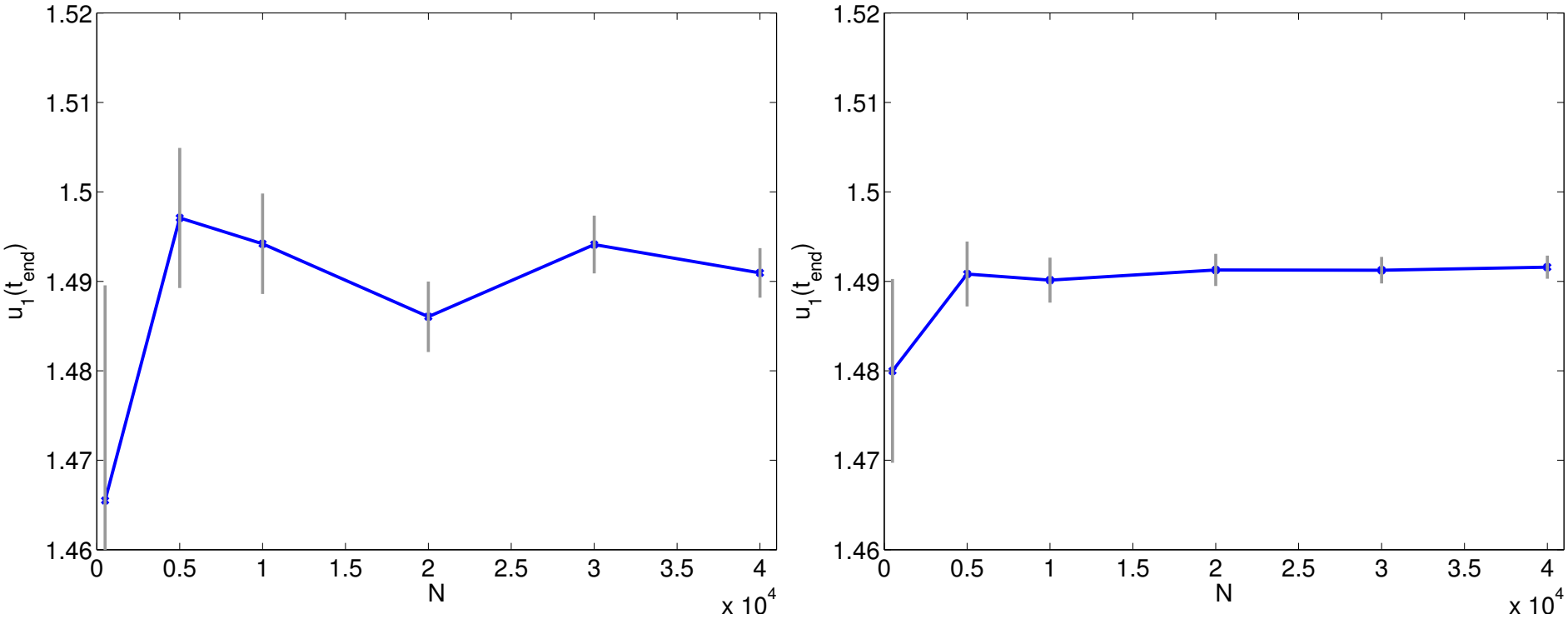
- Denote by $X_k = \phi(\boldsymbol{u}^{\Delta t}(T))$ the basic samples, by $\tilde{X}_k$ the antithetic counterparts. Note that all pairs of samples are independent except each sample and its antithetic counterpart.
- We estimate $\overline{S}_{2N}$ using the sample variance.
- To estimate $\frac{1}{2}\left(\overline{S}_N + \overline{\tilde{S}}_N\right)$ by (2.3), 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.

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.,

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

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

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

  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

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

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

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

giving

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

- 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

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

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

- 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 **Var** Q_h is approximately constant independent of h, this error balance requires

$$N \gtrsim \epsilon^{-2} \qquad \text{and} \qquad h \lesssim \epsilon^{1/\alpha}.$$

- Since the cost per sample was assumed to satisfy $\mathscr{C}(Q_h^{(i)}) \lesssim h^{-\gamma}$, this gives

$$\mathscr{C}(\hat{Q}_{h,N}^{\mathsf{MC}}) \lesssim N h^{-\gamma},$$

whereby the total cost of achieving a RMSE of $O(\epsilon)$ using a standard MC estimator is

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

- 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

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

  where

$$h_\ell = s^{-1} h_{\ell-1}, \qquad h_0 > 0, \quad \ell = 1, \ldots, L, \quad s \in \mathbb{N}\backslash\{1\}. \tag{2.5}$$

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

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

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

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

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

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

and

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

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

- The associated MSE then has the standard decomposition

$$\mathbf{E}\left[\left(\widehat{Q}_{h,\{N_\ell\}}^{\mathsf{MLMC}} - \mathbf{E}\left[Q\right]\right)^2\right] = \sum_{\ell=0}^{L}\frac{\mathbf{Var}\, Y_\ell}{N_\ell} + \mathbf{E}\left[Q_h - Q\right]^2 \tag{2.6}$$

into estimation variance and bias.

- To achieve a balanced RMSE of ϵ, note that the the bias term in (2.6) is the same as for the standard MC estimator, leading again to a choise 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 than for standard MC for the following two reasons:
- If $Q_h \to Q$ also in mean square, then $\mathbf{Var}\, Y_\ell = \mathbf{Var}(Q_{h_\ell} - Q_{h_{\ell-1}}) \to 0$ as $\ell \to \infty$, allowing for smaller and smaller sample sizes N_ℓ on finer and finer levels.
- As $\epsilon \to 0$, the discretization parameter h_0 on the coarsest level can remain fixed, leading to fixed cost per sample there.

- The cost of the MLMC estimator is

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

- Treating the N_ℓ as continuous variables, the variance of the MLMC estimator is minimized for a fixed cost for

$$N_\ell \simeq \sqrt{\frac{\mathbf{Var}\, Y_\ell}{\mathscr{C}_\ell}} \tag{2.7}$$

  with the implied constant chosen to make the total variance equal to $\epsilon^2/2$.
- This results in a total cost on level ℓ proportional to $\sqrt{\mathscr{C}_\ell \mathbf{Var}\, Y_\ell}$ and therefore

$$\mathscr{C}(\hat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}}) \lesssim \sum_{\ell=0}^{L} \sqrt{\mathscr{C}_\ell \mathbf{Var}\, Y_\ell}$$

- If $\mathbf{Var}\, Y_\ell$ decays faster than $\mathscr{C}_\ell$ increases, the cost on level $\ell = 0$ dominates, and, since $N_0 \eqsim \epsilon^{-2}$, the cost ratio of MLMC to ML estimation is approximately
$$\frac{\mathscr{C}_0}{\mathscr{C}_L} \eqsim \left(\frac{h_L}{h_0}\right)^\gamma.$$
- 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
$$\frac{\mathbf{Var}\, Y_L}{\mathbf{Var}\, Y_0},$$
which is $O(\epsilon^2)$ if h_0 is such that $\mathbf{Var}\, Y_0 \eqsim \mathbf{Var}\, Q_{h_0}$.

Theorem 2.3

Let $\{h_\ell\}_{\ell=0}^L$ satisfy (2.5), $\epsilon < \exp(-1)$, and assume there exist constants $\alpha, \beta, \gamma, \delta, c_{\mathsf{M1}}, c_{\mathsf{M2}}, c_{\mathsf{M4}} > 0$ such that $\alpha \geqslant \min\{\beta, \gamma/\delta\}$ and $\delta \in (\frac{1}{2}, 1]$.
Assume further that

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

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

(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) $\mathscr{C}(\widehat{Y}_\ell) \leqslant N_\ell h_\ell^{-\gamma}$.

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

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

where the constant c depends on $c_{\mathsf{M1}}, c_{\mathsf{M2}}$ 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 $h \leqslant h^\star$ 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 (2.7).
4 Evaluate extra samples at each level as needed for the new N_ℓ.
5 **if** $L \geqslant 1$ **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

$$-\nabla\cdot(a\nabla u) = f, \quad \text{on } D\subset\mathbb{R}^2, \qquad u_{|\partial D} = 0, \tag{2.8}$$

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 2.4

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

$$0 < a_{\min}(\omega) \leqslant a(\boldsymbol{x},\omega) \leqslant a_{\max}(\omega) < \infty \qquad \text{a.e. in } D, \tag{2.9}$$

where

$$a_{\min}(\omega) := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} a(\boldsymbol{x},\omega), \qquad a_{\max}(\omega) := \operatorname*{ess\,sup}_{\boldsymbol{x}\in D} a(\boldsymbol{x},\omega). \tag{2.10}$$

For any realization ω for which Assumption 2.4 holds and $f(\omega) \in L^2(D)$, we may apply the Lax-Milgram lemma and obtain a unique solution of (2.8).

Theorem 2.5

Let Assumption 2.4 hold and $f(\cdot,\omega) \in L^2(D)$ **P**-a.s. Then (2.8) 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 2.6

Under Assumption 2.4, 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 $p \geqslant 1$ implies

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

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

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

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

Assumption 2.7

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

$$|u|_{L^4(\Omega;H^2(D))} \leqslant K_2 \|f\|_{L^2(\Omega;L^2(D)}.$$

Theorem 2.8

Under the conditions of Theorem 2.5 together with Assumption 2.7 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

$$\|u - u_h\|_{L^2(\Omega;H_0^1(D))} \leqslant Kh \|a_{\min}^{-1/2} a_{\max}^{1/2}\|_{L^4(\Omega;\mathbb{R})} \|f\|_{L^2(\Omega)}.$$

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

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

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

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

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

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

which is $O(h)$ by Theorem 2.8.

Theorem 2.9

Under the conditions of Theorem 2.6 there holds

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

with a constant K independent of h.

Corollary 2.10

Under the conditions of Theorem 2.6 there holds for any $\epsilon > 0$

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

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

- **Result:** The total error of estimating the mean $\mathbf{E}[u]$ of the solution of (2.8) 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 2.3, 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.