Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Ninth lecture, June 23, 2025

We continue studying the *uniform and affine model*: let $D \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded Lipschitz domain, let $f \in L^2(D)$, and let $U := [-1/2, 1/2]^{\mathbb{N}} := \{(a_j)_{j\geq 1} : -1/2 \leq a_j \leq 1/2\}$ be a set of parameters. Consider the problem of finding, for all $\boldsymbol{y} \in U$, $u(\cdot, \boldsymbol{y}) \in H_0^1(D)$ such that

$$\int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla u(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \text{for all } v \in H_0^1(D),$$

where the diffusion coefficient has the parameterization

$$a(\boldsymbol{x}, \boldsymbol{y}) := a_0(\boldsymbol{x}) + \sum_{j=1}^{\infty} y_j \psi_j(\boldsymbol{x}), \quad \boldsymbol{x} \in D, \ \boldsymbol{y} \in U,$$

where we assume

(A1) $a_0 \in L^\infty(D)$ and $\psi_j \in L^\infty(D)$ for all $j \in \mathbb{N}$,

(A2) there exist $a_{\min}, a_{\max} > 0$ s.t. $0 < a_{\min} \leq a(\boldsymbol{x}, \boldsymbol{y}) \leq a_{\max} < \infty$ for all $\boldsymbol{x} \in D$ and $\boldsymbol{y} \in U$,

(A3) $\sum_{j=1}^{\infty} \|\psi_j\|_{L^\infty(D)}^p < \infty$ for some $p \in (0,1)$.

(Note that (A3) implies that $\sum_{j=1}^{\infty} \|\psi_j\|_{L^\infty(D)} < \infty$.)

Let $u_s(\cdot,\boldsymbol{y}) := u_s(\cdot,(y_1,\ldots,y_s,0,0,\ldots))$ denote the dimensionally-truncated PDE solution for $\boldsymbol{y}\in U$ (we sometimes also write $u_s(\cdot,\boldsymbol{y})$ for $\boldsymbol{y}\in[-1/2,1/2]^s$), and let $u_{s,h}(\cdot,\boldsymbol{y})\in V_h$ denote the dimensionally-truncated FE solution in the FE space spanned by piecewise linear FE basis functions. Let $G\colon H_0^1(D)\to\mathbb{R}$ be a bounded linear functional.

During the last lecture, we split the overall approximation error as

$$\begin{aligned}
&\bigg|\int_{[-1/2,1/2]^{\mathbb{N}}} G(u(\cdot,\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}-\frac{1}{n}\sum_{i=1}^n G(u_{s,h}(\cdot,\boldsymbol{t}_i))\bigg|\\
&\leq\bigg|\int_{[-1/2,1/2]^{\mathbb{N}}}(G(u(\cdot,\boldsymbol{y})-u_s(\cdot,\boldsymbol{y})))\,\mathrm{d}\boldsymbol{y}\bigg| && \text{(dimension-truncation error)}\\
&+\bigg|\int_{[-1/2,1/2]^{s}} G(u_s(\cdot,\boldsymbol{y})-u_{s,h}(\cdot,\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}\bigg| && \text{(finite element error)}\\
&+\bigg|\int_{[-1/2,1/2]^{s}} G(u_{s,h}(\cdot,\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}-\frac{1}{n}\sum_{i=1}^n G(u_{s,h}(\cdot,\boldsymbol{t}_i))\bigg|, && \text{(cubature error)}
\end{aligned}$$

and found that it is possible to construct a QMC point set $\boldsymbol{t}_i := \{\frac{i\boldsymbol{z}}{n}\}$ satisfying the QMC cubature error rate $\mathcal{O}(\varphi(n)^{\max\{-1/p+1/2,-1+\delta\}})$, where the implied coefficient is independent of s, n, and h, and $\delta\in(0,1/2)$ is arbitrary. Let us consider the other error contributions next.

Some auxiliary results

Neumann series: "Sufficiently small perturbations of the identity are still invertible"

We will require the following well-known generalization of the geometric series formula, named after 19^{th} century mathematician Carl Neumann.

Theorem (Neumann series)

Let H be a Hilbert space and let $A \in \mathcal{L}(H)$ be a bounded linear functional with operator norm $\|A\| < 1$. Then $I - A$ is invertible in $\mathcal{L}(H)$ with

$$(I - A)^{-1} = I + A + \cdots + A^n + \cdots = \sum_{k=0}^{\infty} A^k,$$

and this series converges in operator norm.

Proof. Let $B_{m,n} := \sum_{k=m}^{n} A^k$, $m < n$. Since $\|A\| < 1$, we have

$$\|B_{m,n}\| \leq \sum_{k=m}^{n} \|A\|^k = \|A\|^m \sum_{k=0}^{m-n} \|A\|^k = \|A\|^m \frac{1 - \|A\|^{n-m+1}}{1 - \|A\|} \xrightarrow{m,n\to\infty} 0.$$

$\therefore$ The partial sums $\sum_{k=0}^{n} A^k$ form a Cauchy sequence in $\mathcal{L}(H)$.

Since H is a Hilbert space, $\mathcal{L}(H)$ is a Banach space and the limit

$$B := \lim_{n\to\infty} \sum_{k=0}^{n} A^k \in \mathcal{L}(H)$$

exists. We need to prove that $(I - A)B = I = B(I - A)$. Let

$$B_n := I + A + \cdots + A^n.$$

Then

$$(I - A)B_n = I - A^{n+1},$$
$$B_n(I - A) = I - A^{n+1},$$

and since $\|A\| < 1$, $\|A^{n+1}\| \leq \|A\|^{n+1} \xrightarrow{n\to\infty} 0$, we thus obtain

$$I - A^{n+1} \xrightarrow{n\to\infty} I \quad \text{in } \mathcal{L}(H)$$

and

$$(I - A)B = \lim_{n\to\infty} (I - A)B_n = I = \lim_{n\to\infty} B_n(I - A) = B(I - A). \quad \square$$

Multinomial theorem

The multinomial theorem is a generalization of Newton's binomial formula. Using multi-index notation, it can be expressed as

$$(x_1 + \cdots + x_s)^k = \sum_{\substack{|\boldsymbol{\nu}|=k \\ \boldsymbol{\nu}\in\mathbb{N}_0^s}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{x}^{\boldsymbol{\nu}}.$$

In fact, if $\boldsymbol{x} := (x_j)_{j=1}^{\infty} \in \ell^1$, then we have

$$\left(\sum_{j=1}^{\infty} x_j\right)^k = \sum_{\substack{|\boldsymbol{\nu}|=k \\ \boldsymbol{\nu}\in\mathscr{F}}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{x}^{\boldsymbol{\nu}}$$

and we will later require the following special case:

$$\left(\sum_{j=s+1}^{\infty} x_j\right)^k = \sum_{\substack{|\boldsymbol{\nu}|=k \\ \boldsymbol{\nu}\in\mathscr{F} \\ \nu_j=0\ \forall j\leq s}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{x}^{\boldsymbol{\nu}}. \tag{1}$$

The following lemma frequently appears in the context of best N-term approximation.

Lemma (Stechkin's lemma)

Let Λ be a countable index set, let $0 < p \le q < \infty$, and let $(a_{\nu})_{\nu \in \Lambda}$ be a sequence. Let $\emptyset \neq \Lambda_N \subset \Lambda$ be a set of indices containing the N largest terms of the sequence $(a_{\nu})_{\nu \in \Lambda}$. Then

$$\left(\sum_{\nu \in \Lambda \setminus \Lambda_N} |a_{\nu}|^q \right)^{1/q} \le N^{-r} \left(\sum_{\nu \in \Lambda} |a_{\nu}|^p \right)^{1/p}, \quad r = \frac{1}{p} - \frac{1}{q}.$$

Proof. WLOG, we can relabel the a-sequence so that $(a_j)_{j \ge 1}$ is non-increasing, i.e., $a_{j+1} \le a_j$ for all $j \ge 1$. We obtain

$$\left(\sum_{j=N+1}^{\infty} |a_j|^q \right)^{1/q} = \left(\sum_{j=N+1}^{\infty} |a_j|^{q-p} |a_j|^p \right)^{1/q} \le |a_N|^{1-p/q} \left(\sum_{j=N+1}^{\infty} |a_j|^p \right)^{1/q}$$

$$\le |a_N|^{1-p/q} \left(\sum_{j \ge 1}^{\infty} |a_j|^p \right)^{1/q}.$$

The key is to bound $|a_N|^{1-p/q}$ in terms of N.

Standard technique: the monotonicity of the a-sequence implies that

$$N|a_N|^p = |a_N|^p + \cdots + |a_N|^p \leq |a_1|^p + \cdots + |a_N|^p \leq \sum_{j\geq 1} |a_j|^p$$

$$\Rightarrow \quad |a_N|^p \leq N^{-1} \sum_{j\geq 1} |a_j|^p.$$

Hence

$$|a_N|^{1-p/q} = |a_N|^{pr} \leq N^{-r} \left(\sum_{j\geq 1} |a_j|^p \right)^r.$$

Plugging this into the inequality on the previous page yields

$$\left(\sum_{j=N+1}^{\infty} |a_j|^q \right)^{1/q} \leq |a_N|^{1-p/q} \left(\sum_{j\geq 1}^{\infty} |a_j|^p \right)^{1/q} \leq N^{-r} \left(\sum_{j\geq 1} |a_j|^p \right)^{r+1/q}$$

$$= N^{-r} \left(\sum_{j\geq 1} |a_j|^p \right)^{1/p},$$

where the final equality follows from the definition $r = 1/p - 1/q$. □

Dimension truncation error

Remark about infinite-dimensional integrals

Recall that $U := [-1/2, 1/2]^{\mathbb{N}}$. We will be discussing infinite-dimensional Lebesgue integrals of the form

$$\int_U f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y},$$

where we have the infinite tensor product measure

$$\mathrm{d}\boldsymbol{y} := \bigotimes_{j=1}^{\infty} \mathrm{d}y_j.$$

The σ-algebra $\mathcal{F}$ for $\mathrm{d}\boldsymbol{y}$ is generated by finite rectangles $\prod_{j=1}^{\infty} S_j$, where only a finite number of S_j are different from $[-1/2, 1/2]$ and those that are different are contained in $[-1/2, 1/2]$. The resulting triplet $(U, \mathcal{F}, \mathrm{d}\boldsymbol{y})$ is a probability space.

For in-depth measure-theoretic considerations cf., e.g., “Measure Theory” by Halmos.

For the purposes of this course, we can regard infinite-dimensional integrals as limits of finite-dimensional integrals in the following sense:

$$\int_U f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \lim_{s\to\infty} \int_{[-1/2,1/2]^s} f(y_1,\ldots,y_s,0,0,\ldots)\,\mathrm{d}y_1\cdots\mathrm{d}y_s. \tag{2}$$

The justification for this can be found, e.g., in "Infinite-dimensional integration and the multivariate decomposition method" by Kuo, Nuyens, Plaskota, Sloan, and Wasilkowski (J. Comput. Appl. Math., 2017). The result is stated below without proof. (Homework: verify that the conditions of the following theorem are valid for our PDE model problem.)

Theorem (Kuo *et al.* 2017)

Let $f\colon U\to\mathbb{R}$ be integrable w.r.t. the measure $\mathrm{d}\boldsymbol{y} := \bigotimes_{j=1}^{\infty}\mathrm{d}y_j$ which satisfies

$$\lim_{s\to\infty} f(y_1,\ldots,y_s,0,0,\ldots) = f(\boldsymbol{y}) \quad \text{for a.e. } \boldsymbol{y}\in U,$$

$$|f(y_1,\ldots,y_s,0,0,\ldots)| \le |g(\boldsymbol{y})| \quad \text{for a.e. } \boldsymbol{y}\in U$$

for some integrable function $g\colon U\to\mathbb{R}$ w.r.t. the measure $\mathrm{d}\boldsymbol{y}$. Then the characterization (2) holds.

The following rate was proved in "Dimension truncation in QMC for affine-parametric operator equations" by Gantner (MCQMC 2016).

Theorem (Dimension truncation error)

Suppose that the assumptions $(\mathrm{A1})$–$(\mathrm{A3})$ *hold and* $\|\psi_1\|_{L^\infty(D)} \geq \|\psi_2\|_{L^\infty(D)} \geq \|\psi_3\|_{L^\infty(D)} \geq \cdots$. *Then for every* $f \in L^2(D)$ *and every bounded linear functional* $G\colon H_0^1(D) \to \mathbb{R}$*, there holds*

$$\left| \int_U G(u(\cdot,\boldsymbol{y}) - u_s(\cdot,\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} \right| \leq C \frac{\|f\|_{L^2(D)} \|G\|_{H_0^1(D)\to\mathbb{R}}}{a_{\min}} s^{-\frac{2}{p}+1},$$

where the constant $C > 0$ *is independent of* s*,* f*, and* G*.*

Intermezzo

The dimension truncation proof is based on recasting the variational formulation as an affine-parametric operator equation. Specifically, if $u(\cdot, \boldsymbol{y})$ denotes the parametric PDE solution and f the source term, we require for the analysis the (linear) *forward operator*

$$A(\boldsymbol{y})\colon u(\cdot, \boldsymbol{y}) \mapsto f$$

and the *solution operator*

$$A(\boldsymbol{y})^{-1}\colon f \mapsto u(\cdot, \boldsymbol{y}).$$

To this end, we need to be careful with the function space setting (the domains and codomains of $A(\boldsymbol{y})$ and $A(\boldsymbol{y})^{-1}$).

First of all, let us denote the dual space of $H_0^1(D)$ as

$$H^{-1}(D) := (H_0^1(D))' := \{F\colon H_0^1(D) \to \mathbb{R} \mid F \text{ is linear and bounded}\}.$$

(This is a Hilbert space as a consequence of Riesz representation theorem.)

Let $F \in H^{-1}(D)$ and $v \in H_0^1(D)$. Then the *duality pairing* of F and v is defined as

$$\langle F, v\rangle_{H^{-1}(D),H_0^1(D)} := F(v).$$

In a certain sense, the element $F \in H^{-1}(D)$ is defined by its *action* on the elements of $H_0^1(D)$. For example, fix some $f \in L^2(D)$. Then (weighted) integration over (parts of) the domain D, e.g.,

$$\langle F, v\rangle_{H^{-1}(D),H_0^1(D)} := \int_D f(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

would be an example of an element of $H^{-1}(D)$.

Let $\boldsymbol{y} \in U$ and consider the bilinear form

$$B_{\boldsymbol{y}}(v, w) = \int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla v(\boldsymbol{x}) \cdot \nabla w(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \quad v, w \in H_0^1(D).$$

Now

$$B_{\boldsymbol{y}}(v, w) \leq a_{\max} \|v\|_{H_0^1(D)} \|w\|_{H_0^1(D)}, \quad v, w \in H_0^1(D), \qquad \text{(boundedness)}$$

$$|B_{\boldsymbol{y}}(v, v)| \geq a_{\min} \|v\|^2_{H_0^1(D)}, \quad v \in H_0^1(D). \qquad \text{(coercivity)}$$

Then the Lax–Milgram lemma implies that for any $F \in H^{-1}(D)$, there exists a unique element $u(\cdot, \boldsymbol{y}) \in H_0^1(D)$ such that

$$B_{\boldsymbol{y}}(u(\cdot, \boldsymbol{y}), v) = F(v) \quad \text{for all } v \in H_0^1(D)$$

and

$$\|u(\cdot, \boldsymbol{y})\|_{H_0^1(D)} \leq \frac{\|F\|_{H^{-1}(D)}}{a_{\min}}.$$

Especially, the linear map $A(\boldsymbol{y})\colon H_0^1(D) \to H^{-1}(D)$, $u(\boldsymbol{y}) \mapsto F$, is boundedly invertible[†] with

$$\|A(\boldsymbol{y})\|_{H_0^1(D) \to H^{-1}(D)} \leq a_{\max} \quad \text{and} \quad \|A(\boldsymbol{y})^{-1}\|_{H^{-1}(D) \to H_0^1(D)} \leq \tfrac{1}{a_{\min}}.$$

[†] Not trivial! See, e.g., Remark 2.7 in "Theory and Practice of Finite Elements" by Ern and Guermond.

Proof (dimension truncation). Let us introduce the operators $A(\boldsymbol{y}), A^s(\boldsymbol{y})\colon H_0^1(D) \to H^{-1}(D)$,

$$A(\boldsymbol{y}) := B_0 + \sum_{j=1}^{\infty} y_j B_j \quad \text{and} \quad A^s(\boldsymbol{y}) := B_0 + \sum_{j=1}^{s} y_j B_j,$$

where $B_j\colon H_0^1(D) \to H^{-1}(D)$ are defined by setting

$$\langle B_0 v, w\rangle_{H^{-1}(D),H_0^1(D)} := \langle a_0 \nabla v, \nabla w\rangle_{L^2(D)},$$

$$\langle B_j v, w\rangle_{H^{-1}(D),H_0^1(D)} := \langle \psi_j \nabla v, \nabla w\rangle_{L^2(D)} \quad \text{for } j \geq 1.$$

The variational problem

$$\int_D a(\boldsymbol{x},\boldsymbol{y}) \nabla u(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \langle F, v\rangle_{H^{-1}(D),H_0^1(D)} \quad \text{for all } v \in H_0^1(D),$$

$$a(\boldsymbol{x},\boldsymbol{y}) = a_0(\boldsymbol{x}) + \sum_{j=1}^{\infty} y_j \psi_j(\boldsymbol{x}),$$

where $F \in H^{-1}(D)$, can be expressed as an affine-parametric parametric operator equation

$$A(\boldsymbol{y}) u(\cdot,\boldsymbol{y}) = F.$$

Our assumptions (A1)–(A3) ensure that both $A(\boldsymbol{y})$ and $A^s(\boldsymbol{y})$ are boundedly invertible linear maps for all $\boldsymbol{y} \in U$.

Suppose that $1 \leq s < s'$. As a consequence of the *a priori* bound for the PDE, we have

$$\int_D G(u(\boldsymbol{y}) - u_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} \leq \frac{2\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}}$$
$$= \frac{2\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \frac{s^{-2/p+1}}{s^{-2/p+1}} \leq \frac{2\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \frac{s^{-2/p+1}}{(s')^{-2/p+1}}.$$

Thus it is sufficient to prove the claim for $s \geq s'$ with s' large enough. To this end, we assume that $s \geq s'$ where s' is chosen to be large enough such that

$$\sum_{j=s+1}^{\infty} b_j \leq \frac{1}{2} \quad \text{for all } s \geq s'. \tag{3}$$

For future reference, note that (3) also implies for all $s \geq s'$ that

$$b_j \leq \frac{1}{2} \quad \text{for all } j \geq s+1 \quad \text{and} \quad \sum_{j=s+1}^{\infty} b_j^2 \leq \sum_{j=s+1}^{\infty} b_j \leq \frac{1}{2}. \tag{4}$$

We also have for all $\boldsymbol{y} \in U$ that

$$\|A(\boldsymbol{y})\|_{H_0^1(D)\to H^{-1}(D)} \le a_{\max}, \quad \|A^s(\boldsymbol{y})\|_{H_0^1(D)\to H^{-1}(D)} \le a_{\max}$$
$$\|A(\boldsymbol{y})^{-1}\|_{H^{-1}(D)\to H_0^1(D)} \le \frac{1}{a_{\min}}, \quad \|A^s(\boldsymbol{y})^{-1}\|_{H^{-1}(D)\to H_0^1(D)} \le \frac{1}{a_{\min}}.$$

For brevity, let us denote

$$u(\boldsymbol{y}) := u(\cdot,\boldsymbol{y}), \quad \boldsymbol{y} \in U,$$
$$u_s(\boldsymbol{y}) := u_s(\cdot,\boldsymbol{y}), \quad \boldsymbol{y} \in U.$$

Now $u(\boldsymbol{y}) = A(\boldsymbol{y})^{-1}F$, $u_s(\boldsymbol{y}) = A^s(\boldsymbol{y})^{-1}F$, and we can write

$$A(\boldsymbol{y}) - A^s(\boldsymbol{y}) = \sum_{j=s+1}^{\infty} y_j B_j, \quad \boldsymbol{y} \in U,\ s \in \mathbb{N}.$$

Let $w \in H_0^1(D)$. Then

$$\|A^s(\boldsymbol{y})^{-1} B_j w\|_{H_0^1(D)} \leq \frac{\|B_j w\|_{H^{-1}(D)}}{a_{\min}}$$
$$= \frac{1}{a_{\min}} \sup_{v \in H_0^1(D) \setminus \{0\}} \frac{\langle \psi_j \nabla w, \nabla v \rangle_{L^2(D)}}{\|v\|_{H_0^1(D)}} \leq b_j \|w\|_{H_0^1(D)},$$

where the sequence $\boldsymbol{b} = (b_j)_{j \geq 1}$ is defined as $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$. In consequence,

$$\sup_{\boldsymbol{y} \in U} \|A^s(\boldsymbol{y})^{-1} B_j\|_{\mathcal{L}(H_0^1(D))} \leq b_j,$$

$$\sup_{\boldsymbol{y} \in U} \|A^s(\boldsymbol{y})^{-1}(A(\boldsymbol{y}) - A^s(\boldsymbol{y}))\|_{\mathcal{L}(H_0^1(D))} \leq \sum_{j=s+1}^{\infty} b_j \overset{(3)}{\leq} \frac{1}{2} < 1.$$

It follows from the previous discussion and the assumption $s \geq s'$ that the Neumann series

$$\begin{aligned} A(\boldsymbol{y})^{-1} &= (I + A^s(\boldsymbol{y})^{-1}(A(\boldsymbol{y}) - A^s(\boldsymbol{y})))^{-1} A^s(\boldsymbol{y})^{-1} \\ &= \sum_{k=0}^{\infty} (-A^s(\boldsymbol{y})^{-1}(A(\boldsymbol{y}) - A^s(\boldsymbol{y})))^k A^s(\boldsymbol{y})^{-1} \end{aligned}$$

is well-defined. Moreover, we have the representation

$$\begin{aligned} &\int_U G(u(\boldsymbol{y}) - u_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} = \int_U G((A(\boldsymbol{y})^{-1} - A^s(\boldsymbol{y})^{-1})f)\,\mathrm{d}\boldsymbol{y} \\ &= \sum_{k=1}^{\infty} \int_U G((-A^s(\boldsymbol{y})^{-1}(A(\boldsymbol{y}) - A^s(\boldsymbol{y})))^k u_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} \\ &= \sum_{k=1}^{\infty} (-1)^k \int_U G\bigg(\bigg(\sum_{j=s+1}^{\infty} y_j A^s(\boldsymbol{y})^{-1} B_j\bigg)^k u_s(\boldsymbol{y})\bigg)\,\mathrm{d}\boldsymbol{y}. \end{aligned}$$

The integrand can be expanded as

$$\left(\sum_{j=s+1}^{\infty} y_j A^s(\boldsymbol{y})^{-1} B_j\right)^k = \sum_{\eta_1,\ldots,\eta_k=s+1}^{\infty} \left(\prod_{i=1}^{k} y_{\eta_i}\right)\left(\prod_{i=1}^{k} A^s(\boldsymbol{y})^{-1} B_{\eta_i}\right),$$

where the second product symbol is assumed to respect the order of the noncommutative operators. By Fubini's theorem, we obtain

$$\int_U G\left(\left(\sum_{j=s+1}^{\infty} y_j A^s(\boldsymbol{y})^{-1} B_j\right)^k u_s(\boldsymbol{y})\right) \mathrm{d}\boldsymbol{y}$$

$$= \sum_{\eta_1,\ldots,\eta_k=s+1}^{\infty} \underbrace{\left(\int_U \prod_{i=1}^{k} y_{\eta_i}\,\mathrm{d}\boldsymbol{y}\right)}_{=:I_1}\underbrace{\left(\int_{U_s} G\left(\left(\prod_{i=1}^{k} A^s(\boldsymbol{y})^{-1} B_{\eta_i}\right) u_s(\boldsymbol{y})\right) \mathrm{d}\boldsymbol{y}_{\{1:s\}}\right)}_{=:I_2}.$$

- $I_1 \geq 0$ can be written as a product of univariate integrals of the form $0 \leq \int_{-1/2}^{1/2} y_j^m \,\mathrm{d}y_j \leq 1$, $m \in \mathbb{N}$. Note that this vanishes when $m = 1$.
- $|I_2| \leq \|G\|_{H^{-1}(D)} \big(\prod_{i=1}^{k} \sup_{\boldsymbol{y}\in U} \|A^s(\boldsymbol{y})^{-1} B_{\eta_i}\|\big) \|u_s(\boldsymbol{y})\|_{H_0^1(D)}$
  $\leq \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \big(\prod_{i=1}^{k} b_{\eta_i}\big).$

Earlier we arrived at

$$\int_U G(u(\boldsymbol{y}) - u_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} = \sum_{k=1}^{\infty}(-1)^k \int_U G\left(\left(\sum_{j=s+1}^{\infty} y_j A^s(\boldsymbol{y})^{-1} B_j\right)^k u_s(\boldsymbol{y})\right)\mathrm{d}\boldsymbol{y}$$

We can estimate the summands as

$$\left|(-1)^k \int_U G\left(\left(\sum_{j=s+1}^{\infty} y_j A^s(\boldsymbol{y})^{-1} B_j\right)^k u_s(\boldsymbol{y})\right)\mathrm{d}\boldsymbol{y}\right|$$

$$\leq \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \sum_{\eta_1,\ldots,\eta_k=s+1}^{\infty} \left(\int_U \prod_{k=1}^{k} y_{\eta_i}\,\mathrm{d}\boldsymbol{y}\right)\left(\prod_{i=1}^{k} b_{\eta_i}\right)$$

$$= \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \int_U \sum_{\eta_1,\ldots,\eta_k=s+1}^{\infty} \left(\prod_{k=1}^{k} y_{\eta_i}\right)\left(\prod_{i=1}^{k} b_{\eta_i}\right)\mathrm{d}\boldsymbol{y}$$

$$= \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \int_U \left(\sum_{j=s+1}^{\infty} y_j b_j\right)^k \mathrm{d}\boldsymbol{y}$$

$$\overset{(1)}{=} \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \int_U \sum_{\substack{|\boldsymbol{\nu}|=k \\ \nu_j=0\ \forall j\leq s}} \frac{k!}{\boldsymbol{\nu}!}\left(\prod_{j=s+1}^{\infty} y_j^{\nu_j}\right)\left(\prod_{j=s+1}^{\infty} b_j^{\nu_j}\right)\mathrm{d}\boldsymbol{y}.$$

The integrals vanish whenever $\boldsymbol{\nu}$ contains an element equal to 1, hence

$$\left|(-1)^k \int_U G\Big(\Big(\sum_{j=s+1}^{\infty} y_j A^s(\boldsymbol{y})^{-1} B_j\Big)^k u_s(\boldsymbol{y})\Big)\,\mathrm{d}\boldsymbol{y}\right| \leq \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \sum_{\substack{|\boldsymbol{\nu}|=k\\ \nu_j=0\ \forall j\leq s\\ \nu_j\neq 1\ \forall j>s}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{b}^{\boldsymbol{\nu}}.$$

We arrive at (note that the summand corresponding to $k=1$ vanishes!)

$$\left|\int_U G(u(\boldsymbol{y})-u_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}\right| \leq \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \sum_{k=1}^{\infty} \sum_{\substack{|\boldsymbol{\nu}|=k\\ \nu_j=0\ \forall j\leq s\\ \nu_j\neq 1\ \forall j>s}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{b}^{\boldsymbol{\nu}}$$

$$= \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}} \Bigg[\sum_{k=k'}^{\infty} \sum_{\substack{|\boldsymbol{\nu}|=k\\ \nu_j=0\ \forall j\leq s\\ \nu_j\neq 1\ \forall j>s}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{b}^{\boldsymbol{\nu}} + \sum_{k=2}^{k'-1} \sum_{\substack{|\boldsymbol{\nu}|=k\\ \nu_j=0\ \forall j\leq s\\ \nu_j\neq 1\ \forall j>s}} \frac{k!}{\boldsymbol{\nu}!}\boldsymbol{b}^{\boldsymbol{\nu}}\Bigg],$$

where we split the sum into two w.r.t. $k'\geq 3$ to be specified later.

The sum over $k \geq k'$ can be bounded using the geometric series as

$$\sum_{k=k'}^{\infty} \sum_{\substack{|\boldsymbol{\nu}|=k \\ \nu_j=0\ \forall j \leq s \\ \nu_j \neq 1\ \forall j>s}} \frac{k!}{\boldsymbol{\nu}!} \boldsymbol{b}^{\boldsymbol{\nu}} \leq \sum_{k=k'}^{\infty} \Big(\sum_{j=s+1}^{\infty} b_j \Big)^k$$

$$\leq \Big(\sum_{j=s+1}^{\infty} b_j \Big)^{k'} \frac{1}{1 - \sum_{j=s+1}^{\infty} b_j} \leq C s^{k'(-1/p+1)},$$

where Stechkin's lemma yields $\sum_{j=s+1}^{\infty} b_j \leq \big(\sum_{j=1}^{\infty} b_j^p \big)^{1/p} s^{-1/p+1}$ and the resulting constant $C_1 := 2\big(\sum_{j=1}^{\infty} b_j^p \big)^{k'/p}$ is independent of s, f, and G.

On the other hand, for the sum over $2 \leq k < k'$, we estimate

$$\sum_{k=2}^{k'-1} \sum_{\substack{|\boldsymbol{\nu}|=k \\ \nu_j=0\ \forall j\leq s \\ \nu_j\neq 1\ \forall j>s}} \frac{k!}{\boldsymbol{\nu}!} \boldsymbol{b}^{\boldsymbol{\nu}} \leq (k'-1)! \sum_{k=2}^{k'-1} \sum_{\substack{|\boldsymbol{\nu}|=k \\ \nu_j=0\ \forall j\leq s \\ \nu_j\neq 1\ \forall j>s}} \boldsymbol{b}^{\boldsymbol{\nu}}.$$

For each $2 \leq k < k'$, we obtain

$$\sum_{\substack{|\boldsymbol{\nu}|=k \\ \nu_j=0\ \forall j\leq s \\ \nu_j\neq 1\ \forall j>s}} \boldsymbol{b}^{\boldsymbol{\nu}} \leq \sum_{\substack{0\neq|\boldsymbol{\nu}|_\infty\leq k \\ \nu_j=0\ \forall j\leq s \\ \nu_j\neq 1\ \forall j>s}} \boldsymbol{b}^{\boldsymbol{\nu}} = \prod_{j=s+1}^{\infty} \left(1 + \sum_{\ell=2}^{k} b_j^{\ell}\right) - 1$$

$$= \prod_{j=s+1}^{\infty} \left(1 + b_j^2 \frac{1-b_j^{j-1}}{1-b_j}\right) - 1 \leq \prod_{j=s+1}^{\infty} (1 + 2b_j^2) - 1$$

$$\leq \exp\left(2 \sum_{j=s+1}^{\infty} b_j^2\right) - 1 \leq C_2 s^{-2/p+1}, \quad C_2 := 2(\mathrm{e}-1)\Big(\sum_{j=1}^{\infty} b_j^p\Big)^{1/p},$$

where we used $\mathrm{e}^x \leq 1 + (\mathrm{e}-1)x$ for $x \in [0,1]$ and Stechkin's lemma $\sum_{j=s+1}^{\infty} b_j^2 \leq \big(\sum_{j=1}^{\infty} b_j^p\big)^{1/p} s^{-2/p+1}$. C_2 is independent of s, f, and G.

Putting everything together, we conclude that

$$\left| \int_U G(u(\boldsymbol{y}) - u_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} \right| \leq \frac{\|G\|_{H^{-1}(D)}\|F\|_{H^{-1}(D)}}{a_{\min}}\big(C_1 s^{k'(-1/p+1)} + k'!(k'-2)C_2 s^{-2/p+1}\big).$$

The two terms can be balanced by choosing $k' := \lceil (2-p)/(1-p) \rceil$, where $\lceil x \rceil := \min\{k \in \mathbb{Z} \mid k \geq x\}$ is the ceiling function. (Note that $k' \geq 3$ for all $p \in (0,1)$.)

Since we already know that the result holds for all $s \leq s'$, the assertion for all $s \geq 1$ follows by a trivial adjustment of the constant factors.

Finally, if the source term $f \in L^2(D)$, we can associate it with an element $F \in H^{-1}(D)$ defined by

$$\langle F, v \rangle_{H^{-1}(D), H_0^1(D)} := \int_D f(\boldsymbol{x}) v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad v \in H_0^1(D).$$

Especially, $\|F\|_{H^{-1}(D)} \leq C_P \|f\|_{L^2(D)}$, where $C_P > 0$ is the Poincaré constant. □

Finite element error

Suppose that $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, is a bounded, convex polyhedral domain.

Let $\{V_h\}_h$ be a family of finite element subspaces of $H_0^1(D)$, indexed by the mesh size $h > 0$ and spanned by continuous, piecewise linear finite element basis functions over a sequence of regular, simplicial meshes in D obtained from an initial, regular triangulation of D by recursive, uniform bisection of simplices.

In this setup, it is known (cf., e.g., Gilbarg and Trudinger) that for functions $v \in H_0^1(D) \cap H^2(D)$, there exists a constant $C_1 > 0$ such that

$$\inf_{v_h \in V_h} \|v - v_h\|_{H_0^1(D)} \leq C_1 h \|v\|_{H_0^1(D) \cap H^2(D)} \quad \text{as } h \to 0, \tag{5}$$

where $\|v\|_{H_0^1(D)\cap H^2(D)} := (\|v\|_{L^2(D)}^2 + \|\Delta v\|_{L^2(D)}^2)^{1/2}$.

Note that we need higher $H^2(D)$ regularity of the PDE solution in order to derive the asymptotic convergence rate as $h \to \infty$. This can be ensured, e.g., when the diffusion coefficient is Lipschitz, $f \in L^2(D)$, and the domain D is a bounded, convex polyhedron.

Proposition (Elliptic regularity)

Suppose that $a_0 \in W^{1,\infty}(D)$ *and* $\psi_j \in W^{1,\infty}(D)$ *for all* $j \geq 1$ *such that* $C_\psi := \sum_{j\geq 1} \|\psi_j\|_{W^{1,\infty}(D)} < \infty$, *where*

$$\|v\|_{W^{1,\infty}(D)} := \max\{\|v\|_{L^\infty(D)}, \|\nabla v\|_{L^\infty(D)}\}.$$

Then there exists a constant $C_2 > 0$ *independent of* $\boldsymbol{y}$ *and* f *such that the solution* $u(\cdot,\boldsymbol{y}) \in H_0^1(D)$ *of the parametric PDE problem satisfies*

$$\|u(\cdot,\boldsymbol{y})\|_{H_0^1(D)\cap H^2(D)} \leq C_2\|f\|_{L^2(D)} \quad \text{for all } \boldsymbol{y} \in U. \tag{6}$$

Proof (sketch). Standard ellipticity theory implies that $u(\cdot,\boldsymbol{y}) \in H_0^1(D)$ is such that $\exists \Delta u(\cdot,\boldsymbol{y}) \in L^2(D)$ for all $\boldsymbol{y} \in U$. Since now $\|a(\cdot,\boldsymbol{y})\|_{W^{1,\infty}(D)} < \infty$ for all $\boldsymbol{y} \in U$, we obtain

$$\begin{aligned}
&-\nabla\cdot(a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})) = f(\boldsymbol{x}) \qquad\qquad (\nabla\cdot(\psi\nabla\varphi) = \nabla\psi\cdot\nabla\varphi + \psi\Delta\varphi)\\
&\Leftrightarrow -a(\boldsymbol{x},\boldsymbol{y})\Delta u(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla a(\boldsymbol{x},\boldsymbol{y})\cdot\nabla u(\boldsymbol{x},\boldsymbol{y})\\
&\Rightarrow \|\Delta u(\cdot,\boldsymbol{y})\|_{L^2(D)} \leq \frac{\|f\|_{L^2(D)}}{a_{\min}} + \frac{\|\nabla a(\cdot,\boldsymbol{y})\|_{L^\infty(D)}}{a_{\min}}\|u(\cdot,\boldsymbol{y})\|_{H_0^1(D)}\\
&\quad \leq \frac{\|f\|_{L^2(D)}}{a_{\min}} + \frac{\|a_0\|_{W^{1,\infty}(D)} + C_\psi}{a_{\min}}\,\frac{C_P\|f\|_{L^2(D)}}{a_{\min}} =: C_2\|f\|_{L^2(D)}. \qquad \square
\end{aligned}$$

Dimensionally-truncated finite element solution

Let $a_s(\boldsymbol{x},\boldsymbol{y}) := a(\boldsymbol{x},(y_1,\ldots,y_s,0,0,\ldots))$ for $\boldsymbol{y} \in U$. For $\boldsymbol{y} \in U$, $u_{s,h}(\cdot,\boldsymbol{y}) \in V_h$ is the dimensionally-truncated finite element solution if

$$\int_D a_s(\boldsymbol{x},\boldsymbol{y}) \nabla u_{s,h}(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \text{for all } v \in V_h.$$

Finite element error in $H_0^1(D)$

Recall that by Céa's lemma, the finite element solution is a *quasi-optimal* approximation in the following sense:

$$\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H_0^1(D)} \leq C(\boldsymbol{y}) \inf_{v_h \in V_h} \|u_s(\cdot,\boldsymbol{y}) - v_h\|_{H_0^1(D)},$$

where the constant $C(\boldsymbol{y}) := \frac{\sup_{\boldsymbol{x}\in D} a(\boldsymbol{x},\boldsymbol{y})}{\inf_{\boldsymbol{x}\in D} a(\boldsymbol{x},\boldsymbol{y})} \leq \frac{a_{\max}}{a_{\min}} =: C_3 < \infty$ can be bounded independently of $\boldsymbol{y} \in U$ due to our uniform ellipticity assumption. Combining this with the approximation property (5) and the elliptic regularity shift (6) yields

$$\begin{aligned} \|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H_0^1(D)} &\leq C_3 \inf_{v_h \in V_h} \|u_s(\cdot,\boldsymbol{y}) - v_h\|_{H_0^1(D)} \\ &\overset{(5)}{\leq} C_3 C_1 h \|u_s(\cdot,\boldsymbol{y})\|_{H^2(D)\cap H_0^1(D)} \\ &\overset{(6)}{\leq} C_3 C_1 C_2 h \|f\|_{L^2(D)} \quad \text{as } h \to 0. \end{aligned} \tag{7}$$

However, if we measure the error in the $L^2(D)$ norm, the finite element convergence rate can be improved by an order of magnitude.

Finite element error in $L^2(D)$

Proposition

Under the same assumptions as the previous proposition, there exists a constant $C > 0$ independent of s, h, f, and $\boldsymbol{y}$ such that

$$\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{L^2(D)} \leq Ch^2\|f\|_{L^2(D)} \quad \text{as } h \to 0.$$

Proof. Let $g \in L^2(D)$. For $\boldsymbol{y} \in U$, let $u_{g,s}(\cdot,\boldsymbol{y}) \in H_0^1(D)$ denote the solution to

$$\int_D a_s(\boldsymbol{x},\boldsymbol{y})\nabla u_{g,s}(\cdot,\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_D g(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{for all } v \in H_0^1(D),$$

where $a_s(\cdot,\boldsymbol{y}) := a(\cdot,(y_1,\ldots,y_s,0,0,\ldots))$. We test this against $v = u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})$ and let $v_h \in V_h$ be arbitrary.

It follows from Galerkin orthogonality of the finite element solution that

$$\begin{aligned}
&\langle g, u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\rangle_{L^2(D)} \\
&= \int_D a_s(\boldsymbol{x},\boldsymbol{y})\nabla u_{g,s}(\boldsymbol{x},\boldsymbol{y})\cdot\nabla(u_s(\boldsymbol{x},\boldsymbol{y}) - u_{s,h}(\boldsymbol{x},\boldsymbol{y}))\,\mathrm{d}\boldsymbol{x} \\
&= \int_D a_s(\boldsymbol{x},\boldsymbol{y})\nabla(u_{g,s}(\boldsymbol{x},\boldsymbol{y}) - v_h(\boldsymbol{x}))\cdot\nabla(u_s(\boldsymbol{x},\boldsymbol{y}) - u_{s,h}(\boldsymbol{x},\boldsymbol{y}))\,\mathrm{d}\boldsymbol{x} \\
&\leq a_{\max}\|u_{g,s}(\cdot,\boldsymbol{y}) - v_h\|_{H_0^1(D)}\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H_0^1(D)}.
\end{aligned}$$

In consequence,

$$\begin{aligned}
&\langle g, u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\rangle_{L^2(D)} \\
&\leq a_{\max}\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H_0^1(D)} \inf_{v_h\in V_h}\|u_{g,s}(\cdot,\boldsymbol{y}) - v_h\|_{H_0^1(D)},
\end{aligned} \tag{8}$$

where $g \in L^2(D)$ is arbitrary. We now use the *Aubin–Nitsche trick*: recall from the exercises of week 2(!) that the following identity holds

$$\|F\|_{L^2(D)} = \sup_{\substack{g\in L^2(D)\\ \|g\|_{L^2(D)}\leq 1}} \langle g, F\rangle_{L^2(D)} \quad \text{for all } F \in L^2(D).$$

We take the supremum over $\{g \in L^2(D) : \|g\|_{L^2(D)} \leq 1\}$ in (8) to obtain...

$$
\begin{aligned}
&\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{L^2(D)} \\
&= \sup_{\substack{g\in L^2(D)\\ \|g\|_{L^2(D)}\leq 1}} \langle g, u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\rangle_{L^2(D)} \\
&\leq a_{\max} \underbrace{\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H_0^1(D)}}_{\overset{(7)}{\leq} C_3 C_1 C_2 h \|f\|_{L^2(D)}} \sup_{\substack{g\in L^2(D)\\ \|g\|_{L^2(D)}\leq 1}} \underbrace{\Big(\inf_{v_h\in V_h} \|u_{g,s}(\cdot,\boldsymbol{y}) - v_h\|_{H_0^1(D)}\Big)}_{\substack{\overset{(5)}{\leq} C_1 h \|u_{g,s}(\cdot,\boldsymbol{y})\|_{H_0^1(D)\cap H^2(D)} \\ \overset{(6)}{\leq} C_1 C_2 h \|g\|_{L^2(D)}}} \\
&\leq Ch^2 \|f\|_{L^2(D)},
\end{aligned}
$$

where the constant $C := a_{\max}(C_1 C_2)^2 C_3$ is independent of s, h, f, and $\boldsymbol{y}$. □

Note especially that if $G\colon L^2(D) \to \mathbb{R}$ is a bounded linear operator, then

$$
\int_U |G(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}))|\,\mathrm{d}\boldsymbol{y} \leq C\|G\|_{L^2(D)\to\mathbb{R}} \|f\|_{L^2(D)} h^2,
$$

where $C > 0$ is independent of s, h, and f.

Overall error

Let $I(F) := \int_U F(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$.

Theorem

Let $D \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded polyhedron, assume (A1)–(A3), $\|\psi_1\|_{L^\infty(D)} \geq \|\psi_2\|_{L^\infty(D)} \geq \|\psi_3\|_{L^\infty(D)} \geq \cdots$, and suppose that $a_0 \in W^{1,\infty}(D)$ and $\psi_j \in W^{1,\infty}(D)$ with $\sum_{j=1}^\infty \|\psi_j\|_{W^{1,\infty}(D)} < \infty$. Let $G\colon L^2(D) \to \mathbb{R}$ be a bounded linear functional and define $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$. Then using the CBC algorithm with the POD weights

$$\gamma_{\mathfrak{u}} := \Big(|\mathfrak{u}|! \prod_{j\in\mathfrak{u}} \frac{b_j}{\sqrt{2\zeta(2\lambda)/(2\pi^2)^\lambda}}\Big)^{2/(1+\lambda)}, \quad \lambda := \begin{cases} \frac{p}{2-p} & \text{if } p \in (2/3,1), \\ \frac{1}{2-2\delta} & \text{if } p \in (0,2/3], \end{cases}$$

as inputs to construct a randomly shifted rank-1 lattice rule $Q_{n,s}^{\boldsymbol{\Delta}}(F) := \sum_{k=0}^{n-1} F(\{\frac{k\boldsymbol{z}}{n} + \boldsymbol{\Delta}\} - \frac{\boldsymbol{1}}{\boldsymbol{2}})$, $\boldsymbol{\Delta} \in [0,1]^s$, we have the overall error

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}}|I(G(u)) - Q_{n,s}^{\boldsymbol{\Delta}}(G(u_{s,h}))|^2} \leq C(\varphi(n)^{\max\{-1/p+1/2,-1+\delta\}} + s^{-2/p+1} + h^2),$$

where the constant $C > 0$ is independent of s, n, and h.

Proof. We have the total error decomposition[†]

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}}[|I(G(u)) - Q_{n,s}^{\boldsymbol{\Delta}}(u_{s,h})|^2] \leq\ & 9|(I - I_s)(G(u))|^2 \\ &+ 9|I_s(G(u_s - u_{s,h}))|^2 \\ &+ 9\mathbb{E}_{\boldsymbol{\Delta}}[|I_s(G(u_{s,h})) - Q_{n,s}^{\boldsymbol{\Delta}}(G(u_{s,h}))|^2].\end{aligned}$$

We have already proved, under the stated assumptions, that there hold

$$\begin{aligned}&|(I - I_s)(G(u))| = \mathcal{O}(s^{-2/p+1}),\\ &|I_s(G(u_s - u_{s,h}))| = \mathcal{O}(h^2),\\ &\mathbb{E}_{\boldsymbol{\Delta}}[|I_s(G(u_{s,h})) - Q_{n,s}^{\boldsymbol{\Delta}}(G(u_{s,h}))|^2] = \mathcal{O}(n^{\max\{-1/p+1/2,-1+\delta\}}),\end{aligned}$$

from which the claim immediately follows. □

[†]Let $a, b, c \geq 0$. Then
$a + b + c \leq 3\max\{a, b, c\} = 3\sqrt{\max\{a, b, c\}^2} \leq 3\sqrt{a^2 + b^2 + c^2}$.

Extension of QMC theory to the full PDE solution *without* a bounded linear quantity of interest G

Earlier, we discussed the QMC approximation for integrals of the form

$$\mathbb{E}[G(u_s)] = \int_{U_s} G(u_s(\cdot, \boldsymbol{y}))\,\mathrm{d}\boldsymbol{y},$$

where $G\colon H_0^1(D) \to \mathbb{R}$ (or $G\colon L^2(D) \to \mathbb{R}$) is a bounded linear functional (quantity of interest).

But what if we wanted to approximate

$$\mathbb{E}[u_s(\boldsymbol{x}, \cdot)] = \int_{U_s} u_s(\boldsymbol{x}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$

without a linear quantity of interest instead?

Idea: recall the variational characterization

$$\|f\|_{L^2(D)} = \sup_{\substack{G \in L^2(D) \\ \|G\|_{L^2(D)} \leq 1}} \langle G, f \rangle_{L^2(D)}$$

of the L^2 norm from earlier.

By Fubini's theorem, we have that

$$\begin{aligned}
\|I_s(u_s) - Q^{\boldsymbol{\Delta}}_{n,s}(u_s)\|_{L^2(D)} &= \sup_{\substack{G\in L^2(D)\\ \|G\|_{L^2(D)}\leq 1}} |\langle G, I_s(u_s) - Q^{\boldsymbol{\Delta}}_{n,s}(u_s)\rangle_{L^2(D)}| \\
&= \sup_{\substack{G\in L^2(D)\\ \|G\|_{L^2(D)}\leq 1}} |I_s(\langle G, u_s\rangle_{L^2(D)}) - Q^{\boldsymbol{\Delta}}_{n,s}(\langle G, u_s\rangle_{L^2(D)})| \\
&\leq e_{n,s}(\boldsymbol{z};\boldsymbol{\Delta}) \sup_{\substack{G\in L^2(D)\\ \|G\|_{L^2(D)}\leq 1}} \|\langle G, u_s\rangle_{L^2(D)}\|_{s,\boldsymbol{\gamma}},
\end{aligned}$$

where $e_{n,s}(\boldsymbol{z};\boldsymbol{\Delta})$ denotes the worst-case error of the shifted lattice $\{\boldsymbol{t}_i + \boldsymbol{\Delta} : i \in \{1,\ldots,n\}\}$. Especially:

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}}\|I_s(u_s) - Q^{\boldsymbol{\Delta}}_{n,s}(u_s)\|^2_{L^2(D)}} \leq e^{\rm sh}_{n,s}(\boldsymbol{z}) \sup_{\substack{G\in L^2(D)\\ \|G\|_{L^2(D)}\leq 1}} \|\langle G, u_s\rangle_{L^2(D)}\|_{s,\boldsymbol{\gamma}}.$$

The *shift-averaged worst-case error* $e^{\rm sh}_{n,s}(\boldsymbol{z})$ is precisely the same object that we have considered in the past, i.e.,
$[e^{\rm sh}_{n,s}(\boldsymbol{z})]^2 = \frac{1}{n}\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s\}}\gamma_{\mathfrak{u}}\sum_{k=0}^{n-1}\prod_{j\in\mathfrak{u}} B_2\big(\{\frac{kz_j}{n}\}\big)$.

In summary, even in this setting, we have the CBC search criterion

$$[e_{n,s}^{\mathrm{sh}}(\boldsymbol{z})]^2 = \frac{1}{n} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \sum_{k=0}^{n-1} \prod_{j \in \mathfrak{u}} B_2\left(\left\{ \frac{k z_j}{n} \right\}\right).$$

The generating vector obtained using the CBC algorithm satisfies the estimate

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} \| I_s(u_s) - Q_{n,s}^{\boldsymbol{\Delta}}(u_s) \|_{L^2(D)}^2} \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/\lambda} \times \sup_{\substack{G \in L^2(D) \\ \|G\|_{L^2(D)} \leq 1}} \| \langle G, u_s \rangle_{L^2(D)} \|_{s,\boldsymbol{\gamma}}$$

for all $\lambda \in (1/2, 1]$.

Precisely the same analysis that we carried out before shows that choosing the weights

$$\gamma_{\mathfrak{u}} := \left(|\mathfrak{u}|! \prod_{j\in\mathfrak{u}} \frac{b_j}{\sqrt{2\zeta(2\lambda)/(2\pi^2)^\lambda}}\right)^{2/(1+\lambda)}, \quad \lambda := \begin{cases} \frac{p}{2-p} & \text{if } p \in (2/3, 1), \\ \frac{1}{2-2\delta} & \text{if } p \in (0, 2/3], \end{cases}$$

with arbitrary $\delta > 0$, yields the QMC convergence rate

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}}\|I_s(u_s) - Q_{n,s}^{\boldsymbol{\Delta}}(u_s)\|_{L^2(D)}^2} = \mathcal{O}(\varphi(n)^{\max\{-1/p+1/2,-1+\delta\}}),$$

where the implied coefficient is independent of the dimension s.

Naturally, the dimensionally-truncated PDE solution in the above formula can be replaced by the dimensionally-truncated FE solution $u_{s,h}$ (provided that we use a conforming FE method, i.e., the domain D is a polygon and we use, e.g., piecewise linear finite element basis functions to span the finite element space V_h).