

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

First lecture, April 15, 2025
Second lecture, April 22, 2025
Preliminaries

Practical matters

- Lectures on Mondays at 10:15-11:45 in A6/032 by Vesa Kaarnioja.
- Exercises on Tuesdays at 10:15-11:45 in A6/032 by Vesa Kaarnioja.
- *There will be no lectures on April 14, April 21, and June 9.*
- *The first and second lecture will be held April 15 and April 22 in room A3/120 in place of the exercise session.*
- Exercise sheets will be published regularly on the course Whiteboard page. Please submit your solutions to the exercises before the deadlines specified on each exercise sheet.
- The conditions for completing this course are
 - (1) *successfully earning a cumulative 60% of points from the exercises* (active participation + regular attendance), and
 - (2) *successfully passing the course exam.*

The course evaluation is based on the oral exam at the end of the course.

Exercise guidelines

- Solutions to exercises can be submitted either via email or by handing in your solutions at the exercise session by the specified deadline. Late submissions will *not* be considered.
- Please present your calculations clearly and neatly, providing explanation for all steps.
- Ensure that your arguments are coherent and presented in an orderly fashion. Organize your solutions logically, starting from the problem statement and proceeding step-by-step to the solution.
- Typeset or write your solutions in clear handwriting for easy readability.
- Avoid ambiguity in your solutions: consider the perspective or the reader and ensure that your solutions are understandable from their point of view (i.e., the reader should not have to guess what you have written).
- Use appropriate mathematical notation and terminology.
- Double-check your solutions for errors and correctness before submission. Aim for precision and accuracy in your mathematical expressions and calculations.
- In programming tasks, ensure that your program executes successfully. Include the source code as well as the output of the program as part of your submission.

Uncertainty in groundwater flow

Risk analysis of radwaste disposal or CO₂ sequestration.

Darcy's law

$$\mathbf{q}(\mathbf{x}) + \mathbf{a}(\mathbf{x}) \nabla p(\mathbf{x}) = \mathbf{f}(\mathbf{x})$$

mass conservation law

$$\nabla \cdot \mathbf{q}(\mathbf{x}) = 0$$

in $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$

together with boundary conditions

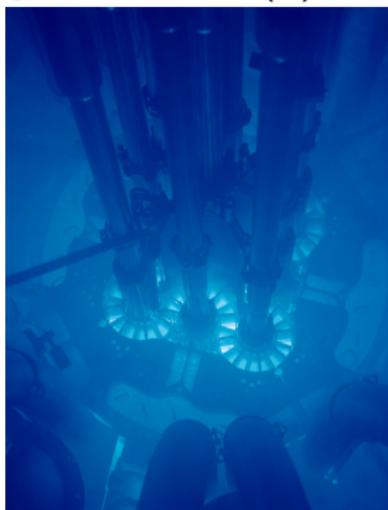


Uncertainty in $\mathbf{a}(\mathbf{x}, \omega)$ leads to uncertainty in $\mathbf{q}(\mathbf{x}, \omega)$ and $p(\mathbf{x}, \omega)$

Criticality problem for nuclear reactors

$$-\nabla \cdot (\underbrace{a(\mathbf{x})}_{\text{diffusion}} \nabla u(\mathbf{x})) + \underbrace{b(\mathbf{x})}_{\text{absorption}} u(\mathbf{x}) = \lambda \underbrace{c(\mathbf{x})}_{\text{fission}} u(\mathbf{x})$$

- The smallest eigenvalue $\lambda_1 \in \mathbb{R}$ measures *criticality* of a reactor.
- Eigenfunction $u_1(\mathbf{x})$ is the *neutron flux* at the point \mathbf{x} .



- $\lambda_1 \approx 1 \Rightarrow$ operating efficiently
- $\lambda_1 > 1 \Rightarrow$ not self-sustaining
- $\lambda_1 < 1 \Rightarrow$ supercritical

Source: Argonne National
Laboratory on Flickr

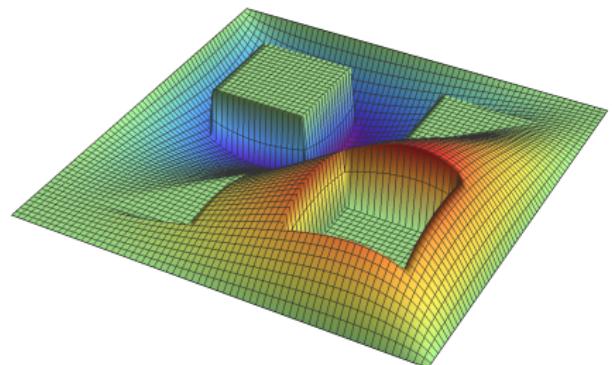
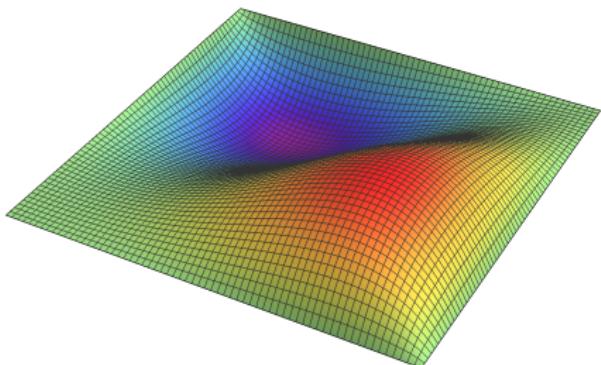
Optimization under uncertainty

Find $\min_{z \in L^2(D)} J(u, z)$,

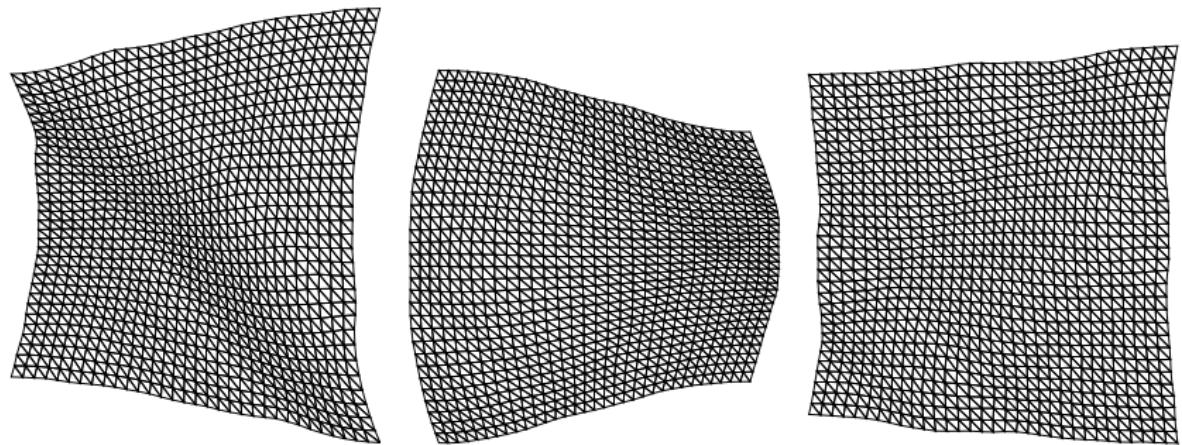
$$J(u, z) := \frac{1}{2} \int_{\Omega} \int_D (u(\mathbf{x}, \omega) - g(\mathbf{x}))^2 d\mathbf{x} d\mathbb{P}(\omega) + \frac{\alpha}{2} \int_D z(\mathbf{x})^2 d\mathbf{x},$$

subject to

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = z(\mathbf{x}), & \mathbf{x} \in D, \text{ a.e. } \omega \in \Omega \\ u(\mathbf{x}, \omega) = 0, & \mathbf{x} \in \partial D, \text{ a.e. } \omega \in \Omega, \\ z_{\min}(\mathbf{x}) \leq z(\mathbf{x}) \leq z_{\max}(\mathbf{x}), & \text{a.e. } \mathbf{x} \in D. \end{cases}$$



Domain uncertainty quantification

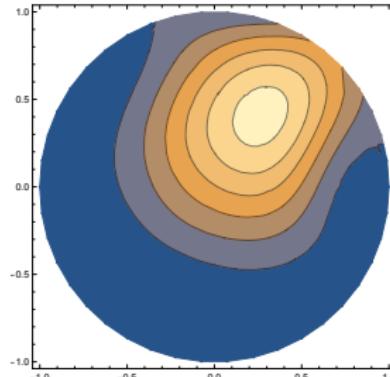
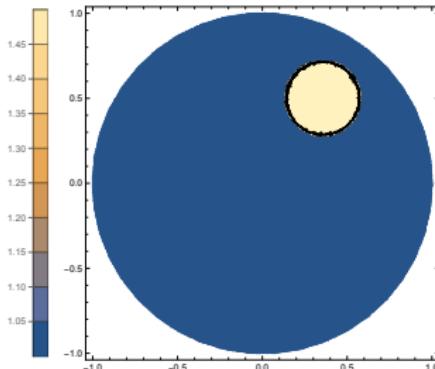
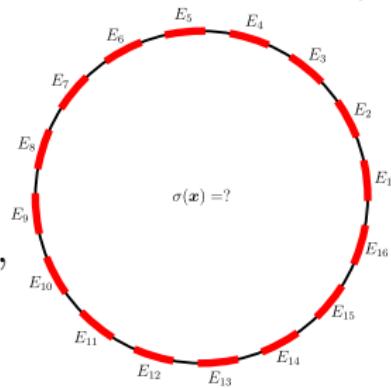


Three realizations of a random spatial domain

Electrical impedance tomography

Use measurements of current and voltage collected at electrodes covering part of the boundary to infer the interior conductivity of an object/body.

$$\begin{cases} \nabla \cdot (\sigma \nabla u) = 0 & \text{in } D, \\ \sigma \frac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \partial D \setminus \bigcup_{k=1}^L \overline{E_k}, \\ u + z_k \sigma \frac{\partial u}{\partial \mathbf{n}} = U_k & \text{on } E_k, \ k \in \{1, \dots, L\}, \\ \int_{E_k} \sigma \frac{\partial u}{\partial \mathbf{n}} dS = I_k, & k \in \{1, \dots, L\}, \end{cases}$$



Consider the elliptic PDE problem:

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) = f(\mathbf{x}) & \text{for } \mathbf{x} \in D, \\ +\text{boundary conditions.} \end{cases}$$

In practice, one or several of the material/system parameters may be uncertain or incompletely known and modeled as random fields:

- PDE coefficient a may be uncertain;
- Source term f may be uncertain;
- Boundary conditions may be uncertain;
- The domain D itself may be uncertain.

In forward uncertainty quantification, one is interested in assessing how uncertainties in the inputs of a mathematical model affect the output.

⇒ If the uncertain inputs are modeled as random fields, then the output of the PDE is also a random field. One may be interested in assessing the statistical response of the system, for example, the expectation or variance of the PDE solution (or some other quantity of interest thereof).

High-dimensional numerical integration

$$\int_{[0,1]^s} f(\mathbf{y}) \, d\mathbf{y} \approx \sum_{i=1}^n w_i f(\mathbf{t}_i)$$

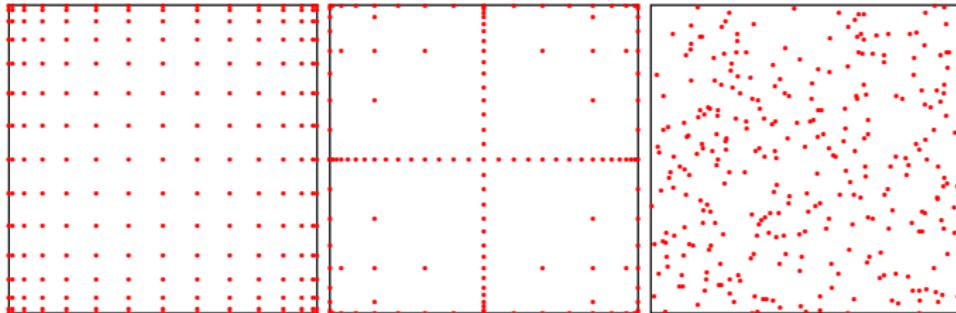


Figure: Tensor product grid, sparse grid, Monte Carlo nodes (not QMC rules)

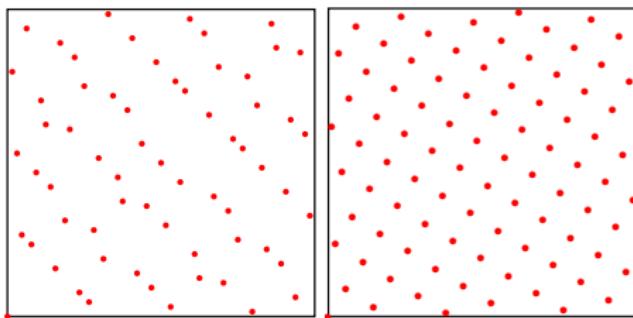


Figure: Sobol' points, lattice rule (examples of QMC rules)

Quasi-Monte Carlo (QMC) methods are a class of *equal weight* cubature rules

$$\int_{[0,1]^s} f(\mathbf{y}) d\mathbf{y} \approx \frac{1}{n} \sum_{i=1}^n f(\mathbf{t}_i),$$

where $(\mathbf{t}_i)_{i=1}^n$ is an ensemble of *deterministic* nodes in $[0, 1]^s$.

The nodes $(\mathbf{t}_i)_{i=1}^n$ are NOT random! Instead, they are *deterministically chosen*.

QMC methods exploit the smoothness and anisotropy of an integrand in order to achieve better-than-Monte Carlo rates.

Course contents

- Preliminaries: Hilbert spaces, Sobolev spaces, elliptic partial differential equations (PDEs)
- Finite element (FE) method
- Modeling random field inputs
- Elliptic PDEs with random coefficients
- Quasi-Monte Carlo (QMC) methods
- QMC-FE methods for uncertainty quantification of elliptic PDEs with random coefficients

Preliminary functional analysis

Inner product space

A real vector space X is an *inner product space* if there exists a mapping $\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{R}$ satisfying

- $\langle ax_1 + bx_2, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$ for all $x_1, x_2, y \in X$ and $a, b \in \mathbb{R}$;
- $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in X$;
- $\langle x, x \rangle \geq 0$ for all $x \in X$, where equality holds iff $x = 0$.

A mapping $\langle \cdot, \cdot \rangle$ satisfying these conditions is called an *inner product*.

Example

i) $\mathbb{R}^n = \{(x_1, \dots, x_n) \mid x_k \in \mathbb{R}\}$. Then the inner product is the Euclidean dot product

$$\langle x, y \rangle = \sum_{k=1}^n x_k y_k, \quad x = (x_1, \dots, x_n), \quad y = (y_1, \dots, y_n).$$

ii) Let $X = C([a, b]) = \{f \mid f : [a, b] \rightarrow \mathbb{R} \text{ is continuous}\}$ and define

$$\langle f, g \rangle = \int_a^b f(x)g(x) dx.$$

Then this is an inner product on $C([a, b])$.

iii) Let $X = \ell^2(\mathbb{R}) = \{(z_k)_{k=1}^{\infty} \mid \sum_{k=1}^{\infty} |z_k|^2 < \infty\}$. Then $\ell^2(\mathbb{R})$ is an inner product space when

$$\langle x, y \rangle = \sum_{k=1}^{\infty} x_k y_k, \quad x = (x_1, x_2, \dots), \quad y = (y_1, y_2, \dots).$$

Definition

A real vector space X is a *normed space* if there exists a mapping $\|\cdot\|: X \rightarrow \mathbb{R}$ satisfying

- $\|ax\| = |a|\|x\|$ for all $a \in \mathbb{R}$ and $x \in X$;
- $\|x\| \geq 0$ for all $x \in X$, where equality holds iff $x = 0$.
- $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (triangle inequality).

If X is an inner product space, then it is a normed space in a canonical way with the induced norm $\|\cdot\|: X \rightarrow \mathbb{R}$ defined by

$$\|x\| = \sqrt{\langle x, x \rangle}, \quad x \in X.$$

The first two postulates follow immediately from the properties of inner product spaces, the triangle inequality follows from the Cauchy–Schwarz inequality.

Proposition (Cauchy–Schwarz inequality)

If $(X, \langle \cdot, \cdot \rangle)$ is an inner product space, then

$$|\langle x, y \rangle| \leq \|x\| \|y\| \quad \text{for all } x, y \in X.$$

Proof. Let $x, y \in X$ and $t \in \mathbb{R}$. If $x = 0$ or $y = 0$, then the claim is trivial. Suppose that $x \neq 0 \neq y$. Then

$$0 \leq \langle x + ty, x + ty \rangle = \|x\|^2 + 2t\langle x, y \rangle + t^2\|y\|^2.$$

This is a second degree polynomial w.r.t. t with at most 1 real root. Hence,

$$\begin{aligned} \text{discriminant} \leq 0 &\Leftrightarrow 4|\langle x, y \rangle|^2 - 4\|x\|^2\|y\|^2 \leq 0 \\ &\Leftrightarrow |\langle x, y \rangle|^2 \leq \|x\|^2\|y\|^2. \end{aligned}$$

Note that if $y = ax$, $a \in \mathbb{R}$, then discriminant = 0 and Cauchy–Schwarz holds with equality. □

The triangle inequality is an immediate consequence of Cauchy–Schwarz:

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2|\langle x, y \rangle| \leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \\ &= (\|x\| + \|y\|)^2 \quad \text{for all } x, y \in X. \end{aligned}$$

For our purposes, having an inner product is not enough. We need to know that these spaces are also *complete* normed spaces.

Definition (Cauchy sequence)

A sequence $(x_k)_{k=1}^{\infty}$ of elements of $(X, \|\cdot\|)$ is called a *Cauchy sequence* if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|x_m - x_n\| < \varepsilon.$$

Definition (Complete space)

A normed space $(X, \|\cdot\|)$ is *complete* if all Cauchy sequences in X converge to an element of X .

Definition (Banach space)

A normed space $(X, \|\cdot\|)$ which is complete with respect to $\|\cdot\|$ is a *Banach space*.

Definition (Hilbert space)

An inner product space $(H, \langle \cdot, \cdot \rangle)$ which is complete with respect to $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ defined by the inner product is a *Hilbert space*.

Example

i) \mathbb{R}^n and $\ell^2(\mathbb{R})$ are complete.

ii) $C([a, b])$ is *not* complete w.r.t. the norm

$$\|f\|^2 = \int_a^b |f(x)|^2 dx.$$

Let $a = -1$, $b = 1$, and define

$$f_n(x) := \begin{cases} 0, & -1 \leq x < 0, \\ nx, & 0 \leq x \leq \frac{1}{n}, \\ 1, & \frac{1}{n} < x \leq 1. \end{cases}$$

Then f_n is continuous, and if $H(x) = \chi_{[0,1]}(x) = \begin{cases} 0, & -1 \leq x \leq 0, \\ 1, & 0 < x \leq 1, \end{cases}$ we have

$$\begin{aligned} \int_{-1}^1 |f_n(x) - H(x)|^2 dx &= \int_0^{1/n} |nx - 1|^2 dx = \int_0^{1/n} (n^2 x^2 - 2nx + 1) dx \\ &= \left[\frac{n^2 x^3}{3} - nx^2 + x \right]_{x=0}^{x=1/n} = \frac{1}{3n} - \frac{1}{n} + \frac{1}{n} = \frac{1}{3n} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

We have $\|f_n - H\| \rightarrow 0$, but $H \notin C([-1, 1])$.

However, note that $C([a, b])$ is complete w.r.t. the sup-norm $\|f\|_\infty = \sup_{a \leq x \leq b} |f(x)|$, but $\|\cdot\|_\infty \neq \|\cdot\|$ and there is no inner product inducing $\|\cdot\|_\infty$ -norm (exercise).

If one wishes to consider function spaces equipped with inner product norms, one is led to L^2 spaces.

Definition

Let $D \subset \mathbb{R}^n$ be a Lebesgue measurable set. Then

$$L^2(D) := \{f \mid f: D \rightarrow \mathbb{R} \text{ measurable}, \int_D |f(x)|^2 dx < \infty\}.$$

We define the inner product

$$\langle f, g \rangle_{L^2(D)} = \int_D f(x)g(x) dx, \quad (1)$$

which induces the norm

$$\|f\|_{L^2(D)} = \left(\int_D |f(x)|^2 dx \right)^{1/2}.$$

Theorem

$L^2(D)$ is a Hilbert space with the inner product (1).

Remark. In practice, we treat the elements of $L^2(D)$ (resp. $L^p(D)$) as functions. Strictly speaking, elements of $L^2(D)$ (resp. $L^p(D)$) are equivalence classes of measurable functions that are equal almost everywhere on D . That is, if $f, g \in L^2(D)$ and $f(x) = g(x)$ for almost every $x \in D$, then f and g represent the same element in $L^2(D)$. This identification ensures that $L^2(D)$ is a true normed space (and in fact a Hilbert space), since the norm is zero if and only if the function is zero almost everywhere.

Bounded linear operators in Hilbert spaces

Definition

Let X and Y be normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. A linear operator $A: X \rightarrow Y$ is said to be *bounded* if there exists $C > 0$ such that

$$\|Ax\|_Y \leq C\|x\|_X \quad \text{for all } x \in X.$$

Lemma

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces. Then a linear operator $A: X \rightarrow Y$ is bounded iff

$$\|A\| := \|A\|_{X \rightarrow Y} := \sup_{\|x\|_X \leq 1} \|Ax\|_Y < \infty. \quad (\text{operator norm})$$

Proof. " \Rightarrow " If there is $C > 0$ s.t. $\|Ax\|_Y \leq C\|x\|_X$ for all $x \in X$, then clearly

$$\|A\| = \sup_{\|x\|_X \leq 1} \|Ax\|_Y \leq C.$$

" \Leftarrow " Let $\|A\| < \infty$. Since $\|\frac{x}{\|x\|_X}\|_X = 1$ for all $x \neq 0$, from the linearity of A we infer

$$\frac{\|Ax\|_Y}{\|x\|_X} = \|A(\frac{x}{\|x\|_X})\|_Y \leq \|A\| \quad \text{for all } x \in X.$$

This implies the important estimate

$$\|Ax\|_Y \leq \|A\|\|x\|_X \quad \text{for all } x \in X.$$

A linear operator is continuous precisely when it is bounded.

Proposition

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces and $A: X \rightarrow Y$ a linear operator. Then the following are equivalent:

- (i) A is a bounded operator;
- (ii) A is continuous (in X);
- (iii) A is continuous at one point $x_0 \in X$.

Proof. (i) \Rightarrow (ii): if $x, y \in X$ and $\varepsilon > 0$, then

$$\|x - y\|_X \leq \frac{\varepsilon}{\|A\|} =: \delta \quad \Rightarrow \quad \|Ax - Ay\|_Y \stackrel{A \text{ linear}}{=} \|A(x - y)\|_Y \leq \|A\| \|x - y\|_X \leq \varepsilon.$$

(ii) \Rightarrow (iii): trivial.

(iii) \Rightarrow (i): let A be continuous at $x_0 \in X$. By definition, there exists $\delta > 0$ such that

$$\|y - x_0\|_X \leq \delta \quad \Rightarrow \quad \|Ay - Ax_0\|_Y \leq 1.$$

If $x \in X$ is such that $\|x\|_X \leq \delta$, then by taking $y = x + x_0$:

$$\|Ax\|_Y = \|A(x + x_0) - Ax_0\|_Y \leq 1.$$

On the other hand, for any $\|x\|_X \leq 1$, there holds $\|\delta x\|_X = \delta \|x\|_X \leq \delta$ and thus

$$\delta \|Ax\|_Y = \|A(\delta x)\|_Y \leq 1, \quad \text{i.e.,} \quad \|Ax\|_Y \leq \frac{1}{\delta} \quad \text{for all } \|x\|_X \leq 1.$$

Therefore $\|A\| \leq \frac{1}{\delta}$, meaning that A is bounded. □



Let H be a real Hilbert space.

Definition

Two elements $x, y \in H$ are said to be *orthogonal* if $\langle x, y \rangle = 0$.

Let $M \subset H$ be a subset. The orthogonal complement of M in H is defined as

$$M^\perp := \{y \in H \mid \langle x, y \rangle = 0 \text{ for all } x \in M\}.$$

We state the following easy consequences.

Lemma

For any subset $M \subset H$, M^\perp is a closed subspace of H and $M \subset (M^\perp)^\perp$.

Lemma

If M is a subspace of H , then $(M^\perp)^\perp = \overline{M}$.

Proof. Exercise. □

Proposition (Hilbert projection theorem)

Let M be a nonempty, closed, and convex[†] subset of a real Hilbert space H . Then there exists a unique element $x_0 \in M$ satisfying

$$\|x_0\| \leq \|x\| \quad \text{for all } x \in M.$$

Proof. Let $\delta = \inf\{\|x\| \mid x \in M\}$. We use the parallelogram identity

$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$ (exercise) applied to vectors $u = \frac{1}{2}x$ and $v = \frac{1}{2}y$, $x, y \in M$, to obtain

$$\frac{1}{4}\|x - y\|^2 = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \left\|\frac{x + y}{2}\right\|^2.$$

Due to convexity $\frac{1}{2}(x + y) \in M$, so

$$\|x - y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 - 4\delta^2 \quad \text{for all } x, y \in M. \tag{2}$$

Existence: let $(x_k)_{k=1}^{\infty} \subset M$ s.t. $\|x_k\| \xrightarrow{k \rightarrow \infty} \delta$. Substituting $x \leftarrow x_n$ and $y \leftarrow x_m$ in (2) yields $\|x_n - x_m\|^2 \leq 2\|x_n\|^2 + 2\|x_m\|^2 - 4\delta^2$, since $\frac{1}{2}(x_n + x_m) \in M$ for all n, m . Thus $\|x_n - x_m\| \rightarrow 0$ as $n, m \rightarrow \infty$. $(x_k)_{k=1}^{\infty}$ is Cauchy in the Hilbert space H , so there exists $x_0 := \lim_{k \rightarrow \infty} x_k \in H$. Since $\|\cdot\|$ is continuous, $\|x_0\| = \lim_{k \rightarrow \infty} \|x_k\| = \delta$. Since M is closed and $(x_k)_{k=1}^{\infty} \subset M$, the limit $x_0 \in M$.

Uniqueness: If $\|x\| = \|y\| = \delta \Rightarrow \|x - y\|^2 \leq 0$ by (2) and so $x = y$. □

[†] $tx + (1 - t)y \in M$ for all $x, y \in M$, $t \in (0, 1)$.

Corollary

Let H be a real Hilbert space, M a nonempty, closed, and convex subset of H , and $x \in H$. Then there exists a unique element $y_0 \in M$ such that

$$\|x - y_0\| = \inf\{\|x - y\| \mid y \in M\}.$$

Proof. The set $x - M := \{x - y \mid y \in M\}$ is closed and convex, and $\min\{\|x - y\| \mid x - y \in x - M\} = \min\{\|x - y\| \mid y \in M\}$. The claim follows from the previous result. □

Proposition (Orthogonal decomposition)

If M is a closed subspace of a real Hilbert space H , then

$$H = M \oplus M^\perp,$$

which means that every element $y \in H$ can be uniquely represented as

$$y = x + x^\perp, \quad x \in M, \quad x^\perp \in M^\perp.$$

Proof. It suffices to prove that $M \cap M^\perp = \{0\}$ and $M + M^\perp = H$.

- If $x \in M \cap M^\perp$, then $0 = \langle x, x \rangle = \|x\|^2$ (i.e., $x \perp x$) so $x = 0$.
 $\therefore M \cap M^\perp = \{0\}$.
- Let $x \in H$. The Hilbert projection theorem guarantees that there exists a unique $y_0 \in M$ such that

$$\|x - y_0\| \leq \|x - y\| \quad \text{for all } y \in M. \quad (3)$$

Let $x_0 = x - y_0$ so that $x = y_0 + x_0 \in M + x_0$. It remains to show that $x_0 \in M^\perp$.

The inequality (3) can be written as

$$\|x_0\| \leq \|z\| \quad \text{for all } z \in x - M.$$

Since $y_0 \in M$ and M is a vector space, $y_0 + M = M$ and $M = -M$ which implies $x - M = x + M = y_0 + x_0 + M = x_0 + M$. The previous inequality can be recast as

$$\|x_0\| \leq \|z\| \quad \text{for all } z \in x_0 + M \quad \Leftrightarrow \quad \|x_0\| \leq \|x_0 + z\| \quad \text{for all } z \in M.$$

This statement is true if and only if $\langle x_0, z \rangle = 0$ for all $z \in M$. Therefore $x_0 \in M^\perp$. □

Let M be a closed subspace. The orthogonal decomposition implies that every element $y \in H$ can be uniquely represented as

$$y = x + x^\perp, \quad x \in M, \quad x^\perp \in M^\perp.$$

Lemma

Let $M \subset H$ be a closed subspace. The mapping $P_M: H \rightarrow M$, $y \mapsto x$, is an orthogonal projection, i.e., $P_M^2 = P_M$ and $\text{Ran}(P_M) \perp \text{Ran}(I - P_M)$. It satisfies the following properties:

- P_M is linear;
- $\|P_M\| = 1$ if $M \neq \{0\}$;
- $I - P_M = P_{M^\perp}$;
- $\|y - P_M y\| \leq \|y - z\|$ for all $z \in M$;
- $y \in M \Rightarrow P_M y = y$, $(I - P_M)y = 0$;
 $y \in M^\perp \Rightarrow P_M y = 0$, $(I - P_M)y = y$.
- $\|y\|^2 = \|P_M y\|^2 + \|(I - P_M)y\|^2$ (Pythagoras)

Proof. See for example [Rudin, Real and Complex Analysis, pp. 34–35].



Example

Let H_1 and H_2 be real Hilbert spaces and let $A: H_1 \rightarrow H_2$ be a continuous linear operator.

The kernel (or null space) of operator A is defined as

$$\text{Ker}(A) := \{x \in H_1 \mid Ax = 0\}.$$

The range (or image) of operator A is defined as

$$\text{Ran}(A) := \{y \in H_2 \mid y = Ax, x \in H_1\}.$$

Then we have the following:

- $\text{Ker}(A)$ is a *closed* subspace of H_1 , and $\text{Ran}(A)$ is a subspace of H_2 .
- $H_1 = \text{Ker}(A) \oplus (\text{Ker}(A))^\perp$.
- $H_2 = \overline{\text{Ran}(A)} \oplus (\text{Ran}(A))^\perp$.

We denote

$$\mathcal{L}(X, Y) := \{A \mid A: X \rightarrow Y \text{ is bounded and linear}\}.$$

Proposition

Let X and Y be normed spaces. If Y is complete, then $\mathcal{L}(X, Y)$ is complete w.r.t. the operator norm (i.e., it is a Banach space).

Proof. Let $x \in X$ and assume that $A_k \in \mathcal{L}(X, Y)$, $k \in \mathbb{N}$, is a Cauchy sequence. If $x = 0$, then $A_k 0 = 0$ and the limit $A(0) := \lim_{k \rightarrow \infty} A_k 0 = 0$ trivially exists. On the other hand, if $x \neq 0$, then for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$m, n > N \quad \Rightarrow \quad \|A_m - A_n\| < \frac{\varepsilon}{\|x\|}.$$

Especially,

$$\|A_m x - A_n x\|_Y \leq \|A_m - A_n\| \|x\|_X < \varepsilon \quad \text{when } m, n > N,$$

so $(A_k x)$ is a Cauchy sequence in Y and therefore the limit

$$A(x) := \lim_{k \rightarrow \infty} A_k x$$

exists.

It is easy to see that $A(x) := \lim_{k \rightarrow \infty} A_k x$ is linear. It is also bounded: there exists $N \in \mathbb{N}$ such that

$$m, n > N \Rightarrow \|A_m - A_n\| < 1.$$

Fix $m > N$. Then for all $n > m$,

$$\|A_n\| < 1 + \|A_m\|$$

and thus

$$\|A_n x\|_Y \leq (1 + \|A_m\|) \|x\|_X.$$

But $\|Ax\|_Y = \lim_{n \rightarrow \infty} \|A_n x\|_Y \leq (1 + \|A_m\|) \|x\|_X$. Therefore A is bounded.

Finally, we need to show that $\|A_n - A\| \rightarrow 0$ as $n \rightarrow \infty$. Since we assumed $(A_k)_{k=1}^{\infty}$ to be Cauchy, let $\varepsilon > 0$ be s.t. for $m, n > N$, there holds $\|A_m - A_n\| < \varepsilon$. Then

$$\begin{aligned} \|(A - A_n)x\|_Y &= \lim_{m \rightarrow \infty} \|A_m x - A_n x\|_Y \leq \varepsilon \|x\|_X \quad \text{for all } x \in X \\ \Rightarrow \|A - A_n\| &< \varepsilon. \end{aligned}$$

Hence $\|A - A_n\| \rightarrow 0$ as $n \rightarrow \infty$.

□

If $X = H_1$ and $Y = H_2$ are Hilbert spaces, then $\mathcal{L}(H_1, H_2)$ is a complete normed space.

Definition

Let H be a Hilbert space. The space $H' := \mathcal{L}(H, \mathbb{R})$ is called the *topological dual space* of H .

Corollary

If H is a Hilbert space, then H' is complete w.r.t. the operator norm.

Proof. This is an immediate consequence of the previous proposition since \mathbb{R} is a complete Hilbert space. □

Remark. In general, $\mathcal{L}(H_1, H_2)$ is *not* a Hilbert space even when both H_1 and H_2 are. However, in the special case $H' = \mathcal{L}(H, \mathbb{R})$ it turns out that indeed one can associate an inner product that induces the operator norm $\|\cdot\|$ – meaning that H' is a Hilbert space! This is made possible by the *Riesz representation theorem*.

Existence results

Proposition (Riesz representation theorem)

Let H be a real Hilbert space. If $A: H \rightarrow \mathbb{R}$ is a bounded linear functional, i.e., A is linear and there exists $C > 0$ such that

$$|A(x)| \leq C\|x\| \quad \text{for all } x \in H,$$

then there exists a unique $y \in H$ such that

$$A(x) = \langle x, y \rangle \quad \text{for all } x \in H.$$

Proof. If $A \equiv 0$, then $y = 0$ and this is unique. Suppose $A \neq 0$ and let

$$M := \text{Ker}(A) = \{x \in H \mid A(x) = 0\}.$$

Since A is continuous, M is a closed subspace of H . Furthermore, by the orthogonal decomposition $H = M \oplus M^\perp$, our assumption $A \neq 0$ implies that $M \neq H \Rightarrow M^\perp \neq \{0\}$.

Let $x \in H$ and $z \in M^\perp$ with $\|z\| = 1$. Define

$$u := A(x)z - A(z)x.$$

Then

$$A(u) = A(x)A(z) - A(z)A(x) = 0$$

meaning that $u \in M$. In particular $\langle u, z \rangle = \langle A(x)z - A(z)x, z \rangle = 0$ and

$$\begin{aligned} A(x) &= A(x) \underbrace{\langle z, z \rangle}_{=\|z\|^2=1} = \langle A(x)z, z \rangle \\ &= \langle A(z)x, z \rangle = A(z)\langle x, z \rangle = \langle x, zA(z) \rangle. \end{aligned}$$

\therefore The element $y = zA(z)$ satisfies $A(x) = \langle x, y \rangle$.

To prove uniqueness, suppose that there exist $y_1, y_2 \in H$ such that

$$A(x) = \langle x, y_1 \rangle = \langle x, y_2 \rangle.$$

Then $\langle x, y_1 - y_2 \rangle = 0$ for all $x \in H$. Choose $x = y_1 - y_2$. Then

$$0 = \langle y_1 - y_2, y_1 - y_2 \rangle = \|y_1 - y_2\|^2 \Leftrightarrow y_1 = y_2.$$



The Riesz operator

Let $x \in H$ and consider the linear mapping $f_x: H \rightarrow \mathbb{R}$, $z \mapsto \langle z, x \rangle_H$. Note that $f_x \in H'$ since it follows from the Cauchy–Schwarz inequality that

$$|f_x(z)| = |\langle z, x \rangle_H| \leq \|z\|_H \|x\|_H \quad \text{for all } z \in H. \quad (4)$$

Now define the *Riesz operator* $R_H: H \rightarrow H'$ as $x \mapsto f_x$.

- R_H is linear: $R_H(ax_1 + bx_2) = f_{ax_1+bx_2} = \langle \cdot, ax_1 + bx_2 \rangle_H = a\langle \cdot, x_1 \rangle_H + b\langle \cdot, x_2 \rangle_H = af_{x_1} + bf_{x_2} = aR_Hx_1 + bR_Hx_2$ for $x_1, x_2 \in H$, $a, b \in \mathbb{R}$.
- R_H is an isometry ($\|R_Hx\|_{H'} = \|x\|_H$): it follows from (4) that $\|R_Hx\|_{H'} = \|f_x\|_{H'} = \sup_{\|z\|_H \leq 1} |\langle z, x \rangle_H| \leq \|x\|_H$. The other direction follows from $\|x\|_H^2 = \langle x, x \rangle_H = f_x(x) = |f_x(x)| \leq \|f_x\|_{H'} \|x\|_H = \|R_Hx\|_{H'} \|x\|_H$.
- R_H is injective (one-to-one): let $R_Hx = R_Hy$ for some $x, y \in H$. From linearity, $R_H(x - y) = 0 \Rightarrow f_{x-y} = 0 \Rightarrow \langle x - y, z \rangle_H = 0$ for all $z \in H \Rightarrow x = y$.
- R_H is surjective (onto): by Riesz representation theorem, given $A \in H'$, there exists a unique $x \in H$ satisfying $A(z) = \langle z, x \rangle_H = f_x(z)$ for all $z \in H$. In other words, $A = \langle \cdot, x \rangle_H = f_x = R_Hx$.

∴ The Riesz operator $R_H: H \rightarrow H'$ is a bijective linear operator isometry.

Lemma

Let H be a Hilbert space. The dual space $H' := \mathcal{L}(H, \mathbb{R})$ is a Hilbert space induced by

$$\|A\|_{H'} := \sup_{\|x\|_H \leq 1} |Ax| = \sqrt{\langle A, A \rangle_{H'}}, \quad \langle A, B \rangle_{H'} := \langle R_H^{-1}A, R_H^{-1}B \rangle_H.$$

Adjoint operator

Proposition

Let H_1 and H_2 be real Hilbert spaces and suppose that $A \in \mathcal{L}(H_1, H_2)$. Then there exists a unique bounded linear operator $A^* : H_2 \rightarrow H_1$, called the adjoint of A , satisfying $\langle Ax, y \rangle_{H_2} = \langle x, A^*y \rangle_{H_1}$. Moreover, $\|A\|_{H_1 \rightarrow H_2} = \|A^*\|_{H_2 \rightarrow H_1}$.

Proof. Let $y \in H_2$ and consider $T_y : H_1 \rightarrow \mathbb{R}$, $x \mapsto \langle Ax, y \rangle_{H_2}$. Clearly, T_y is linear and bounded so by the Riesz representation theorem there exists a unique $z \in H_1$ s.t.

$$\langle Ax, y \rangle_{H_2} = T_y(x) = \langle x, z \rangle_{H_1} \quad \text{for all } x \in H_1.$$

Define $A^*y := z$.

- Let $a, b \in \mathbb{R}$ and $y_1, y_2 \in H_2$. Linearity follows from

$$\langle x, A^*(ay_1 + by_2) \rangle = \langle Ax, ay_1 + by_2 \rangle = a\langle Ax, y_1 \rangle + b\langle Ax, y_2 \rangle =$$

$a\langle x, A^*y_1 \rangle + b\langle x, A^*y_2 \rangle = \langle x, aA^*y_1 + bA^*y_2 \rangle$. Since $x \in H_1$ was arbitrary,

$$A^*(ay_1 + by_2) = aA^*y_1 + bA^*y_2.$$

- $\|A^*\|_{H_2 \rightarrow H_1} = \sup_{\|y\|_{H_2} \leq 1} \|A^*y\|_{H_1} \stackrel{(*)}{=} \sup_{\|y\|_{H_2} \leq 1} \sup_{\|x\|_{H_1} \leq 1} |\langle A^*y, x \rangle|$
 $= \sup_{\|y\|_{H_2} \leq 1} \sup_{\|x\|_{H_1} \leq 1} |\langle y, Ax \rangle| \stackrel{(*)}{=} \sup_{\|x\|_{H_1} \leq 1} \|Ax\|_{H_2} = \|A\|_{H_1 \rightarrow H_2} < \infty.$ □

(*) Let $\Lambda \in \mathcal{L}(H, K)$, H, K Hilbert spaces. Cauchy–Schwarz: $\sup_{\|y\|_K \leq 1} |\langle \Lambda x, y \rangle_K| \leq \|\Lambda x\|_K$.

Other direction: $\sup_{\|y\|_K \leq 1} |\langle \Lambda x, y \rangle_K| \geq |\langle \Lambda x, \frac{1}{\|\Lambda x\|_K} \Lambda x \rangle_K| = \|\Lambda x\|_K$.

$\therefore \|\Lambda x\|_K = \sup_{\|y\|_K \leq 1} |\langle \Lambda x, y \rangle_K|$.

Some properties of the adjoint operator

Proposition

Let H_1 and H_2 be real Hilbert spaces and suppose that $A, B \in \mathcal{L}(H_1, H_2)$. Then

- (i) $\|A^*A\|_{H_1 \rightarrow H_1} = \|A\|_{H_1 \rightarrow H_2}^2$;
- (ii) $A^{**} = A$, where $A^{**} = (A^*)^*$;
- (iii) $(c_1A + c_2B)^* = c_1A^* + c_2B^*$, $c_1, c_2 \in \mathbb{R}$.

Proof. (i) Let $x \in H_1$, $\|x\|_{H_1} = 1$. By the Cauchy–Schwarz inequality,

$$\|Ax\|_{H_2}^2 = \langle Ax, Ax \rangle_{H_2} = \langle x, A^*Ax \rangle_{H_1} \leq \|A^*Ax\|_{H_1} \Rightarrow \|A\|_{H_1 \rightarrow H_2}^2 \leq \|A^*A\|_{H_1 \rightarrow H_1}.$$

Other direction: $\|A^*A\| \leq \|A^*\| \cdot \|A\| = \|A\|^2$.

(ii) If $x \in H_1$ and $y \in H_2$, then

$$\langle A^{**}x, y \rangle_{H_2} = \langle x, A^*y \rangle_{H_1} = \langle A^*y, x \rangle_{H_1} = \langle y, Ax \rangle_{H_2} = \langle Ax, y \rangle_{H_2}.$$

Hence $\langle A^{**}x - Ax, y \rangle_{H_2} = 0$ for all $y \in H_2 \Rightarrow A^{**}x = Ax$ for all $x \in H_1 \Rightarrow A^{**} = A$.

(iii) Let $x \in H_1$ and $y \in H_2$. Then

$$\begin{aligned}\langle (c_1A + c_2B)^*y, x \rangle_{H_1} &= \langle y, (c_1A + c_2B)x \rangle_{H_2} = c_1\langle y, Ax \rangle_{H_2} + c_2\langle y, Bx \rangle_{H_2} \\ &= c_1\langle A^*y, x \rangle_{H_1} + c_2\langle B^*y, x \rangle_{H_1} = \langle (c_1A^* + c_2B^*)y, x \rangle_{H_1}.\end{aligned}$$

Similarly to the previous part, we conclude that $(c_1A + c_2B)^* = c_1A^* + c_2B^*$. □

Self-adjoint operators

Definition

Let H be a Hilbert space. The operator $A \in \mathcal{L}(H) := \mathcal{L}(H, H)$ is called *self-adjoint* if $A^* = A$, i.e.,

$$\langle Ax, y \rangle = \langle x, Ay \rangle \quad \text{for all } x, y \in H.$$

Example

Let H be a Hilbert space and let $A, B \in \mathcal{L}(H)$ be self-adjoint operators. Then

- (i) $A + B$ is self-adjoint.
- (ii) if $c \in \mathbb{R}$, then cA is self-adjoint.
- (iii) if $AB = BA$, then AB is self-adjoint.

Parts (i) and (ii) follow immediately from part (iii) on the previous slide. If $x, y \in H$, then

$$\langle ABx, y \rangle = \langle BAx, y \rangle = \langle Ax, By \rangle = \langle x, ABy \rangle \quad \Rightarrow \quad (AB)^* = AB.$$

Example

Let H be a Hilbert space and $M \subset H$ a closed subspace. Then the orthogonal projections $P_M: H \rightarrow M$ and $I - P_M =: P_{M^\perp}: H \rightarrow M^\perp$ are self-adjoint.

Lax–Milgram lemma

Proposition (Lax–Milgram lemma)

Let H be a real Hilbert space and let $B: H \times H \rightarrow \mathbb{R}$ be a bilinear mapping[†] with $C, c > 0$ such that

$$|B(u, v)| \leq C\|u\| \cdot \|v\| \quad \text{for all } u, v \in H, \quad (\text{boundedness})$$

$$B(u, u) \geq c\|u\|^2 \quad \text{for all } u \in H. \quad (\text{coercivity})$$

Let $F: H \rightarrow \mathbb{R}$ be a bounded linear mapping. Then there exists a unique element $u \in H$ satisfying

$$B(u, v) = F(v) \quad \text{for all } v \in H$$

and

$$\|u\| \leq \frac{1}{c}\|F\|.$$

[†] $B(u + v, w) = B(u, w) + B(v, w)$, $B(au, v) = aB(u, v)$,
 $B(u, v + w) = B(u, v) + B(u, w)$, $B(u, av) = aB(u, v)$
for all $u, v, w \in H$ and $a \in \mathbb{R}$.

Proof. We split the proof into several steps.

Step 1. Let $v \in H$ be fixed. Then the mapping

$$T: w \mapsto B(v, w), \quad H \rightarrow \mathbb{R},$$

is bounded and linear. It follows from the Riesz representation theorem that there exists a unique element $a \in H$ with

$$Tw = \langle a, w \rangle \quad \text{for all } w \in H.$$

Let us define the mapping $A: H \rightarrow H$ by setting

$$Av = a.$$

Then

$$B(v, w) = \langle Av, w \rangle \quad \text{for all } v, w \in H.$$

Step 2. We show that the mapping $A: H \rightarrow H$ is linear and bounded.
Clearly,

$$\begin{aligned}\langle A(c_1v_1 + c_2v_2), w \rangle &= B(c_1v_1 + c_2v_2, w) \\&= c_1B(v_1, w) + c_2B(v_2, w) \\&= \langle c_1Av_1 + c_2Av_2, w \rangle\end{aligned}$$

for all $w \in H$, so $A(c_1v_1 + c_2v_2) = c_1Av_1 + c_2Av_2$. Moreover,

$$\begin{aligned}\|Av\|^2 &= \langle Av, Av \rangle \\&= B(v, Av) \\&\leq C\|v\|\|Av\|\end{aligned}$$

which implies that

$$\|Av\| \leq C\|v\|.$$

Step 3. We show that

$$\begin{cases} A \text{ is one-to-one,} \\ \text{Ran}(A) = AH \text{ is closed in } H. \end{cases}$$

We begin by noting that

$$c\|v\|^2 \leq B(v, v) = \langle Av, v \rangle \leq \|Av\|\|v\|$$

and thus

$$\|Av\| \geq c\|v\| \quad \text{for all } v \in H. \tag{5}$$

Especially

$$Av = Aw \Rightarrow A(v - w) = 0 \Rightarrow 0 = \|A(v - w)\| \geq c\|v - w\| \geq 0 \Rightarrow v = w$$

so A is one-to-one.

To see that $\text{Ran}(A)$ is closed, let $y_j = Ax_j \in \text{Ran}(A)$. The goal is to show that $y := \lim_{j \rightarrow \infty} y_j \in \text{Ran}(A)$. We observe that

$$\lim_{j,k \rightarrow \infty} \|x_j - x_k\| \stackrel{(5)}{\leq} \lim_{j,k \rightarrow \infty} \frac{1}{c} \|y_j - y_k\| = 0,$$

i.e., $(x_j)_{j=1}^{\infty}$ is Cauchy and $x := \lim_{j \rightarrow \infty} x_j \in H$ exists by completeness. Moreover,

$$\lim_{j \rightarrow \infty} \|Ax_j - Ax\| \leq \lim_{j \rightarrow \infty} \|A\| \|x_j - x\| \leq C \lim_{j \rightarrow \infty} \|x_j - x\| = 0$$

and therefore

$$y = \lim_{j \rightarrow \infty} Ax_j = Ax \in \text{Ran}(A).$$

Step 4. We show that $\overline{\text{Ran}(A)} = H$. We prove this by contradiction: suppose that $\text{Ran}(A) = \overline{\text{Ran}(A)} \neq H$. Then there exists $w \in \text{Ran}(A)^\perp$, $w \neq 0$.[†] This implies that

$$\|w\|^2 \leq \frac{1}{c} B(w, w) = \frac{1}{c} \langle Aw, w \rangle = 0,$$

i.e., $w = 0$. This contradiction shows that $\text{Ran}(A) = H$. Therefore $A: H \rightarrow H$ is a continuous bijection.

Step 5. Existence of a solution. We use the Riesz representation theorem: since $F: H \rightarrow \mathbb{R}$ is linear and continuous, there exists $b \in H$ such that

$$F(v) = \langle b, v \rangle \quad \text{for all } v \in H.$$

Define $u := A^{-1}b$. Hence

$$\begin{aligned} Au = b &\Leftrightarrow \langle Au, v \rangle = \langle b, v \rangle \quad \text{for all } v \in H \\ &\Leftrightarrow B(u, v) = F(v) \quad \text{for all } v \in H. \end{aligned}$$

[†]Since $(\text{Ran}(A)^\perp)^\perp = \overline{\text{Ran}(A)} \neq H \Rightarrow (\text{Ran}(A))^\perp \neq \{0\}$.

Step 6. Uniqueness. Suppose that

$$\begin{aligned}B(u_1, w) &= F(w) \quad \text{for all } w \in H, \\B(u_2, w) &= F(w) \quad \text{for all } w \in H.\end{aligned}$$

Let $u := u_1 - u_2$. By linearity,

$$B(u, w) = 0 \quad \text{for all } w \in H.$$

The coercivity of B implies that

$$\|u\|^2 \leq \frac{1}{c} B(u, u) = 0$$

so that $u = 0$, i.e., $u_1 = u_2$.

Step 7. A priori bound. If $B(u, w) = F(w)$ for all $w \in H$, then by setting $w = u$ we obtain

$$\|u\|^2 \leq \frac{1}{c} B(u, u) = \frac{1}{c} F(u) \leq \frac{1}{c} \|F\| \|u\|$$

which immediately yields

$$\|u\| \leq \frac{1}{c} \|F\|.$$



Density argument

Lemma

Let X, Y be Banach spaces and let $Z \subset X$ be a dense subspace. If $T: Z \rightarrow Y$ is a linear mapping such that

$$\|Tx\|_Y \leq C\|x\|_X, \quad x \in Z, \tag{6}$$

then there exists a unique extension $\tilde{T}: X \rightarrow Y$ with $\tilde{T}|_Z = T$ and

$$\|\tilde{T}x\|_Y \leq C\|x\|_X, \quad x \in X. \tag{7}$$

Moreover, if (6) holds with equality, then so does (7).

Proof. Let $x \in X$. Because $Z \subset X$ is dense, there exists a sequence $(z_k)_{k=1}^{\infty} \subset Z$ s.t. $\|z_k - x\|_X \xrightarrow{k \rightarrow \infty} 0$. Let $\varepsilon > 0$. Since $(z_k)_{k=1}^{\infty}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ s.t.

$$m, n \geq N \quad \Rightarrow \quad \|z_m - z_n\|_X < \frac{\varepsilon}{C}.$$

Then there holds

$$\|Tz_m - Tz_n\|_Y = \|T(z_m - z_n)\|_Y \leq C\|z_m - z_n\|_X < \varepsilon,$$

which means that $(Tz_k)_{k=1}^{\infty}$ is a Cauchy sequence in Y . Since Y is complete, there exists $y := \lim_{k \rightarrow \infty} Tz_k$. Hence we may define $\tilde{T}: X \rightarrow Y$ by setting $\tilde{T}(x) = y$.

We begin by showing that \tilde{T} is well-defined. Let $(z_k)_{k=1}^{\infty}$, $(\tilde{z}_k)_{k=1}^{\infty}$ be two sequences in Z s.t. $z_k, \tilde{z}_k \xrightarrow{k \rightarrow \infty} x$ in X . Then

$$\|Tz_k - T\tilde{z}_k\|_Y = \|T(z_k - \tilde{z}_k)\|_Y \leq C\|z_k - \tilde{z}_k\| \leq C\|z_k - x\| + C\|\tilde{z}_k - x\| \xrightarrow{k \rightarrow \infty} 0.$$

Recalling that $\tilde{T}(x) := \lim_{k \rightarrow \infty} Tz_k$, we obtain

$$\|T\tilde{z}_k - \tilde{T}(x)\| \leq \|T\tilde{z}_k - Tz_k\| + \|Tz_k - \tilde{T}(x)\| \xrightarrow{k \rightarrow \infty} 0,$$

showing that \tilde{T} is well-defined.

Next we show that \tilde{T} is linear. Let $x, \tilde{x} \in X$ and $a, b \in \mathbb{R}$. Let $Z \ni z_k \xrightarrow{k \rightarrow \infty} x$ and $Z \ni \tilde{z}_k \xrightarrow{k \rightarrow \infty} \tilde{x}$. Now $az_k + b\tilde{z}_k \in Z$ and $Z \ni az_k + b\tilde{z}_k \rightarrow ax + b\tilde{x}$. Thus

$$\tilde{T}(ax + b\tilde{x}) = \lim_{k \rightarrow \infty} T(az_k + b\tilde{z}_k) = a \lim_{k \rightarrow \infty} Tz_k + b \lim_{k \rightarrow \infty} T\tilde{z}_k = a\tilde{T}x + b\tilde{T}\tilde{x},$$

since the limit is linear.[†]

Since the norm is continuous,

$$\|\tilde{T}x\| = \left\| \lim_{k \rightarrow \infty} Tx_k \right\| = \lim_{k \rightarrow \infty} \|Tx_k\| \leq C \lim_{k \rightarrow \infty} \|x_k\| = C\|x\|.$$

Finally, $\tilde{T}|_Z = T$ holds by construction and the uniqueness of the limit $Tz_k \rightarrow y$ ensures that there cannot exist another mapping $L: X \rightarrow Y$ s.t. $L|_Z = T$ and $\|Lx\| \leq C\|x\|$. \square

[†]Let $y := \lim_{k \rightarrow \infty} Tz_k$ and $\tilde{y} := \lim_{k \rightarrow \infty} T\tilde{z}_k$.

Then $\|T(az_k + b\tilde{z}_k) - ay - b\tilde{y}\| \leq a\|Tz_k - y\| + b\|T\tilde{z}_k - \tilde{y}\| \rightarrow 0$.

Hence $\lim_{k \rightarrow \infty} T(az_k + b\tilde{z}_k) = a \lim_{k \rightarrow \infty} Tz_k + b \lim_{k \rightarrow \infty} T\tilde{z}_k$.

Multi-index notation

A vector of the form $\alpha := (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ is called a *multi-index*. We denote the j^{th} component of multi-index α by α_j .

The *order* (or *modulus*) of a multi-index is defined as

$$|\alpha| := \alpha_1 + \cdots + \alpha_d.$$

Let $x := (x_j)_{j=1}^d \in \mathbb{R}^d$. We define the monomial notation

$$x^\alpha := \prod_{j=1}^d x_j^{\alpha_j},$$

where $0^0 := 1$, and the corresponding formula for partial derivatives

$$\partial^\alpha := \partial_x^\alpha := \prod_{j=1}^d \frac{\partial^{\alpha_j}}{\partial x_j^{\alpha_j}}.$$

Other often used multi-index notations include $\binom{\alpha}{\beta} := \prod_{j=1}^d \binom{\alpha_j}{\beta_j}$, $\alpha! := \alpha_1! \cdots \alpha_d!$ (but $|\alpha|! := (\alpha_1 + \cdots + \alpha_d)!)$, etc.

Some function spaces

Let $D \subset \mathbb{R}^d$ be a nonempty open set. Let us recall the following function spaces.

Definition

$$C(D) := \{u: D \rightarrow \mathbb{R} \mid u \text{ is continuous}\},$$

$$C^k(D) := \{u: D \rightarrow \mathbb{R} \mid \exists \partial^\alpha u \text{ is continuous for all } |\alpha| \leq k, \alpha \in \mathbb{N}_0^d\},$$

$$C^\infty(D) := \{u: D \rightarrow \mathbb{R} \mid \exists \partial^\alpha u \text{ is continuous for all } \alpha \in \mathbb{N}_0^d\} = \bigcap_{k=0}^{\infty} C^k(D),$$

$$C_0^k(D) := \{u \in C^k(D) \mid \text{supp}(u) \subset D \text{ is a compact set}\},$$

$$C_0^\infty(D) := \{u \in C^\infty(D) \mid \text{supp}(u) \subset D \text{ is a compact set}\},$$

$$\text{where } \text{supp}(u) := \overline{\{\mathbf{x} \in D \mid u(\mathbf{x}) \neq 0\}},$$

$$L^1(D) := \{u: D \rightarrow \mathbb{R} \mid u \text{ is measurable}, \|u\|_{L^1(D)} := \int_D |u(\mathbf{x})| d\mathbf{x} < \infty\}.$$

Remark. Recall that in the Euclidean space \mathbb{R}^d , a set is compact iff it is closed and bounded. This is the *Heine–Borel theorem*.

Let $D \subset \mathbb{R}^d$ be open.

Definition

$$L^1_{\text{loc}}(D) := \{u: D \rightarrow \mathbb{R} \mid u \in L^1(K) \text{ for all compact } K \subset D\}$$

Example

Let $u \in C^1(D)$. Then integration by parts yields

$$\int_D u(\mathbf{x}) \partial_{x_i} \varphi(\mathbf{x}) \, d\mathbf{x} = - \int_D \partial_{x_i} u(\mathbf{x}) \varphi(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } \varphi \in C_0^\infty(D). \quad (8)$$

If $u \in C^k(D)$ and $\alpha \in \mathbb{N}_0^d$ is a multi-index with order

$|\alpha| := \nu_1 + \cdots + \nu_d = k$, then we obtain from repeated application of (8) that

$$\int_D u(\mathbf{x}) \partial^\alpha \varphi(\mathbf{x}) \, d\mathbf{x} = (-1)^{|\alpha|} \int_D \partial^\alpha u(\mathbf{x}) \varphi(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } \varphi \in C_0^\infty(D).$$

The so-called *weak derivative* is a generalization of the classical derivative.

Definition (Weak derivative)

Let $u, w \in L^1_{\text{loc}}(D)$. We call w the *weak ∂_{x_i} derivative of u* and denote $w = \partial_{x_i} u$ if

$$\int_D w(\mathbf{x})\varphi(\mathbf{x}) d\mathbf{x} = - \int_D u(\mathbf{x})\partial_{x_i}\varphi(\mathbf{x}) d\mathbf{x} \quad \text{for all } \varphi \in C_0^\infty(D).$$

Moreover, we call w the *weak ∂^α derivative of u* and denote $w = \partial^\alpha u$ if

$$\int_D w(\mathbf{x})\varphi(\mathbf{x}) d\mathbf{x} = (-1)^{|\alpha|} \int_D u(\mathbf{x})\partial^\alpha\varphi(\mathbf{x}) d\mathbf{x} \quad \text{for all } \varphi \in C_0^\infty(D).$$

This definition ensures that the integration by parts formula is valid if the weak derivative exists.

Remark. This definition generalizes the classical derivative: if $u \in C^1(D)$, then the weak derivative coincides with the classical one.

Weak derivative

Example

Let $d = 1$, $D = (0, 2)$, and

$$u(x) = \begin{cases} x & \text{if } 0 < x \leq 1, \\ 1 & \text{if } 1 \leq x < 2. \end{cases}$$

Define

$$v(x) = \begin{cases} 1 & \text{if } 0 < x \leq 1, \\ 0 & \text{if } 1 \leq x < 2. \end{cases}$$

We claim $u' = v$ in the weak sense, i.e., $\int_0^2 u(x)\varphi'(x) dx = -\int_0^2 v(x)\varphi(x) dx$ for all $\varphi \in C_0^\infty(D)$. Let $\varphi \in C_0^\infty(D)$ be arbitrary. Then

$$\begin{aligned} \int_0^2 u(x)\varphi'(x) dx &= \int_0^1 x\varphi'(x) dx + \int_1^2 \varphi'(x) dx \\ &= \underbrace{[x\varphi(x)]}_{=\varphi(1)-0} \Big|_{x=0}^{x=1} - \int_0^1 \varphi(x) dx + \underbrace{\varphi(2)}_{=0} - \varphi(1) = - \int_0^1 \varphi(x) dx = - \int_0^2 v(x)\varphi(x) dx \end{aligned}$$

as desired.

Sobolev spaces

Sobolev spaces

Definition

The *Sobolev space* of order k based on $L^2(D)$ is defined by

$$H^k(D) := \{u \in L^2(D) \mid \partial^\alpha u \in L^2(D) \text{ for all } |\alpha| \leq k\},$$

and we equip this space with the norm

$$\|u\|_{H^k(D)} = \left(\sum_{|\alpha| \leq k} \int_D |\partial^\alpha u(x)|^2 dx \right)^{1/2},$$

induced by the inner product

$$\langle u, v \rangle_{H^k(D)} = \sum_{|\alpha| \leq k} \int_D \partial^\alpha u(x) \partial^\alpha v(x) dx.$$

Moreover, we define

$$H_0^k(D) := \text{cl}_{H^k(D)}(C_0^\infty(D)),$$

i.e., $H_0^k(D)$ is the closure of $C_0^\infty(D)$ in the topology of $H^k(D)$.

Proposition

- $\partial^\alpha : H^k(D) \rightarrow H^{k-|\alpha|}$, $k \geq |\alpha|$, is bounded.
- $\partial^\alpha(\partial^\beta u) = \partial^\beta(\partial^\alpha u) = \partial^{\alpha+\beta} u$, $u \in H^{|\alpha|+|\beta|}(D)$, where
 $\alpha + \beta := (\alpha_1 + \beta_1, \dots, \alpha_d + \beta_d)$.

Proposition

$H^k(D)$ is a Hilbert space for all $k \in \mathbb{N}$.

Proof. Let $(u_j)_{j=1}^\infty$ be a Cauchy sequence in $H^k(D)$. Then for all $|\alpha| \leq k$

$$\|D^\alpha u_m - D^\alpha u_n\|_{L^2(D)} \leq \|u_m - u_n\|_{H^k} \xrightarrow{m,n \rightarrow \infty} 0,$$

so $(D^\alpha u_j)_{j=1}^\infty$ is a Cauchy sequence in $L^2(D)$. Since $L^2(D)$ is complete, there exists $f^\alpha \in L^2(D)$ such that $\|f_\alpha - D^\alpha u_j\|_{L^2(D)} \xrightarrow{j \rightarrow \infty} 0$.

Esp. $u_j \xrightarrow{j \rightarrow \infty} f^0 := u$ in $L^2(D)$.

We show that $D^\alpha u \in L^2(D)$ for all $|\alpha| \leq k$, i.e., $u \in H^k(D)$. For $\varphi \in C_0^\infty(D)$,

$$\begin{aligned}\int_D u(\mathbf{x}) \partial^\alpha \varphi(\mathbf{x}) \, d\mathbf{x} &= \lim_{j \rightarrow \infty} \int_D u_j(\mathbf{x}) \partial^\alpha \varphi(\mathbf{x}) \, d\mathbf{x} \\ &= \lim_{j \rightarrow \infty} \int_D (-1)^{|\alpha|} \partial^\alpha u_j(\mathbf{x}) \cdot \varphi(\mathbf{x}) \, d\mathbf{x} \\ &= \int_D (-1)^{|\alpha|} f^\alpha(\mathbf{x}) \cdot \varphi(\mathbf{x}) \, d\mathbf{x}\end{aligned}$$

so $\partial^\alpha u = f^\alpha \in L^2(D)$, $|\alpha| \leq k$. Thus $u \in H^k(D)$.

Finally,

$$\|u_j - u\|_{H^k(D)}^2 = \sum_{|\alpha| \leq k} \|\partial^\alpha u_j - f^\alpha\|_{L^2(D)}^2 \xrightarrow{j \rightarrow \infty} 0$$

which means that

$$\lim_{j \rightarrow \infty} u_j = u \quad \text{in } H^k(D). \quad \square$$

The case when D is a polygon (2D) or a polyhedron (3D) will be of special interest to us. In these cases, the boundary ∂D is not smooth, which needs to be accounted for by our theory. However, it turns out that working with Lipschitz domains is sufficient for our purposes. To this end, we recall the following.

Definition

Let $D \subset \mathbb{R}^d$ be a bounded, open set. A function $u: D \rightarrow \mathbb{R}$ is *Lipschitz continuous* if there exists $L > 0$ such that

$$|u(\mathbf{x}) - u(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}|, \quad \mathbf{x}, \mathbf{y} \in D.$$

Theorem (Rademacher's theorem)

If $D \subset \mathbb{R}^d$ is an open subset and $f: D \rightarrow \mathbb{R}$ is Lipschitz continuous, then f is differentiable almost everywhere in D .

A Lipschitz hypograph $D \subset \mathbb{R}^d$ is a domain of the form

$$D = \{\mathbf{x} \in \mathbb{R}^d \mid x_d > \zeta(\mathbf{x}'), \mathbf{x}' := (x_1, \dots, x_{d-1}) \in \mathbb{R}^{d-1}\}$$

where $\zeta: \mathbb{R}^{d-1} \rightarrow \mathbb{R}$ is a Lipschitz function.

Definition (bounded Lipschitz domain)

An open, bounded set is a *Lipschitz* domain if its boundary ∂D is compact and if there exist $\{W_j\}_{j=1}^N$ and $\{D_j\}_{j=1}^N$ with the following properties:

- (i) $\{W_j\}_{j=1}^N$ is a finite open cover of ∂D , i.e., each $W_j \subset \mathbb{R}^d$ is an open set and $\partial D \subset \bigcup_{j=1}^N W_j$.
- (ii) Each D_j can be transformed into a Lipschitz hypograph by a rotation plus a translation.
- (iii) $W_j \cap D = W_j \cap D_j$ for all $j \in \{1, \dots, N\}$.

The class of Lipschitz domains is broad enough to cover most cases that arise in applications of partial differential equations. For example, any polygon in \mathbb{R}^2 or convex polyhedron in \mathbb{R}^3 is a Lipschitz domain. If $\kappa: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a C^1 diffeomorphism and D is a Lipschitz domain, then the set $\kappa(D)$ is again a Lipschitz domain.

Note that the outer normal vector is defined a.e. at the boundary and it is a.e. continuous.

Trace theorem on Lipschitz domains

Theorem

Let D be a bounded Lipschitz domain and let $\gamma: C^\infty(\overline{D}) \rightarrow C^\infty(\partial D)$ be the trace operator $\gamma u = u|_{\partial D}$. Then the trace operator has a unique extension to a bounded linear operator $\gamma: H^1(D) \rightarrow L^2(\partial D)$.

The significance of the trace theorem is that the boundary values of Sobolev functions belonging to $H^1(D)$ are well-defined in an unambiguous way.

Trace zero functions in $H_0^1(D)$

Theorem

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u \in H^1(D)$, and $\gamma: H^1(D) \rightarrow L^2(\partial D)$ the trace operator. Then

$$u \in H_0^1(D) \iff \gamma u = 0: \partial D \rightarrow \mathbb{R}.$$

Proof. “ \Rightarrow ” follows from previous discussion. “ \Leftarrow ” is more difficult (see, e.g., L.C. Evans “Partial Differential Equations” Section 5.5 for details). □

This implies in particular the characterization $H_0^1(D) = \text{Ker}(\gamma)$, meaning that elements in $H_0^1(D)$ are precisely those elements in $H^1(D)$ with zero trace.

Definition

Let $\|\cdot\|$ and $\|\cdot\|_*$ be two norms in a normed space X . The norms are called *equivalent* if there exist constants $c_1, c_2 > 0$ such that

$$c_1\|x\|_* \leq \|x\| \leq c_2\|x\|_* \quad \text{for all } x \in X.$$

The significance behind this notion lies in the fact that equivalent norms induce the same topology on X . That is, $\|\cdot\|$ and $\|\cdot\|_*$ induce *exactly the same* convergent sequences in X .

For our purposes, let $A: X \rightarrow Y$ be a mapping between two normed spaces. Suppose that $c_X\|\cdot\|_{x,*} \leq \|\cdot\|_x \leq C_X\|\cdot\|_{x,*}$ and $c_Y\|\cdot\|_{Y,*} \leq \|\cdot\|_Y \leq C_Y\|\cdot\|_{Y,*}$ for $c_X, C_X, c_Y, C_Y > 0$. If

$$\|A(x)\|_Y \leq K\|x\|_x \quad \text{for some } x \in X,$$

then

$$\|A(x)\|_{Y,*} \leq \frac{C_X K}{c_Y} \|x\|_{x,*} \quad \text{for some } x \in X.$$

We can change between equivalent norms rather liberally since any results about boundedness, stability, etc., proved using one norm remain true for equivalent norms up to a trivial scaling of the (typically generic) coefficient.

Proposition (Poincaré inequality)

Let $D \subset \mathbb{R}^d$ be a bounded domain. Then there exists $C > 0$ (independently of u) such that

$$\|u\|_{L^2(D)} \leq C \|\nabla u\|_{L^2(D)} \quad \text{for all } u \in H_0^1(D).$$

Proof. Let $\varphi \in C_0^\infty(D)$. Since we assumed D is bounded, we may assume $D \subset [-a, a]^d$ for suitably large $a > 0$. Extending φ by zero outside of D , we obtain

$$\begin{aligned}\varphi(x_1, x_2, \dots, x_d) &= \varphi(x_1, x_2, \dots, x_d) - \varphi(-a, x_2, \dots, x_d) \\ &= \int_{-a}^{x_1} \frac{\partial \varphi}{\partial x_1}(t_1, x_2, \dots, x_d) dt_1.\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}|\varphi(x_1, x_2, \dots, x_d)|^2 &\leq 2a \int_{-a}^a \left| \frac{\partial \varphi}{\partial x_1}(t_1, x_2, \dots, x_d) \right|^2 dt_1 \\ \Rightarrow \quad \int_{-a}^a |\varphi(x_1, x_2, \dots, x_d)|^2 dx_1 &\leq 4a^2 \int_{-a}^a \left| \frac{\partial \varphi}{\partial x_1}(t_1, x_2, \dots, x_d) \right|^2 dt_1.\end{aligned}$$

Repeated integrations w.r.t. x_2, x_3, \dots, x_d over $[-a, a]$ together with the density of $C_0^\infty(D)$ in $H_0^1(D)$ prove the assertion. □

An equivalent norm in $H_0^1(D)$

For all $u \in H_0^1(D)$, the norm

$$\|u\|_{H_0^1(D)} := \|\nabla u\|_{L^2(D)} := \left(\int_D \|\nabla u(\mathbf{x})\|^2 d\mathbf{x} \right)^{1/2}$$

is equivalent to $\|u\|_{H^1(D)} := (\|u\|_{L^2(D)}^2 + \|\nabla u\|_{L^2(D)}^2)^{1/2}$.

This can be seen as an immediate consequence of the Poincaré inequality:

$$\|u\|_{H_0^1(D)}^2 \leq \|u\|_{L^2(D)}^2 + \|\nabla u\|_{L^2(D)}^2 \leq (1 + C^2) \|u\|_{H_0^1(D)}^2.$$

Sobolev inequality

We mention the following result.

Theorem

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain and $k > d/2$. Then

$$H^k(D) \subset C_B(D) := \{v \in C(D) \mid v \text{ is bounded}\}$$

and there is a constant $C > 0$ s.t.

$$\|u\|_{C_B(D)} := \sup_{x \in D} |u(x)| \leq C \|u\|_{H^1(D)} \quad \text{for all } u \in H^1(D).$$

Proof. Cf., e.g., Adams (1975) or Adams and Fournier (2003). □

- If $d = 1$, then $u \in H^1(D)$ has a continuous representative.
- If $d \in \{2, 3\}$, then $u \in H^2(D)$ has a continuous representative.

Elliptic PDEs

Let $D \subset \mathbb{R}^d$ be an open and bounded Lipschitz domain. We consider the problem

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) = f(\mathbf{x}), & \mathbf{x} \in D, \\ u|_{\partial D} = 0, \end{cases} \quad (9)$$

where $f: D \rightarrow \mathbb{R}$ is the *source* and $a: D \rightarrow \mathbb{R}$ is the *diffusion coefficient*.

Uniform ellipticity assumption: There exist constants $a_{\max}, a_{\min} > 0$ such that

$$0 < a_{\min} \leq a(\mathbf{x}) \leq a_{\max} < \infty \quad \text{for all } \mathbf{x} \in D.$$

Definition

Let $a \in C^1(D)$ and $f \in C(D)$. Then $u \in C^2(D)$ is the *classical solution* to (9) if (9) holds for all $\mathbf{x} \in D$ and $u(\mathbf{y}) = 0$, $\mathbf{y} \in \partial D$.

The requirement that f is continuous is usually much too restrictive for practical applications.

Definition (Strong solution)

Let $a: \overline{D} \rightarrow \mathbb{R}$ be Lipschitz and $f \in L^2(D)$. We call $u \in H^2(D) \cap H_0^1(D)$ a *strong solution* to (9) if

$$-\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) = f(\mathbf{x}) \quad \text{for a.e. } \mathbf{x} \in D,$$

where the derivatives are the weak derivatives.

Note that we also have the following.

Lemma

Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then for $u, v \in H^1(D)$,

$$\int_D \partial_{x_j} u(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x} = - \int_D u(\mathbf{x}) \partial_{x_j} v(\mathbf{x}) \, d\mathbf{x} + \int_{\partial D} n_j u|_{\partial D} v|_{\partial D} \, dS,$$

where $\cdot|_{\partial D}: H^1(D) \rightarrow L^2(\partial D)$ is the trace operator.

Proof. The formula holds for $u, v \in C^\infty(\overline{D})$. The assertion follows by exploiting the density of $C^\infty(\overline{D})$ in $H^1(D)$. □

If u is a strong solution to the PDE (9), then for all $v \in C_0^\infty(D)$

$$\begin{aligned}\langle -\nabla \cdot (a \nabla u), v \rangle_{L^2(D)} &= \int_D -\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) v(\mathbf{x}) \, d\mathbf{x} \\ &\stackrel{\dagger}{=} \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} + \int_{\partial D} \underbrace{v(\mathbf{x})(a \nabla u(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}))}_{=0} \, dS \\ &= \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} =: B(u, v).\end{aligned}$$

Define also

$$F(v) := \int_D f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}.$$

This leads us to consider the variational formulation

$$B(u, v) = F(v) \quad \text{for all } v \in C_0^\infty(D).$$

^{\dagger} $\nabla \cdot (v(a \nabla u)) = a \nabla v \cdot \nabla u + v \nabla \cdot (a \nabla u) + \text{divergence theorem}$

The previous discussion motivates us to introduce the so-called *weak solution* to (9).

Definition

Let $a \in L^\infty(D)$ and $f \in L^2(D)$. Then $u \in H_0^1(D)$ is called a weak solution to (9) if

$$B(u, v) = F(v) \quad \text{for all } v \in H_0^1(D), \quad (10)$$

where

$$B(u, v) = \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x}$$

and

$$F(v) = \int_D f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}.$$

Remark. It is sufficient to enforce (10) for all $v \in C_0^\infty(D)$. Moreover, the definition can be extended for arbitrary $F \in (H_0^1(D))' =: H^{-1}(D)$.

Our variational problem is

$$B(u, v) = F(v) \quad \text{for all } v \in H_0^1(D), \quad (11)$$

where $B(u, v) = \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) d\mathbf{x}$ and $F(v) = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x}$.

Let us use the norm $\|v\|_{H_0^1(D)} := \|\nabla v\|_{L^2(D)}$, which is equivalent to the usual Sobolev norm by Poincaré's inequality.

Provided that we have uniform ellipticity, i.e.,
 $0 < a_{\min} \leq a(\mathbf{x}) \leq a_{\max} < \infty$ for all $\mathbf{x} \in D$, then

$$B(u, v) = \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} \leq a_{\max} \|u\|_{H_0^1(D)} \|v\|_{H_0^1(D)}$$

for all $u, v \in H_0^1(D)$ and

$$B(u, u) = \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla u(\mathbf{x}) d\mathbf{x} \geq a_{\min} \|u\|_{H_0^1(D)}^2 \quad \text{for all } u \in H_0^1(D).$$

∴ By the Lax–Milgram lemma, there exists a unique solution $u \in H_0^1(D)$ to (11) s.t. $\|u\|_{H_0^1(D)} \leq \frac{\|F\|_{H^{-1}(D)}}{a_{\min}}$.

When does the weak solution coincide with the strong solution?

If $f \in L^2(D)$, the diffusion coefficient a is smooth enough (e.g., Lipschitz), and the boundary ∂D is “nice enough” (e.g., a convex polyhedron), then $u \in H^2(D) \cap H_0^1(D)$ and the weak solution coincides with the strong solution. These considerations belong to the purview of *elliptic regularity theory*.

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Third lecture, April 28, 2025

Recap: Weak formulation

Let $D \subset \mathbb{R}^d$ be an open and bounded Lipschitz domain. We consider the problem

$$\begin{cases} -\nabla \cdot (a(x)\nabla u(x)) = f(x), & x \in D, \\ u|_{\partial D} = 0, \end{cases} \quad (1)$$

where $f: D \rightarrow \mathbb{R}$ is the *source* and $a: D \rightarrow \mathbb{R}$ is the *diffusion coefficient*.

Uniform ellipticity assumption: There exist constants $a_{\max}, a_{\min} > 0$ such that

$$0 < a_{\min} \leq a(x) \leq a_{\max} < \infty \quad \text{for all } x \in D.$$

Definition

Let $a \in L^\infty(D)$ and $f \in L^2(D)$. Then $u \in H_0^1(D)$ is called a weak solution to (1) if

$$B(u, v) = F(v) \quad \text{for all } v \in H_0^1(D), \quad (2)$$

where

$$B(u, v) = \int_D a(x)\nabla u(x) \cdot \nabla v(x) \, dx$$

and

$$F(v) = \int_D f(x)v(x) \, dx.$$

Galerkin method

1. Let $V_m = \text{span}\{\phi_i\}_{i=1}^m \subset H_0^1(D)$ be a finite-dimensional subspace.
2. Find $u_m \in V_m$ s.t.

$$B(u_m, \phi) = F(\phi) \quad \text{for all } \phi \in V_m. \quad (3)$$

Lemma

The problem (3) has a unique solution which also satisfies the so-called Galerkin orthogonality

$$B(u - u_m, \phi) = 0 \quad \text{for all } \phi \in V_m,$$

where u is the solution to (2).

Proof. The existence of a unique solution is an immediate consequence of the Lax–Milgram lemma applied to (3) in a subspace $V_m \subset H_0^1(D)$. The orthogonality follows from

$$B(u - u_m, \phi) = B(u, \phi) - B(u_m, \phi) = F(\phi) - F(\phi) = 0 \quad \text{for all } \phi \in V_m. \quad \square$$

Let $V_m := \text{span}\{\phi_i\}_{i=1}^m \subset H_0^1(D)$. Note that the problem of finding $u_m \in V_m$ such that

$$B(u_m, \phi) = F(\phi) \quad \text{for all } \phi \in V_m$$

is equivalent to

$$B(u_m, \phi_j) = F(\phi_j) \quad \text{for all } j \in \{1, \dots, m\}.$$

Since $u_m \in V_m$, we can write it as $u_m(\mathbf{x}) = \sum_{i=1}^m c_i \phi_i(\mathbf{x})$ using *undetermined coefficients* $\mathbf{c} = (c_i)_{i=1}^m \subset \mathbb{R}$. Thus the problem of finding $u_m \in V_m$ is equivalent to solving the coefficients \mathbf{c} satisfying

$$\sum_{i=1}^m c_i B(\phi_i, \phi_j) = F(\phi_j) \quad \text{for all } j \in \{1, \dots, m\},$$

which can be expressed as a linear system

$$\mathbf{A}\mathbf{c} = \mathbf{F},$$

where $\mathbf{A} = (A_{i,j})_{i,j=1}^m$ and $\mathbf{F} = (F_i)_{i=1}^m$ are such that

$$A_{i,j} = B(\phi_i, \phi_j) = \int_D a(\mathbf{x}) \nabla \phi_i(\mathbf{x}) \cdot \nabla \phi_j(\mathbf{x}) \, d\mathbf{x}, \quad F_i = \int_D f(\mathbf{x}) \phi_i(\mathbf{x}) \, d\mathbf{x}.$$

The Galerkin solution is a “quasi-optimal” approximation of the weak solution of the PDE in V_m .

Lemma (Céa's lemma)

Let $u \in H_0^1(D)$ be the solution to $B(u, \phi) = F(\phi)$ for all $\phi \in H_0^1(D)$ and let $u_m \in V_m$ be the solution to $B(u_m, \phi) = F(\phi)$ for all $\phi \in V_m$. Then

$$\|u - u_m\|_{H_0^1(D)} \leq \frac{a_{\max}}{a_{\min}} \inf_{v \in V_m} \|u - v\|_{H_0^1(D)}.$$

Proof. Let $v \in V_m$. Then by the coercivity and continuity of B , there holds

$$\begin{aligned} a_{\min} \|u - u_m\|_{H_0^1(D)}^2 &\leq B(u - u_m, u - u_m) \\ &= B(u - u_m, u - v) + B(u - u_m, \underbrace{v - u_m}_{\in V_m}) \\ &\quad \underbrace{}_{=0} \\ &\leq a_{\max} \|u - u_m\|_{H_0^1(D)} \|u - v\|_{H_0^1(D)}. \end{aligned}$$

Hence $a_{\min} \|u - u_m\|_{H_0^1(D)} \leq a_{\max} \|u - v\|_{H_0^1(D)}$. □

Finite element method

One could choose the space $V_m \subset H_0^1(D)$ to be virtually anything. The finite element method is a particular way of constructing this finite dimensional space.

In 2D, we approximate the geometry D by constructing a triangulation.

That is, the computational domain D is represented as the union of non-overlapping triangles called *elements*. The elements are assumed to cover the whole D (and only D). In 2D the elements are typically triangles or quadrilaterals, but they could be practically of any shape. In 3D the elements are typically tetrahedra or hexahedra. Prisms and pyramids are also widely used.

If the domain D is a polyhedron, then the division to elements is accurate. If the domain has, e.g., curved edges, then it cannot be approximated accurately with linear elements. This introduces additional error to the numerical approximation.

A single element is denoted by K . The collection of elements is called a mesh and denoted with \mathcal{T}_h , indexed by the diameter of the maximum element in the mesh. The size of the elements plays a key role in the convergence analysis of the method. For a well-defined method, reducing the size of the elements, i.e., refining the mesh, improves the solution (or at least does not make it worse).

The mesh is a discretization of the domain. It does not define a function space. To define the global space, we define the local space in each of the elements. The global space is a piecewise combination of the local, elementwise spaces.

Assume that the domain D is a 2D polyhedral domain, e.g., unit square, and that it has been divided into triangles. The simplest possible subspace to $H_0^1(D)$ is a piecewise linear, continuous space

$$V_h := \{v \in H_0^1(D) \mid v \in \mathcal{P}^1(K) \ \forall K \in \mathcal{T}_h\}.$$

The continuity is enforced by our requirement that the functions belong to $H^1(D)$.

Let \mathcal{T}_h be a triangulation of the domain D with FE nodes $(\mathbf{n}_i)_{i=1}^N$, where $m < N$ nodes are in the interior of the domain and $N - m$ nodes are on the boundary ∂D . For later convenience, let us denote
 $\text{interior} := \{i \in \{1, \dots, N\} \mid \mathbf{n}_i \notin \partial D\}$.

We can choose piecewise linear basis functions $\phi_i = \phi_{\mathbf{n}_i}$ such that

$$\phi_{\mathbf{n}_i}(\mathbf{n}_j) = \delta_{i,j},$$

that is, $\phi_{\mathbf{n}_i}(\mathbf{n}_i) = 1$ and $\phi_{\mathbf{n}_i}(\mathbf{n}_j) = 0$ whenever $i \neq j$ over the FE nodes $(\mathbf{n}_i)_{i=1}^N$.

Since V_h is finite-dimensional, it is spanned by a set of global basis functions

$$V_h = \text{span}\{\phi_{\mathbf{n}_i}\}_{i \in \text{interior}}.$$

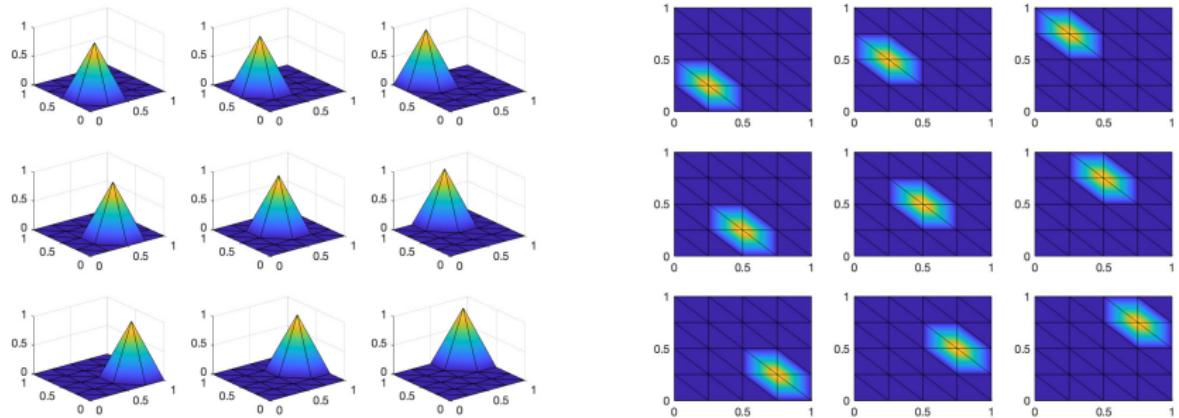


Figure: Left: An illustration of global, piecewise linear FE basis functions spanning V_h over a regular, uniform triangulation of $(0, 1)^2$. Right: Bird's-eye view of the same global FE basis functions.

The goal is to find the FE solution $u_h \in V_h$ such that

$$\int_D a(\mathbf{x}) \nabla u_h(\mathbf{x}) \cdot \nabla \phi(\mathbf{x}) \, d\mathbf{x} = \int_D f(\mathbf{x}) \phi(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } \phi \in V_h.$$

For simplicity, suppose that $f(\mathbf{x}) := \sum_{i=1}^N f_i \phi_{\mathbf{n}_i}(\mathbf{x})$ for some known coefficients $\mathbf{f} := (f_i)_{i=1}^N \subset \mathbb{R}$.[†] We can write $u_h = \sum_{i=1}^N c_i \phi_{\mathbf{n}_i} \in V_h$ and enforce the zero Dirichlet boundary condition by setting $c_i = 0$ for any $\mathbf{n}_i \in \partial D$. Testing the variational formulation against the FE basis functions $\phi = \phi_j$ for all $j \in \text{interior}$:

$$\sum_{i \in \text{interior}} c_i \underbrace{\int_D a(\mathbf{x}) \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) \, d\mathbf{x}}_{=: A_{i,j}} = \sum_{i=1}^N f_i \underbrace{\int_D \phi_{\mathbf{n}_i}(\mathbf{x}) \phi_{\mathbf{n}_j}(\mathbf{x}) \, d\mathbf{x}}_{=: M_{i,j}}.$$

Thus the problem is to solve the FE expansion coefficients

$\mathbf{c} = (c_i)_{i \in \text{interior}}$ from the equation

$$\mathbf{A}_{\text{interior}, \text{interior}} \mathbf{c} = \mathbf{M}_{\text{interior}, :} \mathbf{f},$$

where the matrix $\mathbf{A} = (A_{i,j})_{i,j=1}^N$ is called the *stiffness matrix* and $\mathbf{M} = (M_{i,j})_{i,j=1}^N$ is called the *mass matrix*.

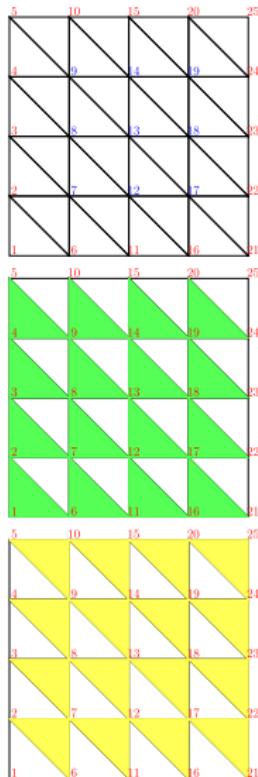
[†]Note that here we do not require $f = 0$ on ∂D ! For general $f \in L^2(D)$, instead of the mass matrix, one would use (Gaussian) quadratures to form the right-hand side.

Our goals in finite element programming:

- Construct a data structure to represent the topology of the finite element mesh.
 - If the FE nodes are given as rows of an array “nodes”, then the elements are triangles with vertices $\text{nodes}[i, :], \text{nodes}[j, :], \text{nodes}[k, :]$ for certain indices i, j, k .
 - We can represent the elements as an array “element”, where each row contains the *indices* corresponding to the nodes which form a triangle in our mesh.
 - Since we focus on homogeneous zero Dirichlet boundary conditions, we can enforce the boundary condition by setting the FE expansion coefficients of the FE solution to be zero at the boundary nodes. This is equivalent to choosing a subspace V_h consisting only of those FE basis functions ϕ_{n_i} corresponding to FE nodes in the interior of the mesh, i.e., $n_i \notin \partial D$. Thus it is helpful to store the indices of the nodes lying in the interior of the domain into a vector called “interior”.
- Assembly of finite element matrices (stiffness and mass matrix).
 - All triangles in our FE mesh can be mapped affinely onto a reference triangle of our choosing (say, $\{(x, y) \in \mathbb{R}^d \mid 0 \leq x \leq 1, 0 \leq y \leq 1 - x\}$) which we can exploit in the construction of the FE matrices.

Finite element programming (in Python)

Triangulation of $D = (0, 1)^2$



```
import numpy as np

def generateFEmesh(level):
    # Create a regular uniform triangulation
    # of the unit square (0,1)**2
    n1 = 2**level+1 # number of nodes in 1D
    # Topology: FE nodes, mesh elements, interior, centers
    X,Y = np.meshgrid(np.arange(0,n1)/(n1-1),
                      np.arange(0,n1)/(n1-1))
    nodes = np.array([X.flatten(),Y.flatten()]).T
    element = [] ; interior = []
    for i in range(0,n1-1):
        for j in range(0,n1-1):
            element.append([j*n1+i,(j+1)*n1+i,j*n1+i+1])
            element.append([(j+1)*n1+i,(j+1)*n1+i+1,j*n1+i+1])
            if i < n1-2 and j < n1-2:
                interior.append((j+1)*n1+i+1)
    centers = np.mean(nodes[element[:]],axis=1)
    return nodes,element,interior,centers
```

Mass matrix

Let $(K_\ell)_{\ell=1}^{nelem}$ be non-overlapping mesh elements s.t. $D = \bigcup_{\ell=1}^{nelem} K_\ell$. Let us first consider constructing the global mass matrix:

$$M_{i,j} = \int_D \phi_{n_i}(\mathbf{x}) \phi_{n_j}(\mathbf{x}) d\mathbf{x} = \sum_{\ell=1}^{nelem} \int_{K_\ell} \phi_{n_i}(\mathbf{x}) \phi_{n_j}(\mathbf{x}) d\mathbf{x}.$$

We can think of the elements of the global mass matrix as a sum of locally defined mass matrices in each “active” element. Recall that we already gave a labeling to the FE nodes earlier, and each row of matrix element contains the indices of FE nodes which form an element in our FE mesh.

```
initialize Mi,j = 0, i, j ∈ {0, …, ncoord – 1}  
for k ∈ {0, …, nelem – 1}, do  
    1. set ind = element[k]  
    2. let K be the element with vertices nodes[ind]  
    3. compute local mass matrix L ∈ ℝ3×3, where  
        Li,j = ∫K φni(x) φnj(x) dx, i, j ∈ {0, 1, 2}  
    4. set Mind,ind = Mind,ind + L  
end for
```

Let us concentrate on step 3.

Local mass matrix

- Let $K \subset \mathbb{R}^2$ be an arbitrary triangle with vertices $\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3$, and $\hat{K} = \{(x_1, x_2) \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}$ the reference triangle.
- Let $\hat{\phi}_1(\mathbf{x}) = 1 - x_1 - x_2$, $\hat{\phi}_2(\mathbf{x}) = x_1$, $\hat{\phi}_3(\mathbf{x}) = x_2$ be the local basis.
- The affine mapping $F_K: \hat{K} \rightarrow K$, $F_K(\mathbf{x}) := B\mathbf{x} + \mathbf{n}_1$, $B = [\mathbf{n}_2 - \mathbf{n}_1, \mathbf{n}_3 - \mathbf{n}_1]$, can be used to write the global basis functions as $\phi_{\mathbf{n}_i}(\mathbf{x}) = \hat{\phi}_i(F_K^{-1}(\mathbf{x}))$. Change of variables:

$$\int_K \phi_{\mathbf{n}_i}(\mathbf{x}) \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x} = |\det B| \int_{\hat{K}} \hat{\phi}_i(\mathbf{x}) \hat{\phi}_j(\mathbf{x}) d\mathbf{x} = \begin{cases} \frac{|\det B|}{12} & \text{if } i = j, \\ \frac{|\det B|}{24} & \text{if } i \neq j \end{cases}$$

that is

$$\left(\int_K \phi_{\mathbf{n}_i}(\mathbf{x}) \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x} \right)_{i,j=1}^3 = |\det B| \begin{pmatrix} \frac{1}{12} & \frac{1}{24} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{12} & \frac{1}{24} \\ \frac{1}{24} & \frac{1}{24} & \frac{1}{12} \end{pmatrix}.$$

Stiffness matrix

We also need to construct the global stiffness matrix

$$A_{i,j} = \int_D a(\mathbf{x}) \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x}.$$

To simplify the analysis, let us suppose that the diffusion coefficient $a(\mathbf{x})$ has been discretized as a piecewise constant function over the mesh elements, i.e.,

$$a(\mathbf{x}) = \sum_{\ell=1}^{\text{nelem}} a_\ell \chi_{K_\ell}(\mathbf{x}), \quad \chi_{K_\ell}(\mathbf{x}) := \begin{cases} 1 & \text{if } \mathbf{x} \in K_\ell, \\ 0 & \text{otherwise.} \end{cases}$$

Here, we can take a_ℓ to be the value of a evaluated at the center point of element K_ℓ . Then

$$A_{i,j} = \sum_{\ell=1}^{\text{nelem}} a_\ell \int_{K_\ell} \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x}.$$

Idea: Precompute the stiffness tensor $S_{i,j,\ell} := \int_{K_\ell} \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x}$. Given \mathbf{a} , the stiffness matrix is a tensor-vector contraction $A = S \times_3 \mathbf{a}$, where \mathbf{a} is a vector containing values of a at element center points.

The *idealized* construction of the stiffness tensor is as follows:

```
initialize Si,j,k = 0, i, j ∈ {0, …, ncoord – 1}, k ∈ {0, …, nelem – 1}
for k ∈ {0, …, nelem – 1}, do
    1. set ind = element[k]
    2. let K be the element with vertices nodes[ind]
    3. compute local stiffness matrix L ∈ ℝ3×3, where
         $L_{i,j} = \int_K \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x}$ , i, j ∈ {0, 1, 2}
    4. set Sind,ind,k = L
end for
```

Problem: Scipy does not support sparse tensors! :(

Workaround: reshape the $n \times n \times m$ tensor into an $n^2 \times m$ matrix!

```
initialize gradi,j,k = 0 for i, j ∈ {0, …, ncoord * ncoord – 1},
k ∈ {0, …, nelem – 1}
for k ∈ {0, …, nelem – 1}, do
    1. set ind = element[k]
    2. let K be the element with the vertices nodes[ind]
    3. compute local stiffness matrix L ∈ ℝ3×3, where
         $L_{i,j} = \int_K \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x}$ , i, j ∈ {0, 1, 2}
    4. initialize dummy = 0ncoord,ncoord; set dummyind,ind = L
    5. set grad[:, k] = dummy.flatten()
end for
```

Local stiffness matrix

- Let $K \subset \mathbb{R}^2$ be an arbitrary triangle with vertices $\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3$, and $\hat{K} = \{(x_1, x_2) \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}$ the reference triangle.
- Let $\hat{\phi}_1(\mathbf{x}) = 1 - x_1 - x_2, \hat{\phi}_2(\mathbf{x}) = x_1, \hat{\phi}_3(\mathbf{x}) = x_2$ be the local basis.
- The affine mapping $F_K: \hat{K} \rightarrow K, F_K(\mathbf{x}) := B\mathbf{x} + \mathbf{n}_1$,
 $B = [\mathbf{n}_2 - \mathbf{n}_1, \mathbf{n}_3 - \mathbf{n}_1]$, can be used to write the global basis
functions as $\phi_{\mathbf{n}_i}(\mathbf{x}) = \hat{\phi}_i(F_K^{-1}(\mathbf{x}))$. Note that there holds
 $\nabla \phi_{\mathbf{n}_i}(\mathbf{x}) = B^{-T}(\nabla \hat{\phi}_i)(F_K^{-1}(\mathbf{x}))$. Change of variables:

$$\int_K \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x} = |\det B| \int_{\hat{K}} B^{-T} \nabla \hat{\phi}_i(\mathbf{x}) \cdot B^{-T} \nabla \hat{\phi}_j(\mathbf{x}) d\mathbf{x}.$$

Define $G^T := (\nabla \hat{\phi}_1, \nabla \hat{\phi}_2, \nabla \hat{\phi}_3) = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$. Then

$$\left(\int_K \nabla \phi_{\mathbf{n}_i}(\mathbf{x}) \cdot \nabla \phi_{\mathbf{n}_j}(\mathbf{x}) d\mathbf{x} \right)_{i,j=1}^3 = \frac{|\det B|}{2} GB^{-1}B^{-T}G^T. \quad (4)$$

Remark. With a bit of linear algebra, one can check that (4) is equal to $\frac{D^T D}{4\text{area}(K)}$, $D := [\mathbf{n}_3 - \mathbf{n}_2, \mathbf{n}_1 - \mathbf{n}_3, \mathbf{n}_2 - \mathbf{n}_1]$.

The shoelace formula

```
def shoelace(g):  
    # Compute the area of a triangle with  
    # vertices g[0], g[1], and g[2]  
    return abs(np.linalg.det([g[0],g[1]])  
              + np.linalg.det([g[1],g[2]])  
              + np.linalg.det([g[2],g[0]]))/2
```

Assembly of the finite element matrices

```
from scipy import sparse

def generateFEmatrices(nodes,element):
    ncoord = len(nodes); nelem = len(element)
    mass_data = []; mass_rows = []; mass_cols = []
    grad_data = []; grad_rows = []; grad_cols = []
    localmass = np.array([[1/12,1/24,1/24],[1/24,1/12,1/24],[1/24,1/24,1/12]])
    for k in range(nelem):
        ind = element[k]; g = nodes[ind]
        detB = abs(np.linalg.det([g[1]-g[0],g[2]-g[0]]))
        Dt = np.array([g[2]-g[1],g[0]-g[2],g[1]-g[0]])
        triarea = shoelace(g)
        localgrad = Dt@Dt.T/4/triarea
        for i in range(3):
            for j in range(3):
                mass_rows.append(ind[i]); mass_cols.append(ind[j]);
                mass_data.append(detB*localmass[i,j])
                grad_rows.append(ind[i]*ncoord+ind[j]); grad_cols.append(k)
                grad_data.append(localgrad[i,j])
    mass = sparse.csr_matrix((mass_data,(mass_rows,mass_cols)),
                             shape=(ncoord,ncoord))
    grad = sparse.csr_matrix((grad_data,(grad_rows,grad_cols)),
                             shape=(ncoord*ncoord,nelem))
    return grad,mass
```

FEM programs in Python

```
level = 5 # discretization level
# Generate FE mesh
nodes,element,interior,centers = generateFEmesh(level)
ncoord = len(nodes) # number of coordinates
# Generate FE matrices
grad,mass = generateFEmatrices(nodes,element)
```

To obtain the stiffness matrix for a piecewise constant diffusion coefficient, we can use the following simple routine.

```
def UpdateStiffness(grad,a):
    # Given vector a containing the values of the diffusion
    # coefficient at the element center points, return the
    # corresponding stiffness matrix
    n = np.sqrt(grad.shape[0]).astype(int)
    vec = grad @ sparse.csr_matrix(a.reshape((a.size,1)))
    return sparse.csr_matrix.reshape(vec,(n,n)).tocsr()
```

Finite element method in 2D – summary

- ① Form a triangulation \mathcal{T}_h of the domain D . Let $(\mathbf{n}_j)_{j=1}^N$ be the finite element nodes. Form the list `interior` containing the indices of interior nodes and the element connectivity matrix `element`. Denote by $m = |\text{interior}|$ the number of degrees of freedom.
- ② Form the stiffness matrix $A \in \mathbb{R}^{m \times m}$ and mass matrix $M \in \mathbb{R}^{N \times N}$.
- ③ Form the loading vector $\mathbf{b} \approx M_{\text{interior},:} \mathbf{f}$, where $\mathbf{f} = [f(\mathbf{n}_1), \dots, f(\mathbf{n}_N)]^T$.
- ④ Solve $\mathbf{c} = (c_j)_{j=1}^m \in \mathbb{R}^m$ from $A\mathbf{c} = \mathbf{b}$.
- ⑤ The finite element solution is given by

$$u_h(\mathbf{x}) = \sum_{j=1}^m c_j \phi_j(\mathbf{x}), \quad \text{where } \phi_j = \phi_{\mathbf{n}_j}.$$

Remark: The global basis functions ϕ_j are typically *never constructed in practice!* Instead, note that $u_h(\mathbf{n}_j) = c_j$. Therefore, the nodal values of the FE solution are precisely the FE expansion coefficients – if one needs to evaluate the FE solution for $\mathbf{x} \in K$, one can use linear interpolation between the vertices of the triangle element K .

Computing norms of finite element solutions

Let $V_h \subset H_0^1(D)$ be a finite element space spanned by piecewise linear, continuous FE basis functions $\{\phi_i\}_{i=1}^m$ in the *interior of the domain*. Let

$$u_h(\mathbf{x}) = \sum_{i=1}^m c_i \phi_i(\mathbf{x}) \in V_h.$$

If $M_{i,j} = \int_D \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) d\mathbf{x}$ is the mass matrix and
 $S_{i,j} = \int_D \nabla \phi_i(\mathbf{x}) \cdot \nabla \phi_j(\mathbf{x}) d\mathbf{x}$ is the stiffness matrix of the Dirichlet Laplacian $-\Delta$ with homogeneous zero Dirichlet boundary conditions, then

$$\|u_h\|_{L^2(D)} = \sqrt{\mathbf{c}^T M \mathbf{c}} \quad \text{and} \quad \|u_h\|_{H_0^1(D)} = \sqrt{\mathbf{c}^T S \mathbf{c}},$$

where $\mathbf{c} = (c_i)_{i=1}^m$. These identities imply that M and S are positive definite.

Numerical example

Consider the elliptic PDE problem

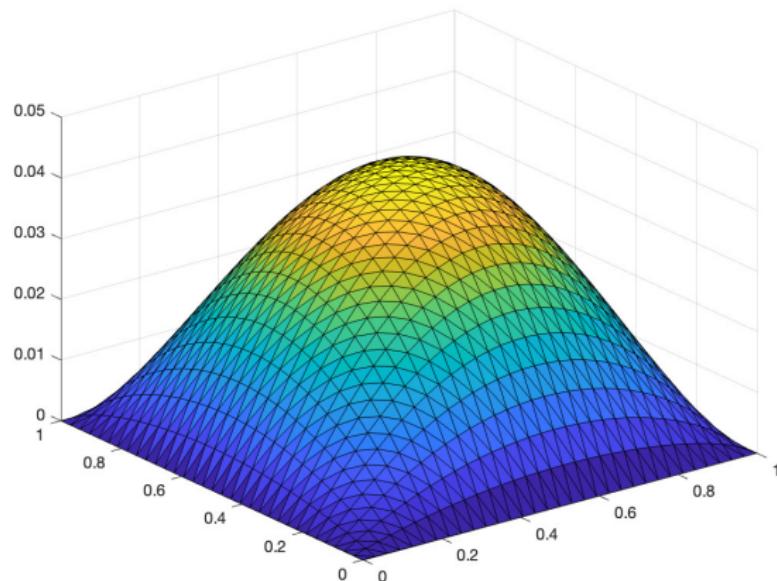
$$\begin{cases} -\nabla \cdot ((1+x^2+y^2)\nabla u(x,y)) = x+y, & (x,y) \in (0,1)^2, \\ u|_{\partial D} = 0. \end{cases}$$

We can solve this problem using the code developed above as follows.

```
level = 5 # discretization level
nodes,element,interior,centers = generateFEmesh(level) # generate FE mesh
ncoord = len(nodes) # number of coordinates
grad,mass = generateFEmatrices(nodes,element) # generate FE matrices
a = lambda x: 1+np.sum(x**2, axis=1) # diffusion coefficient
f = lambda x: np.sum(x, axis=1) # source term
rhs = mass[interior,:]*f(nodes) # precompute the loading vector
aval = a(centers) # evaluate diffusion coefficient at element centers
stiffness = UpdateStiffness(grad,aval) # assemble stiffness matrix
sol = np.zeros(ncoord) # initialize solution vector
# Solve the PDE
sol[interior] = sparse.linalg.spsolve(stiffness[np.ix_(interior,interior)],rhs)

# Visualize the solution
import matplotlib.pyplot as plt
fig = plt.figure(figsize=plt.figaspect(1.0))
ax = fig.add_subplot(1,1,1,projection='3d')
ax.plot_trisurf(nodes[:,0],nodes[:,1],sol,triangles=element,cmap=plt.cm.rainbow)
plt.show()
```

FE solution



This note illustrates possibly the simplest (nontrivial) implementation of conforming h -FEM. “Conforming” means that the FE space V_h is a proper subspace of the solution space $H_0^1(D)$. With this method, the only way to increase the approximation accuracy is by mesh refinement. One could generalize the method in a various number of ways:

- Using higher-order piecewise polynomial basis functions leads to p - and hp -FEM. The idea is to use higher-order polynomials and larger elements in regions of the computational domain where the PDE solution is smooth; conversely, one would use lower order polynomial basis functions and smaller elements near singularities (caused by obtuse angles in the geometry, etc.). A proper refinement strategy with hp -FEM can lead to exponentially convergent implementations.
- One can even use discontinuous basis functions, but the method becomes non-conforming. This has the benefit of improved parallelization and easy adaptation, but the implementation details are significantly more involved.
- Instead of discretizing the diffusion coefficient as a piecewise constant function over the elements, a better approach would be to compute the local stiffness matrices $\int_K a(x) \nabla \phi_{n_i}(x) \cdot \nabla \phi_{n_j}(x) dx$ using Gaussian quadratures for triangles. Similarly, the loading term $\int_K f(x) \phi_{n_i}(x) dx$ could also be computed using a Gaussian quadrature. For simplicity of presentation, the details are omitted.
- One could easily extend the method for more nontrivial boundary conditions: non-homogeneous Dirichlet, Neumann, Robin, mixed boundary conditions, etc. This results in additional “book-keeping” and the details are omitted.
- Many practitioners rely on automated mesh generation using software such as Netgen, etc. When the domain has curved boundaries, one usually either ignores the geometry modeling error (if there is reason to believe it is negligible) or uses, e.g., curved finite elements.
- Instead of using a direct solver like `scipy.sparse.linalg.spsolve` to solve the FE system, algebraic multigrid methods (and/or iterative solvers) can be used to improve the computational complexity. Nonlinear PDEs lead to nonlinear discretized FE systems.

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fourth lecture, May 5, 2025

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space. We consider the problem

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}) & \text{for } \mathbf{x} \in D, \text{ (a.e.) } \omega \in \Omega, \\ u(\mathbf{x}, \omega) = 0 & \text{for } \mathbf{x} \in \partial D, \text{ (a.e.) } \omega \in \Omega, \end{cases}$$

where the diffusion coefficient $a(\cdot, \omega)$ is *random*. In consequence, the solution $u(\cdot, \omega)$ is a random function/field.

In order to analyze $u(\cdot, \omega)$, some approaches might be:

- Monte Carlo methods → slow convergence rate.
- Sparse grid methods → good convergence, poor parallelization.

In certain problems (such as the PDE above) the dependence of u on a tends to be quite smooth (under moderate modeling assumptions).

Quasi-Monte Carlo methods take advantage of this fact and can be used to obtain faster-than-Monte Carlo convergence rates.

Probability measures

Let Ω be a set and let $\mathcal{P}(\Omega) := \{B \mid B \subseteq \Omega\}$ denote its power set. A subset \mathcal{F} of $\mathcal{P}(\Omega)$ is called σ -algebra (or σ -field) if

- ① $\emptyset \in \mathcal{F}$,
- ② $\Omega \setminus A \in \mathcal{F}$ for every $A \in \mathcal{F}$, and
- ③ $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}$ for every countable subset $\{A_n\}_{n \in \mathbb{N}}$ of \mathcal{F} .

A pair (Ω, \mathcal{F}) is called a *measurable space*.

An intuitive way of thinking about σ -algebras is that they contain information. The subsets contained in a σ -algebra represent events for which we can decide, after the observation, whether they happened or not. Hence, \mathcal{F} represents all the information we can get from an experiment. For a topological space Ω (e.g., \mathbb{R}^s), the smallest σ -algebra containing all open sets in Ω is called *Borel σ -algebra* on Ω and it is denoted by $\text{Bor}(\Omega)$.

A function $\mu: \mathcal{F} \rightarrow [0, \infty) \cup \{\infty\}$ is called *probability measure* if

- (i) $\mu(\emptyset) = 0$,
- (ii) for every countable subset $\{A_n\}_{n \in \mathbb{N}} \subset \mathcal{F}$ of pairwise disjoint sets (i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$),

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k),$$

- (iii) and $\mu(\Omega) = 1$.

We call $\mu(A)$ the *probability* of an event $A \in \mathcal{F}$. If $\mu(A) = 1$, we say that the event A occurs *almost surely*. A triple $(\Omega, \mathcal{F}, \mu)$ is called *probability space*. If only properties (i) and (ii) are satisfied, μ is called a *measure*. A measure is called σ -finite if Ω is the countable union of measurable sets with finite measure.

Example

The *Dirac measure* δ_m at a point $m \in \mathbb{R}^s$ is a probability measure on $(\mathbb{R}^s, \text{Bor}(\mathbb{R}^s))$ defined by

$$\delta_m(A) = \begin{cases} 1 & \text{if } m \in A, \\ 0 & \text{if } m \notin A \end{cases} \quad \text{for all } A \in \text{Bor}(\mathbb{R}^s).$$

Example

The Lebesgue measure λ on $(\mathbb{R}^s, \text{Bor}(\mathbb{R}^s))$ is σ -finite, but not a probability measure, since $\lambda(\mathbb{R}^s) = \infty$.

Let μ and ν be two measures on the same measure space. Then μ is said to be *absolutely continuous with respect to ν* (or *dominated by ν*) if $\nu(A) = 0$ implies $\mu(A) = 0$ for each $A \in \mathcal{F}$. We denote this by $\mu \ll \nu$. Measures μ and ν are called *equivalent* if $\mu \ll \nu$ and $\nu \ll \mu$. If μ and ν are supported on disjoint sets, they are called *mutually singular*.

Theorem (Radon–Nikodym)

Let μ and ν be two measures on a measure space (Ω, \mathcal{F}) . If $\mu \ll \nu$ and ν is σ -finite, then there exists a unique ν -integrable function f such that

$$\mu(A) = \int_A f(\omega) \nu(d\omega) \quad \text{for all } A \in \mathcal{F}.$$

The function f is called *Radon–Nikodym derivative* (or *density*) of μ with respect to ν and it is denoted by $\frac{d\mu}{d\nu}$.

Example

If μ is a measure which is absolutely continuous with respect to the Lebesgue measure λ on $(\mathbb{R}^s, \text{Bor}(\mathbb{R}^s))$, then it has a unique density $p \in L^1(\mathbb{R}^s)$ by the Radon–Nikodym theorem.

Example

Let $\mu_1 = \mathcal{U}([0, 1])$ and $\mu_2 = \mathcal{U}([0, 2])$ be uniform probability measures on \mathbb{R} . Then $\mu_1 \ll \mu_2$ with

$$\frac{d\mu_1}{d\mu_2}(t) = \begin{cases} 2 & \text{for } t \in [0, 1], \\ 0 & \text{otherwise,} \end{cases}$$

but μ_2 is not absolutely continuous with respect to μ_1 because $\mu_1([1, 2]) = 0$, whereas $\mu_2([1, 2]) = \frac{1}{2} > 0$.

Random variables

A function $x: \Omega \rightarrow X$ between a probability space $(\Omega, \mathcal{F}, \mu)$ and a measurable space (X, \mathcal{X}) is called a *random variable (with values in X)* if it is measurable, that is, if

$$x^{-1}(A) \in \mathcal{F} \quad \text{for every } A \in \mathcal{X}.$$

Here, $x^{-1}(A) = \{\omega \in \Omega : x(\omega) \in A\}$.

A random variable x induces a probability measure ν on X , defined by

$$\nu(A) := \mu(x^{-1}(A)) \quad \text{for all } A \in \mathcal{X},$$

which is called *probability distribution (or law)* of x . We write $x \sim \nu$ if x is distributed according to ν .

A random variable x connects an event $A \in \mathcal{X}$ with a corresponding event $x^{-1}(A) \in \mathcal{F}$ and assigns the probability of $x^{-1}(A)$ to A . This probability is denoted by

$$\mathbb{P}(x \in A) := \nu(A) = \mu(x^{-1}(A)) = \mu(\{\omega \in \Omega : x(\omega) \in A\}).$$

Now, let x be a random variable with values in $(\mathbb{R}^s, \text{Bor}(\mathbb{R}^s))$ and ν its distribution.

If ν is absolutely continuous with respect to the Lebesgue measure λ on \mathbb{R}^s , then by the Radon–Nikodym theorem there exists a unique $p \in L^1(\mathbb{R}^s)$ such that

$$\nu(A) = \int_A p(x)dx \quad \text{for all } A \in \text{Bor}(\mathbb{R}^s).$$

The function p is called *probability density* of x .

In what follows, we will assume that \mathbb{R}^s -valued random variables have a probability density.

Let \mathbf{x} , \mathbf{x}_1 , and \mathbf{x}_2 be \mathbb{R}^s -valued random variables.

- The *mean* or *expected value* of \mathbf{x} with distribution ν and probability density function p is given by

$$\mathbb{E}[\mathbf{x}] := \int_{\mathbb{R}^s} \mathbf{x} \nu(d\mathbf{x}) = \int_{\mathbb{R}^s} \mathbf{x} p(\mathbf{x}) d\mathbf{x}.$$

- A *mode* $\bar{\mathbf{x}}$ of a random variable \mathbf{x} is defined as a maximizer of its density p , i.e.,

$$\bar{\mathbf{x}} \in \arg \max_{\mathbf{x} \in \mathbb{R}^s} p(\mathbf{x}).$$

- The *covariance* (or *covariance matrix*) of two random variables \mathbf{x}_1 and \mathbf{x}_2 is defined by

$$\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{E} [(\mathbf{x}_1 - \mathbb{E}[\mathbf{x}_1])(\mathbf{x}_2 - \mathbb{E}[\mathbf{x}_2])^T].$$

- The *variance* of random variable \mathbf{x} is its covariance with itself:

$$\text{Var}(\mathbf{x}) = \text{Cov}(\mathbf{x}, \mathbf{x}).$$

- The *characteristic function* $\varphi_{\mathbf{x}}$ of \mathbf{x} is defined by

$$\varphi_{\mathbf{x}}(\mathbf{h}) = \int_{\mathbb{R}^s} \exp(i \mathbf{h}^T \mathbf{x}) \nu(d\mathbf{x}) = \int_{\mathbb{R}^s} \exp(i \mathbf{h}^T \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad \text{for all } \mathbf{h} \in \mathbb{R}^s.$$

A random variable is uniquely determined by its characteristic function.

Gaussian random variables

Let $\mathbf{m} \in \mathbb{R}^s$ and $C \in \mathbb{R}^{s \times s}$ be a symmetric positive semidefinite matrix.[†] An \mathbb{R}^s -valued random variable \mathbf{x} is said to be *Gaussian* (or *normal*) with mean \mathbf{m} and covariance C , denoted by $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, C)$, if its characteristic function $\varphi_{\mathbf{x}}$ is given by

$$\varphi_{\mathbf{x}}(\mathbf{h}) = \exp\left(i\mathbf{h}^T \mathbf{m} - \frac{1}{2}\mathbf{h}^T C \mathbf{h}\right) \quad \text{for all } \mathbf{h} \in \mathbb{R}^s.$$

A Gaussian random variable is completely determined by its mean and its covariance.

Remark: Multivariate Gaussian random variables also have the following characterization. A random vector $\mathbf{x} = (x_1, \dots, x_s)^T$ has a multivariate normal distribution iff $y = a_1 x_1 + \dots + a_s x_s$ is (univariate) normally distributed for all constants $a_1, \dots, a_s \in \mathbb{R}$.

[†]Recall that this means $\xi^T C \xi \geq 0$ for all $\xi \in \mathbb{R}^s$.

- If, in addition, C is positive definite[†], $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, C)$ has the probability density

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{(2\pi)^{s/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^T C^{-1}(\mathbf{x} - \mathbf{m})\right) \\ &= \frac{1}{(2\pi)^{s/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}\|C^{-\frac{1}{2}}(\mathbf{x} - \mathbf{m})\|^2\right). \end{aligned}$$

Note that C is invertible and $C^{-1/2}$ exists due to our assumptions on C .

- The Dirac measure $\delta_{\mathbf{m}}$ at a point $\mathbf{m} \in \mathbb{R}^s$ can be understood as a Gaussian distribution with covariance $C = 0$, i.e., $\delta_{\mathbf{m}} = \mathcal{N}(\mathbf{m}, \mathbf{0})$.
- If $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{m}_1, C_1)$ and $\mathbf{z}_2 \sim \mathcal{N}(\mathbf{m}_2, C_2)$ are independent and $a_1, a_2 \in \mathbb{R}$, then

$$\mathbf{z} = a_1 \mathbf{z}_1 + a_2 \mathbf{z}_2 \sim \mathcal{N}(a_1 \mathbf{m}_1 + a_2 \mathbf{m}_2, a_1^2 C_1 + a_2^2 C_2).$$

- If $\mathbf{z} \sim \mathcal{N}(\mathbf{m}, C)$, $L \in \mathbb{R}^{s \times k}$, and $\mathbf{a} \in \mathbb{R}^s$, then

$$\mathbf{w} = L\mathbf{z} + \mathbf{a} \sim \mathcal{N}(L\mathbf{m} + \mathbf{a}, LCL^T).$$

[†]Recall that this means $\xi^T C \xi > 0$ for all $\xi \in \mathbb{R}^s \setminus \{\mathbf{0}\}$.

Conditional and marginal probability densities

Let \mathbf{x} and \mathbf{y} be random variables with values in \mathbb{R}^s and \mathbb{R}^k , respectively. If the random variable (\mathbf{x}, \mathbf{y}) has a probability density $p_{\mathbf{x}, \mathbf{y}}$, i.e., if

$$\mathbb{P}(\mathbf{x} \in A, \mathbf{y} \in B) = \mathbb{P}((\mathbf{x}, \mathbf{y}) \in A \times B) = \int_{A \times B} p_{\mathbf{x}, \mathbf{y}}(\mathbf{u}, \mathbf{v}) d(\mathbf{u}, \mathbf{v}),$$

for all $A \in \text{Bor}(\mathbb{R}^s)$ and $B \in \text{Bor}(\mathbb{R}^k)$, then $p_{\mathbf{x}, \mathbf{y}}$ is called *joint probability density* of \mathbf{x} and \mathbf{y} . Here $\mathbb{P}(\mathbf{x} \in A, \mathbf{y} \in B) := \mathbb{P}(\mathbf{x} \in A \text{ and } \mathbf{y} \in B)$.

Now, the *marginal probability density* $p_{\mathbf{x}}$ of \mathbf{x} is defined by

$$p_{\mathbf{x}}(\mathbf{u}) = \int_{\mathbb{R}^k} p_{\mathbf{x}, \mathbf{y}}(\mathbf{u}, \mathbf{v}) d\mathbf{v} \quad \text{for all } \mathbf{u} \in \mathbb{R}^s.$$

Analogously, the marginal density of \mathbf{y} is

$$p_{\mathbf{y}}(\mathbf{v}) = \int_{\mathbb{R}^s} p_{\mathbf{x}, \mathbf{y}}(\mathbf{u}, \mathbf{v}) d\mathbf{u} \quad \text{for all } \mathbf{v} \in \mathbb{R}^k.$$

The marginal density of \mathbf{x} is indeed the probability density for \mathbf{x} in the situation that we have no information about the random variable \mathbf{y} , because

$$\begin{aligned}\mathbb{P}(\mathbf{x} \in A) &= \mathbb{P}(\mathbf{x} \in A, \mathbf{y} \in \mathbb{R}^k) = \int_{A \times \mathbb{R}^k} p_{\mathbf{x}, \mathbf{y}}(\mathbf{u}, \mathbf{v}) d(\mathbf{u}, \mathbf{v}) \\ &= \int_A \left(\int_{\mathbb{R}^k} p_{\mathbf{x}, \mathbf{y}}(\mathbf{u}, \mathbf{v}) d\mathbf{v} \right) d\mathbf{u} = \int_A p_{\mathbf{x}}(\mathbf{u}) d\mathbf{u}\end{aligned}$$

for every $A \in \text{Bor}(\mathbb{R}^s)$.

The random variables \mathbf{x} and \mathbf{y} are called *independent* if

$$\mathbb{P}(\mathbf{x} \in A, \mathbf{y} \in B) = \mathbb{P}(\mathbf{x} \in A)\mathbb{P}(\mathbf{y} \in B)$$

for all $A \in \text{Bor}(\mathbb{R}^s)$, $B \in \text{Bor}(\mathbb{R}^k)$ or, equivalently, if

$$p_{\mathbf{x}, \mathbf{y}}(\mathbf{u}, \mathbf{v}) = p_{\mathbf{x}}(\mathbf{u})p_{\mathbf{y}}(\mathbf{v}) \quad \text{almost surely.}$$

Next, we consider the random variable \mathbf{x} in the opposite situation that we know everything about the random variable \mathbf{y} : we have observed it and know what value it has taken.

We say we consider the random variable \mathbf{x} , given that we know the value \mathbf{y}_0 taken by \mathbf{y} , and denote this by $\mathbf{x}|\mathbf{y} = \mathbf{y}_0$. For $\mathbf{y}_0 \in \mathbb{R}^k$ with $p_{\mathbf{y}}(\mathbf{y}_0) > 0$, the *conditional probability density* of $\mathbf{x}|\mathbf{y} = \mathbf{y}_0$, $p_{\mathbf{x}|\mathbf{y}=\mathbf{y}_0}$, is then defined by

$$p_{\mathbf{x}|\mathbf{y}=\mathbf{y}_0}(\mathbf{u}) = \frac{p_{\mathbf{x},\mathbf{y}}(\mathbf{u}, \mathbf{y}_0)}{p_{\mathbf{y}}(\mathbf{y}_0)}.$$

If \mathbf{x} and \mathbf{y} are independent and $p_{\mathbf{y}}(\mathbf{y}_0) > 0$, then

$$p_{\mathbf{x}|\mathbf{y}=\mathbf{y}_0}(\mathbf{u}) = p_{\mathbf{x}}(\mathbf{u}).$$

Representation of random fields

Random field

Definition

Let $D \subset \mathbb{R}^d$ and let $(\Omega, \mathcal{F}, \mu)$ be a probability space. A function $A: D \times \Omega \rightarrow X$ is called a *random field* if $A(\mathbf{x}, \cdot)$ is an X -valued random variable for all $\mathbf{x} \in D$.

Definition

We call a random field $A: D \times \Omega \rightarrow X$ square-integrable if

$$\int_{\Omega} \int_D |A(\mathbf{x}, \omega)|^2 d\mathbf{x} \mu(d\omega) < \infty.$$

Our goal will be to model (infinite-dimensional) input random fields using finite-dimensional expansions with s variables.

Comment on notation: In what follows, s will always refer to the “stochastic dimension” (dimension of the stochastic/parametric space) while d will refer to the “spatial dimension” (dimension of the spatial Lipschitz domain $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$).

Remark: separable Hilbert space

A Hilbert space is said to be *separable* if (and only if) there exists a *countable orthonormal basis* $\{\psi_j\}_{j=1}^{\infty}$ of H with respect to the inner product $\langle \cdot, \cdot \rangle_H$, that is,

$$\langle \psi_j, \psi_k \rangle_H = \delta_{j,k} \quad \text{and} \quad \left\| f - \sum_{j=1}^{\ell} \langle f, \psi_j \rangle_H \psi_j \right\|_H \xrightarrow{\ell \rightarrow \infty} 0 \quad \text{for all } f \in H.$$

This last condition is often written as

$$f = \sum_{j=1}^{\infty} \langle f, \psi_j \rangle_H \psi_j.$$

Note that $\sum_{j=1}^{\ell} \langle f, \psi_j \rangle_H \psi_j$ is precisely the orthogonal projection onto the subspace spanned by $\psi_1, \dots, \psi_{\ell}$.

Mercer's theorem

Let $a(x, \omega)$ be a square-integrable random field with mean

$$\bar{a}(x) = \int_{\Omega} a(x, \omega) \mu(d\omega), \quad x \in D,$$

and a continuous, symmetric, positive definite[†] covariance

$$K(x, x') = \int_{\Omega} (a(x, \omega) - \bar{a}(x))(a(x', \omega) - \bar{a}(x')) \mu(d\omega).$$

Mercer's theorem: if $D \subset \mathbb{R}^d$ is a compact, measurable set with positive Lebesgue measure, then the covariance operator $\mathcal{C}: L^2(D) \rightarrow L^2(D)$,

$$(\mathcal{C}u)(x) = \int_D K(x, x') u(x') dx', \quad x \in D,$$

has a countable sequence of eigenvalues $\{\lambda_k\}_{k \geq 1}$ and corresponding eigenfunctions $\{\psi_k\}_{k \geq 1}$ satisfying $\mathcal{C}\psi_k = \lambda_k \psi_k$ such that $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ and $\lambda_k \rightarrow 0$ and the eigenfunctions form an orthonormal basis for $L^2(D)$.

Note that this means:

$$\int_D K(x, x') \psi_k(x') dx' = \lambda_k \psi_k(x), \quad \int_D \psi_k(x) \psi_\ell(x) dx = \delta_{k,\ell}.$$

[†]In this context, positive definite means: for all choices of finitely many points $x_1, \dots, x_k \in D$, $k \in \mathbb{N}$, the *Gram matrix* $G := [K(x_i, x_j)]_{i,j=1}^k$ is positive semidefinite.

The Karhunen–Loève (KL) expansion of a random field

Theorem

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, let $D \subset \mathbb{R}^d$ be a compact, measurable set with positive Lebesgue measure, and let $a: D \times \Omega \rightarrow \mathbb{R}$ be a square-integrable random field with continuous, symmetric, positive definite covariance $K(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(a(\mathbf{x}, \cdot) - \bar{a}(\mathbf{x}))(a(\mathbf{x}', \cdot) - \bar{a}(\mathbf{x}'))]$. Then the eigensystem $(\lambda_k, \psi_k) \in \mathbb{R}_+ \times L^2(D)$ of the covariance operator $\mathcal{C}: L^2(D) \rightarrow L^2(D)$, as described on the previous slide, can be used to write

$$a(\mathbf{x}, \omega) = \bar{a}(\mathbf{x}) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \psi_k(\mathbf{x}),$$

$$\text{where } \xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D (a(\mathbf{x}, \omega) - \bar{a}(\mathbf{x})) \psi_k(\mathbf{x}) d\mathbf{x},$$

where the convergence is in L^2 w.r.t. the stochastic parameter and uniform in \mathbf{x} . Furthermore, the random variables ξ_k are zero-mean uncorrelated random variables with unit variance, i.e.,

$$\mathbb{E}[\xi_k] = 0 \quad \text{and} \quad \mathbb{E}[\xi_k \xi_\ell] = \delta_{k,\ell}.$$

Proof. WLOG, we can assume that $\bar{a}(x) = 0$.[†] By Mercer's theorem, $\{\psi_k\}_{k=1}^{\infty}$ forms an orthonormal basis on $L^2(D)$ and we can write

$$K(x, x') = \sum_{k=1}^{\infty} \underbrace{\left(\int_D K(x, t) \psi_k(t) dt \right)}_{=\lambda_k \psi_k(x)} \psi_k(x') = \sum_{k=1}^{\infty} \lambda_k \psi_k(x) \psi_k(x').$$

Moreover, the random field a can be expressed using the same eigenbasis:

$$a(x, \omega) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \psi_k(x), \quad \xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D a(x, \omega) \psi_k(x) dx.$$

One easily computes that

$$\mathbb{E}[\xi_k] = \mathbb{E}\left[\frac{1}{\sqrt{\lambda_k}} \int_D a(x, \cdot) \psi_k(x) dx \right] = \frac{1}{\sqrt{\lambda_k}} \int_D \mathbb{E}[a(x, \cdot)] \psi_k(x) dx = 0$$

and

$$\begin{aligned} \mathbb{E}[\xi_k \xi_\ell] &= \mathbb{E}\left[\frac{1}{\lambda_k} \int_D \int_D a(x, \cdot) a(x', \cdot) \psi_k(x) \psi_\ell(x') dx dx' \right] \\ &= \frac{1}{\lambda_k} \int_D \int_D \mathbb{E}[a(x, \cdot) a(x', \cdot)] \psi_k(x) \psi_\ell(x') dx dx' \\ &= \frac{1}{\lambda_k} \int_D \int_D K(x, x') \psi_k(x) \psi_\ell(x') dx dx' = \frac{1}{\lambda_k} \int_D \psi_k(x) \underbrace{\left(\int_D K(x, x') \psi_\ell(x') dx' \right)}_{=\lambda_\ell \psi_\ell(x)} dx = \delta_{k,\ell}, \end{aligned}$$

since $\int_D \psi_k(x) \psi_\ell(x) dx = \delta_{k,\ell}$.

[†]Once the claim has been proved for a zero-mean random field $a(x, \omega)$, the general case follows simply by applying the theorem to $a(x, \omega) \leftarrow a(x, \omega) - \bar{a}(x)$.

Recall from the previous slide that

$$a(x, \omega) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \psi_k(x), \quad \xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D a(x, \omega) \psi_k(x) dx,$$

where $\mathbb{E}[\xi_k] = 0$, and $\mathbb{E}[\xi_k \xi_\ell] = \delta_{k,\ell}$. Let

$$a_s(x, \omega) = \sum_{k=1}^s \sqrt{\lambda_k} \xi_k(\omega) \psi_k(x).$$

$$\begin{aligned} \mathbb{E}[|a(x, \cdot) - a_s(x, \cdot)|^2] &= \mathbb{E}[a(x, \cdot)^2] + \mathbb{E}[a_s(x, \cdot)^2] - 2\mathbb{E}[a(x, \cdot)a_s(x, \cdot)] \\ &= K(x, x) + \mathbb{E}\left[\sum_{k=1}^s \sum_{\ell=1}^s \sqrt{\lambda_k \lambda_\ell} \xi_k(\cdot) \xi_\ell(\cdot) \psi_k(x) \psi_\ell(x)\right] \\ &\quad - 2\mathbb{E}\left[\left(\sum_{\ell=1}^{\infty} \sqrt{\lambda_\ell} \xi_\ell(\cdot) \psi_\ell(x)\right) \left(\sum_{k=1}^s \sqrt{\lambda_k} \xi_k(\cdot) \psi_k(x)\right)\right] \\ &= K(x, x) + \sum_{k=1}^s \sum_{\ell=1}^s \sqrt{\lambda_k \lambda_\ell} \mathbb{E}[\xi_k \xi_\ell] \psi_k(x) \psi_\ell(x) - 2\mathbb{E}\left[\sum_{\ell=1}^{\infty} \sum_{k=1}^s \sqrt{\lambda_k \lambda_\ell} \xi_\ell(\cdot) \xi_k(\cdot) \psi_\ell(x) \psi_k(x)\right] \\ &= K(x, x) + \sum_{k=1}^s \sum_{\ell=1}^s \sqrt{\lambda_k \lambda_\ell} \delta_{k,\ell} \psi_k(x) \psi_\ell(x) - 2 \sum_{\ell=1}^{\infty} \sum_{k=1}^s \sqrt{\lambda_k \lambda_\ell} \mathbb{E}[\xi_\ell \xi_k] \psi_\ell(x) \psi_k(x) \\ &= K(x, x) + \sum_{\ell=1}^s \lambda_\ell \psi_\ell(x)^2 - 2 \sum_{\ell=1}^{\infty} \sum_{k=1}^s \sqrt{\lambda_k \lambda_\ell} \mathbb{E}[\xi_\ell \xi_k] \psi_\ell(x) \psi_k(x) \quad (\mathbb{E}[\xi_\ell \xi_k] = \delta_{\ell,k}) \\ &= K(x, x) - \sum_{\ell=1}^s \lambda_\ell \psi_\ell(x)^2 \rightarrow 0 \quad \text{in the } L^2 \text{ sense by Mercer's theorem.} \quad \square \end{aligned}$$

The Karhunen–Loève (KL) expansion of random field $a(\mathbf{x}, \omega)$ can be written as

$$a(\mathbf{x}, \omega) = \bar{a}(\mathbf{x}) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \psi_k(\mathbf{x}).$$

Remarks:

- The KL expansion minimizes the mean-square truncation error:

$$\left\| a(\mathbf{x}, \omega) - \bar{a}(\mathbf{x}) - \sum_{k=1}^s \sqrt{\lambda_k} \xi_k(\omega) \psi_k(\mathbf{x}) \right\|_{L^2(\Omega, \mu; L^2(D))} = \left(\sum_{k=s+1}^{\infty} \lambda_k \right)^{1/2}.$$

- The random variables ξ_k are centered and uncorrelated, but not necessarily independent.
- If the random field $a(\mathbf{x}, \omega)$ is Gaussian – by definition, this means that $(a(\mathbf{x}_1, \omega), \dots, a(\mathbf{x}_k, \omega))$ is a multivariate Gaussian random variable for all $\mathbf{x}_1, \dots, \mathbf{x}_k \in D$, $k \in \mathbb{N}$ – then the random variables ξ_k are independent.

The utility of the KL expansion comes from the fact that it is an effective method of representing *input* random fields when their covariance structure is known.

Essentially, if the (infinite-dimensional) input random field has a known covariance (which satisfies the conditions of Mercer's theorem), then we can use the KL expansion to find a finite-dimensional approximation, which is optimal in the mean-square error sense.

Example

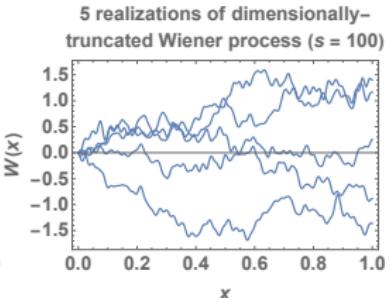
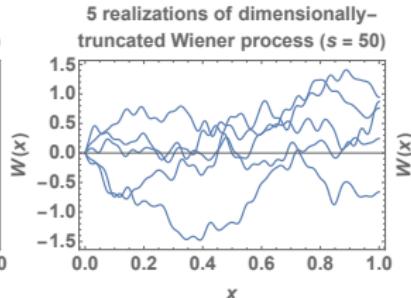
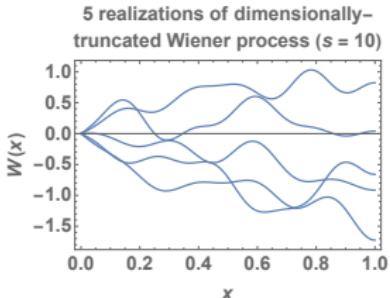
Let us consider the Wiener process over $D = [0, 1]$, which we regard as a centered standard Gaussian random field $W(x, \omega)$ with covariance function $K(x, y) = \min\{x, y\}$, $x, y \in [0, 1]$. It can be shown that

$$\int_0^1 K(x, y) \psi_k(y) dy = \lambda_k \psi_k(x),$$

where $\psi_k(x) = \sqrt{2} \sin((k - \frac{1}{2})\pi x)$, $\lambda_k = \frac{1}{(k - \frac{1}{2})^2 \pi^2}$. Then it has the KL expansion

$$W(x, \omega) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} y_k(\omega) \psi_k(x), \quad y_k \sim \mathcal{N}(0, 1).$$

Let us plot some realizations with the series truncated to $s \in \{10, 50, 100\}$ terms.



Modeling assumptions

In engineering and practical applications, the idea is that we have some *a priori* knowledge/belief that the unknown input random field is distributed according to some known probability distribution with a known covariance.

- If the input random field is Gaussian with a known, nice covariance function[†], then we use the KL expansion to find a reasonable finite-dimensional approximation of true input. Since the KL expansion decorrelates the stochastic variables, and uncorrelated jointly Gaussian random variables are independent, we can exploit the independence of the stochastic variables to parameterize the model problem.
- If the input random field is *not Gaussian*, then the stochastic variables in the KL expansion are uncorrelated *but not necessarily independent*. For the purposes of mathematical analysis, we typically assume that the random variables in the input random field are independent so that we can parameterize the model problem. (Transforming dependent random variables into independent random variables can be done using, e.g., the Rosenblatt transformation, but this is computationally expensive.)

Note especially that in the Gaussian setting we do not need to make any “extra” effort to ensure the independence of the stochastic variables in the KL expansion.

[†]Matérn covariance is an especially popular choice.

Example (Lognormal input random field)

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a Lipschitz domain and consider the PDE problem

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}) & \text{for } \mathbf{x} \in D, \\ u(\cdot, \omega)|_{\partial D} = 0, \end{cases}$$

where $f: D \rightarrow \mathbb{R}$ is a fixed (deterministic) source term. We can model a lognormally distributed random diffusion coefficient $a: D \times \Omega \rightarrow \mathbb{R}$ using the KL expansion, e.g., as

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) \exp \left(\sum_{k=1}^{\infty} y_k(\omega) \psi_k(\mathbf{x}) \right), \quad y_k \sim \mathcal{N}(0, 1),$$

where $a_0 \in L^\infty(D)$ is such that $a_0(\mathbf{x}) > 0$ and the random variables y_k are uncorrelated (and thus independent in the Gaussian case).

Due to the independence, we can consider the above as a *parametric PDE* with $a(\mathbf{x}, \mathbf{y}) \equiv a(\mathbf{x}, \mathbf{y}(\omega))$ and $u(\mathbf{x}, \mathbf{y}) \equiv u(\mathbf{x}, \mathbf{y}(\omega))$, where (formally) $\mathbf{y} \in \mathbb{R}^N$ is a *parametric vector* endowed with a product Gaussian measure.

Example (Uniform and affine input random field)

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a Lipschitz domain, $f: D \rightarrow \mathbb{R}$ is a fixed (deterministic) source term, and consider the PDE problem

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}) & \text{for } \mathbf{x} \in D, \\ u(\cdot, \omega)|_{\partial D} = 0. \end{cases}$$

We can model a uniformly distributed random diffusion coefficient $a: D \times \Omega \rightarrow \mathbb{R}$ using the KL expansion, e.g., as

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \sum_{k=1}^{\infty} y_k(\omega) \psi_k(\mathbf{x}), \quad y_k \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2}),$$

where the random variables y_k are uncorrelated. *Unlike the Gaussian setting, the random variables y_k are generally not independent!*

In numerical analysis, the random variables y_k are often **assumed** to be independent – this allows us to consider the above as a parametric PDE with $a(\mathbf{x}, \mathbf{y}) \equiv a(\mathbf{x}, \mathbf{y}(\omega))$ and $u(\mathbf{x}, \mathbf{y}) \equiv u(\mathbf{x}, \mathbf{y}(\omega))$, where $\mathbf{y} \in [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ is a *parametric vector* endowed with a uniform probability measure.

The Monte Carlo method

A simple technique to approximate the integral

$$I(f) := \int_{\text{supp}(p)} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

is to use a sample average. If we are able to draw an i.i.d. sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ from the probability distribution corresponding to the PDF p then one can consider the Monte Carlo estimate

$$I_n^{\text{MC}}(f) := \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i).$$

Generally speaking, the Law of Large Numbers and the Central Limit Theorem imply that

$$\lim_{n \rightarrow \infty} I_n^{\text{MC}}(f) = I(f) \quad \text{and} \quad \text{Var}(I_n^{\text{MC}}(f) - I(f)) \approx \frac{\text{Var}(f(X))}{n}$$

provided that $f(X)$ has finite mean and variance with X distributed according to the probability distribution corresponding to p .

Model problem 1: uniform and affine model

For the purposes of numerical analysis, it is often desirable to start by analyzing a simpler model. Fix $f \in L^2(D)$, let $U = [-1/2, 1/2]^{\mathbb{N}}$, and consider the problem of finding, for all $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient has the parametrization

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

where $a_0 \in L^\infty(D)$, there exist $a_{\min}, a_{\max} > 0$ s.t. $0 < a_{\min} \leq a(\mathbf{x}, \mathbf{y}) \leq a_{\max} < \infty$ for all $\mathbf{x} \in D$ and $\mathbf{y} \in U$, and the *stochastic fluctuations* $\psi_j: D \rightarrow \mathbb{R}$ are functions of the spatial variable such that

- $\psi_j \in L^\infty(D)$ for all $j \in \mathbb{N}$,
- $\sum_{j=1}^{\infty} \|\psi_j\|_{L^\infty(D)} < \infty$.

Goals: compute $\mathbb{E}[G(u)]$ and $\text{Var}[G(u)]$ for some bounded, linear functional $G: H_0^1(D) \rightarrow \mathbb{R}$ (*quantity of interest*); alternatively, one might be interested in $\mathbb{E}[u(\mathbf{x}, \cdot)]$ and $\text{Var}[u(\mathbf{x}, \cdot)]$ (full PDE solution).

Model problem 2: lognormal model

In many practical applications, it is desirable to model the diffusion coefficient as a lognormal random field. Fix $f \in L^2(D)$, let $U = \mathbb{R}_*^\mathbb{N}$, and consider the problem of finding, for all $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient has the parametrization

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

where $a_0 \in L^\infty(D)$ is such that $a_0(\mathbf{x}) > 0$ and the *stochastic fluctuations* $\psi_j: D \rightarrow \mathbb{R}$ are functions of the spatial variable such that

- $\psi_j \in L^\infty(D)$ for all $j \in \mathbb{N}$,
- $\sum_{j=1}^{\infty} \|\psi_j\|_{L^\infty(D)} < \infty$.

Goals: compute $\mathbb{E}[G(u)]$ and $\text{Var}[G(u)]$ for some bounded, linear functional $G: H_0^1(D) \rightarrow \mathbb{R}$; alternatively, one might be interested in $\mathbb{E}[u(\mathbf{x}, \cdot)]$ and $\text{Var}[u(\mathbf{x}, \cdot)]$.

Here, $\mathbb{R}_*^\mathbb{N} := \{\mathbf{y} \in \mathbb{R}^\mathbb{N} \mid \sum_{j=1}^{\infty} |y_j| \|\psi_j\|_{L^\infty(D)} < \infty\}$. More on this condition later...

Numerical experiment

Let us consider the problem of calculating the (dimensionally-truncated) $\mathbb{E}[u_s(x, \cdot)]$ using the Monte Carlo method. Fix the spatial domain $D = (0, 1)^2$ and source term $f(x) = x_1$. The PDE problem in this case is to find, for all $y \in \mathbb{R}^s$, $u_s(\cdot, y) \in H_0^1(D)$ s.t.

$$\int_D a_s(x, y) \nabla u_s(x, y) \cdot \nabla v(x) dx = \int_D f(x) v(x) dx \quad \text{for all } v \in H_0^1(D)$$

endowed with the (dimensionally-truncated) lognormally parameterized diffusion coefficient

$$a_s(x, y) = \exp \left(\sum_{k=1}^s y_k \psi_k(x) \right), \quad y_k \in \mathbb{R},$$

with stochastic fluctuations $\psi_k(x) := k^{-\vartheta} \sin(\pi k x_1) \sin(\pi k x_2)$ and a fixed decay parameter $\vartheta > 1$. We solve the PDE using a first-order finite element method with mesh size $h = 2^{-5}$ and stochastic dimension $s = 100$. We draw a random sample $y_1, \dots, y_n \sim \mathcal{N}(\mathbf{0}, I_s)$ and compute the Monte Carlo approximation

$$\mathbb{E}[u_{s,h}(x, y)] \approx \frac{1}{n} \sum_{k=1}^n u_{s,h}(x, y_k) = I_n^{\text{MC}}(u_{s,h}(x, \cdot)).$$

We plot the estimated L^2 error by using $I_{n'}^{\text{MC}}(u_{s,h}(x, \cdot))$ for $n' \gg n$ as the reference solution and compute $\|\mathbb{E}[u_{s,h}] - I_n^{\text{MC}}(u_{s,h})\|_{L^2(D)} \approx \|I_{n'}^{\text{MC}}(u_{s,h}) - I_n^{\text{MC}}(u_{s,h})\|_{L^2(D)}$. (To compute the $L^2(D)$ -norm of a function $v_h = \sum_j c_j \phi_j \in V_h$ belonging to a FE space, we use the mass matrix $M_{i,j} = \int_D \phi_i(x) \phi_j(x) dx$ as $\|v_h\|_{L^2} = \sqrt{\mathbf{c}^T M \mathbf{c}}$.)

Appendix

Rosenblatt transformation

In the non-Gaussian setting, the uncorrelated random variables can be made independent using, e.g., the *Rosenblatt transformation*.

The following is an excerpt from “Structural Reliability Analysis and Prediction”, 3rd edition, by R. E. Melchers and A. T. Beck (2018).

A dependent random vector $\mathbf{X} = (X_1, \dots, X_s)$ may be transformed to the *independent* uniformly distributed random vector $\mathbf{U} = (U_1, \dots, U_s)$ through the Rosenblatt (1952) transformation $\mathbf{U} = T\mathbf{X}$ given by

$$u_1 = \mathbb{P}(X_1 \leq x_1) = F_1(x_1),$$

$$u_2 = \mathbb{P}(X_2 \leq x_2 | X_1 = x_1) = F_2(x_2 | x_1),$$

⋮

$$u_s = \mathbb{P}(X_s \leq x_s | X_1 = x_1, \dots, X_{s-1} = x_{s-1}) = F_s(x_s | x_1, \dots, x_{s-1}).$$

If the joint PDF $p_{\mathbf{X}}$ is known, then the conditional CDF F_s can be determined by

$$F_s(x_s | x_1, \dots, x_{s-1}) = \frac{\int_{-\infty}^{x_s} p_{X_1, \dots, X_s}(x_1, \dots, x_{s-1}, t) dt}{p_{X_1, \dots, X_{s-1}}(x_1, \dots, x_{s-1})}.$$

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fifth lecture, May 12, 2025

Today's lecture follows the survey article

-  J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.* **22**:133–288, 2013.
<https://doi.org/10.1017/S0962492913000044>

Notations

- $\{1 : s\} := \{1, 2, \dots, s\}$ for $s \in \mathbb{N}$. We use fraktur letters to denote subsets $\mathfrak{u} \subseteq \{1 : s\}$. We use $|\mathfrak{u}|$ to denote the cardinality of set \mathfrak{u} .
- For $x \geq 0$, we define the fractional part $\{x\} := x - \lfloor x \rfloor = \text{mod}(x, 1)$. For $x < 0$, $\{x\} := x + \lfloor |x| \rfloor$. For $\mathbf{x} \in \mathbb{R}^s$, we define

$$\{\mathbf{x}\} := (\{x_1\}, \{x_2\}, \dots, \{x_s\}).$$

For example, $\{(1.2, 0.5, 2.77)\} = (0.2, 0.5, 0.77)$.

- For $\mathfrak{u} \subseteq \{1 : s\}$, we define $\mathbf{x}_{\mathfrak{u}} = (x_j)_{j \in \mathfrak{u}}$ and

$$\frac{\partial^{|\mathfrak{u}|}}{\partial \mathbf{x}_{\mathfrak{u}}} f(\mathbf{x}) := \prod_{j \in \mathfrak{u}} \frac{\partial}{\partial x_j} f(\mathbf{x}).$$

For example, with $\mathfrak{u} = \{1, 2, 4\}$, we have $|\mathfrak{u}| = 3$, $\mathbf{x}_{\mathfrak{u}} = (x_1, x_2, x_4)$, and

$$\frac{\partial^{|\mathfrak{u}|}}{\partial \mathbf{x}_{\mathfrak{u}}} f(\mathbf{x}) = \frac{\partial^3}{\partial x_1 \partial x_2 \partial x_4} f(\mathbf{x}).$$

Quasi-Monte Carlo methods

Let $f \in C([0, 1]^s)$. We consider the problem of approximating the high-dimensional integral

$$I_s f = \int_{[0,1]^s} f(\mathbf{y}) d\mathbf{y}.$$

Quasi-Monte Carlo (QMC) methods are a class of *equal weight* cubature rules

$$Q_{n,s} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{t}_i),$$

where $(\mathbf{t}_i)_{i=0}^{n-1}$ is an ensemble of *deterministic* nodes in $[0, 1]^s$ (**not** a random sample of $\mathcal{U}([0, 1]^s)$).

QMC methods exploit the smoothness and anisotropy of an integrand in order to achieve better-than-Monte Carlo cubature convergence rates.

Rank-1 lattice rules

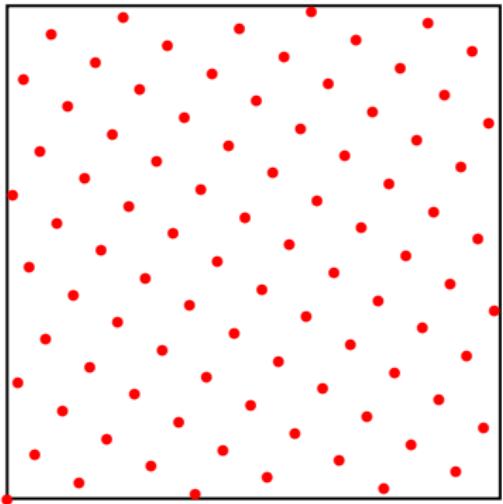
Rank-1 lattice rules

$$Q_{n,s} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{t}_i)$$

have the points

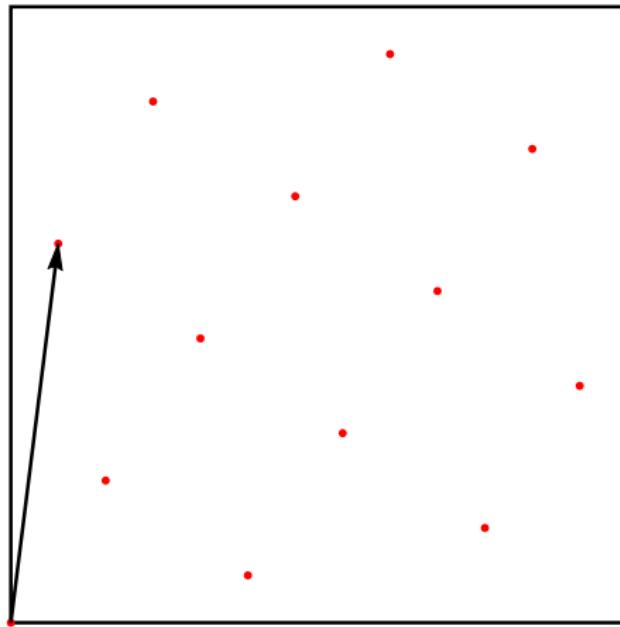
$$\mathbf{t}_i = \text{mod}\left(\frac{i\mathbf{z}}{n}, 1\right), \quad i \in \{0, \dots, n-1\},$$

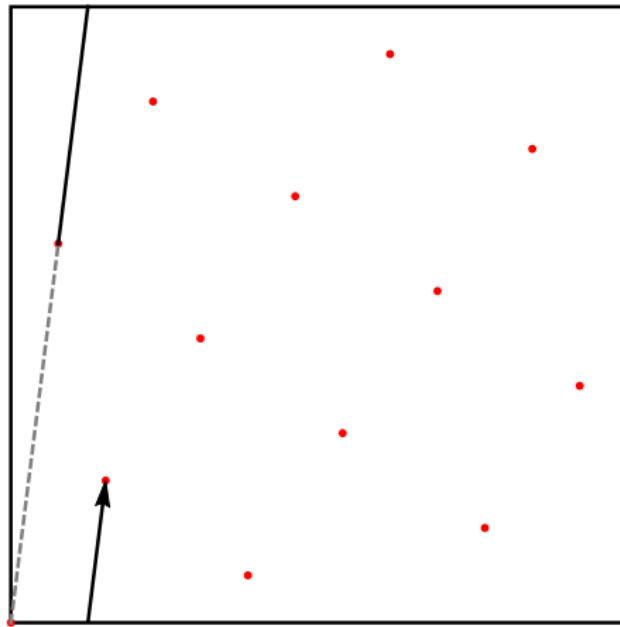
where the entire point set is determined by the *generating vector* $\mathbf{z} \in \mathbb{N}^s$, with all components *coprime* to n .

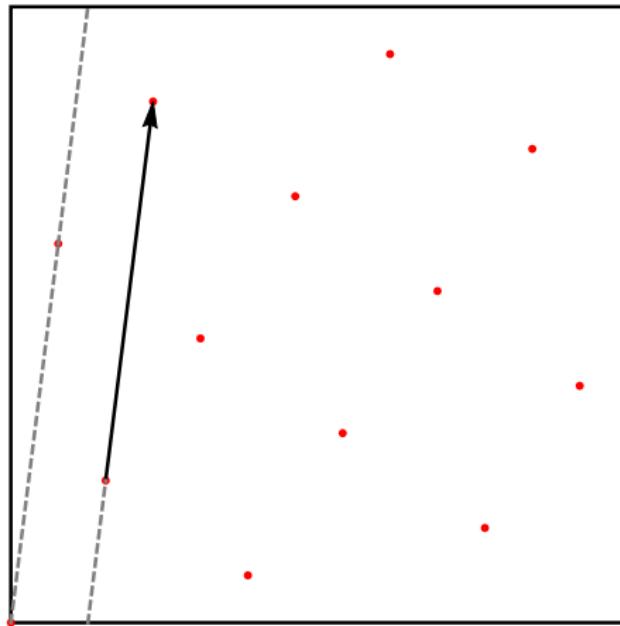


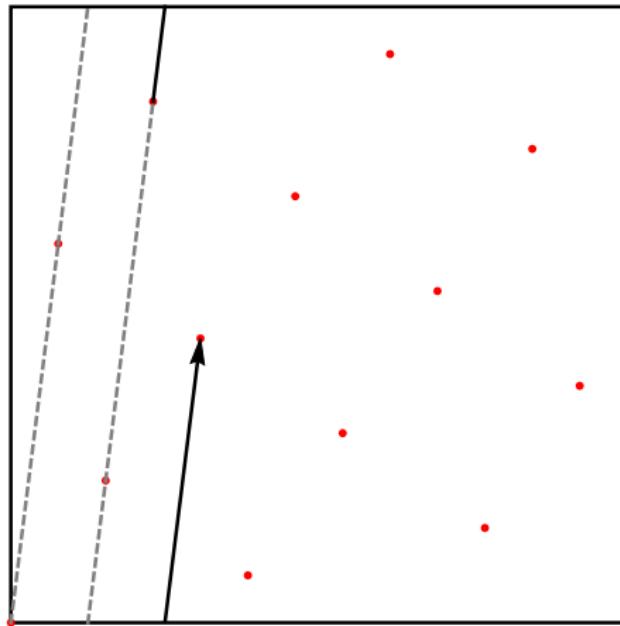
Lattice rule with $\mathbf{z} = (1, 55)$ and $n = 89$
nodes in $[0, 1]^2$

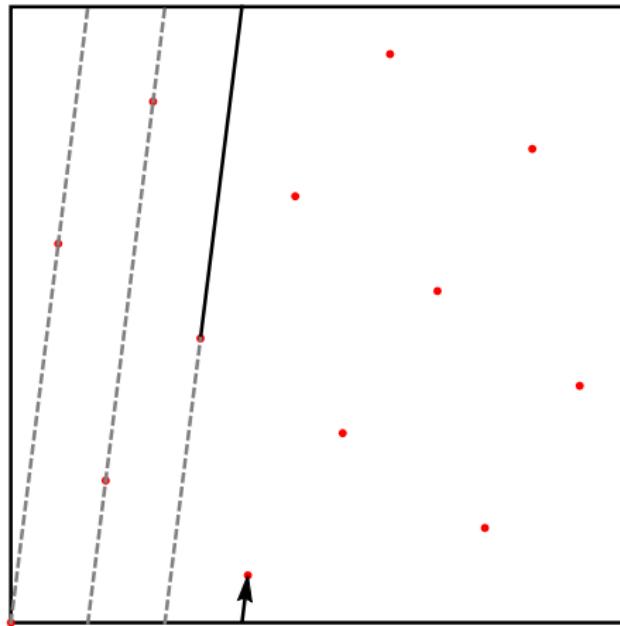
The quality of the lattice rule is determined by the generating vector

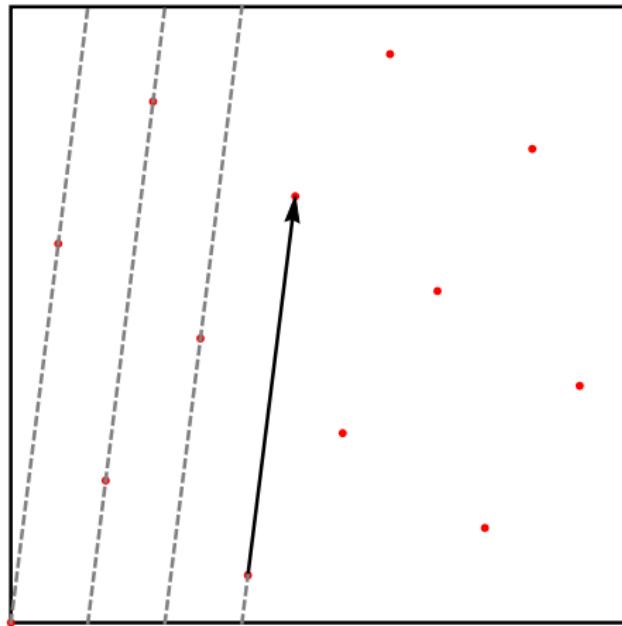


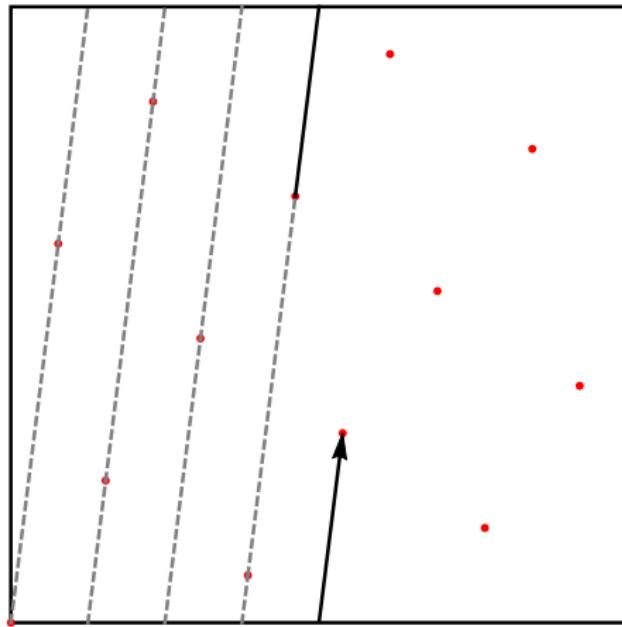


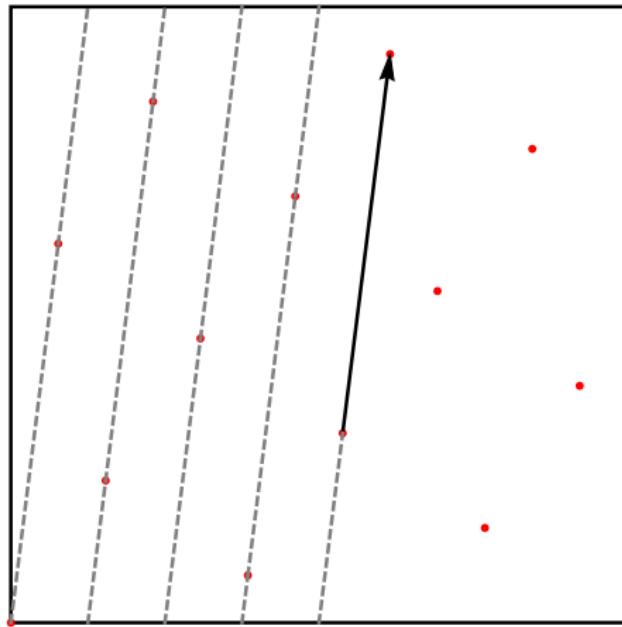


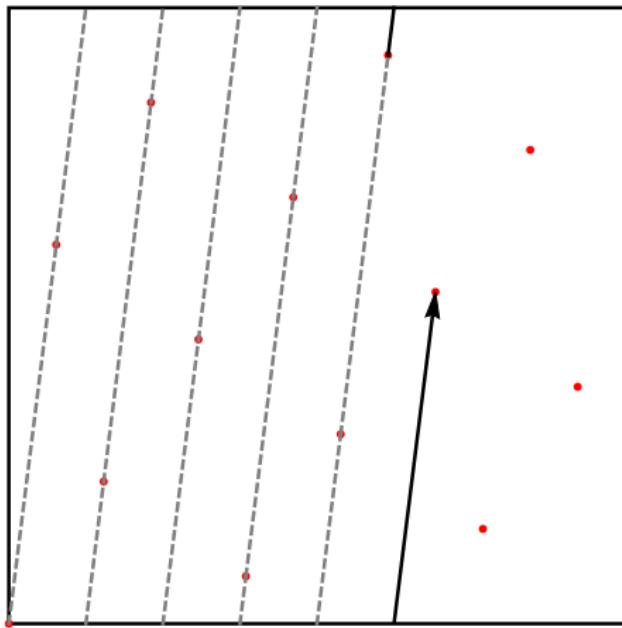


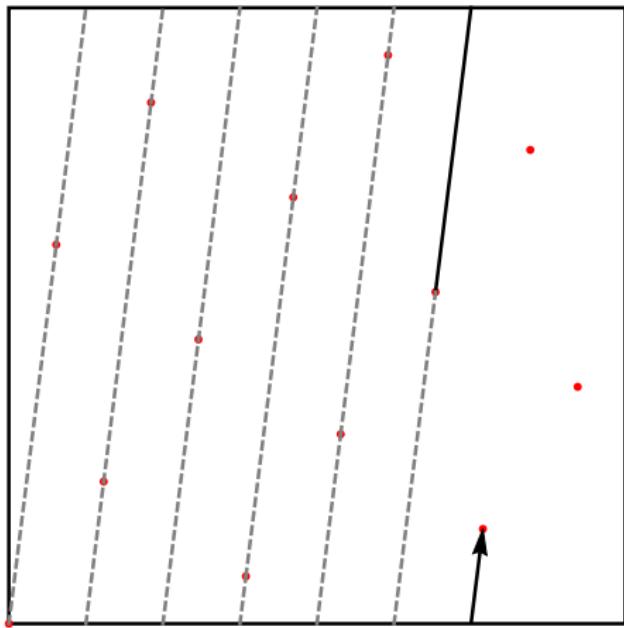


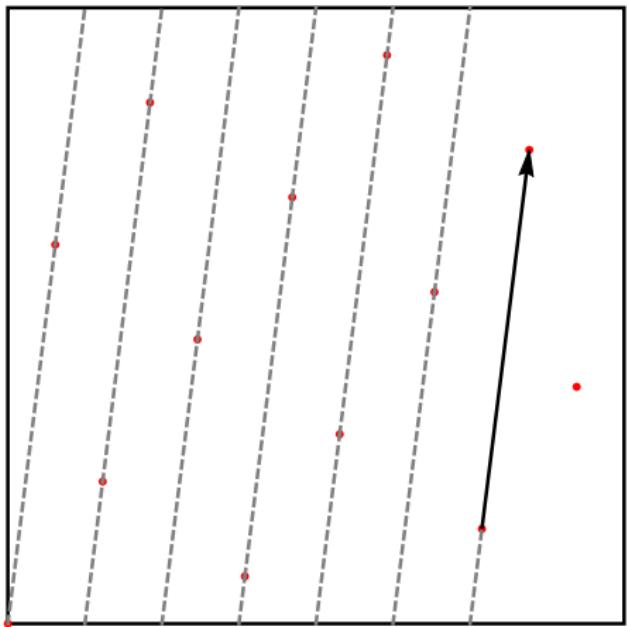


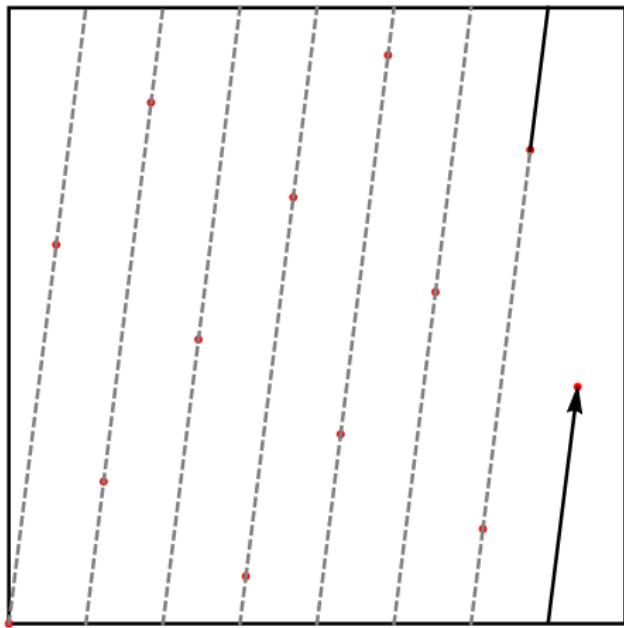












Historical remarks on the development of lattice rules

- Number theorists (Korobov, Zaremba, Hua) in the 1950s and 60s.
- Lattice rules for multiple integration (Sloan and Kachoyan 1987; Sloan and Joe 1994).
- Weighted spaces (Sloan and Woźniakowski 1998; Hickernell 1996).
- Component-by-component (CBC) construction of lattice rules (Kuo, Joe, Sloan 2002).
- Fast CBC algorithm (Cools and Nuyens 2006; Kuo, Cools, and Nuyens 2006).
- Uncertainty quantification of PDEs using QMC methods (Kuo, Schwab, Sloan 2012).

and of course many, many others! (Dick, Giles, Goda, Graham, Kritzer, Niederreiter, Pillichshammer, Wasilkowski, ...)

Brief introduction to the classical theory of lattice rules

Let $f : [0, 1]^s \rightarrow \mathbb{R}$ be an absolutely continuous and 1-periodic function, i.e.,

$$f(y_1, y_2, \dots, y_s) = f(y_1 + 1, y_2, \dots, y_s) = f(y_1, y_2 + 1, \dots, y_s) = \dots,$$

with an absolutely convergent Fourier series

$$f(\mathbf{x}) = \sum_{\mathbf{h} \in \mathbb{Z}^s} \widehat{f}(\mathbf{h}) e^{2\pi i \mathbf{h} \cdot \mathbf{x}}, \quad \widehat{f}(\mathbf{h}) := \int_{[0,1]^s} f(\mathbf{x}) e^{-2\pi i \mathbf{h} \cdot \mathbf{x}} d\mathbf{x}.$$

Then the lattice rule error is precisely the sum of the integrand's Fourier coefficients over the so-called *dual lattice*.

Theorem (Rank-1 lattice rule error)

Under the aforementioned conditions on $f : [0, 1]^s \rightarrow \mathbb{R}$, there holds

$$Q_{n,s}(f) - I_s(f) = \sum_{\mathbf{h} \in \Lambda^\perp \setminus \{\mathbf{0}\}} \widehat{f}(\mathbf{h}),$$

where the dual lattice

$$\Lambda^\perp := \{\mathbf{h} \in \mathbb{Z}^s \mid \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n}\}$$

is determined entirely by the generating vector $\mathbf{z} \in \mathbb{N}^s$ and $n \in \mathbb{N}$.

For future convenience, let us prove a couple of helpful auxiliary identities.

Lemma

Let $\mathbf{h} = (h_1, \dots, h_s) \in \mathbb{Z}^s$ and $n \in \mathbb{N}$. Then

$$\int_{[0,1]^s} e^{2\pi i \mathbf{h} \cdot \mathbf{x}} d\mathbf{x} = \begin{cases} 1 & \text{if } \mathbf{h} = \mathbf{0} \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i k \mathbf{h} \cdot \mathbf{z}/n} = \begin{cases} 1 & \text{if } \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n} \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By Fubini's theorem

$$\int_{[0,1]^s} e^{2\pi i \mathbf{h} \cdot \mathbf{x}} d\mathbf{x} = \prod_{j=1}^s \int_0^1 e^{2\pi i h_j x_j} dx_j, \quad (1)$$

where

$$\int_0^1 e^{2\pi i h_j x_j} dx_j = \begin{cases} \int_0^1 dx_j & \text{if } h_j = 0 \\ \left[\frac{e^{2\pi i h_j x_j}}{2\pi i h_j} \right]_{x_j=0}^{x_j=1} & \text{if } h_j \neq 0 \end{cases} = \begin{cases} 1 & \text{if } h_j = 0 \\ 0 & \text{if } h_j \neq 0. \end{cases}$$

Thus the expression (1) is zero unless $h_1 = h_2 = \dots = h_s = 0$.

To prove the second claim

$$\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i k \mathbf{h} \cdot \mathbf{z} / n} = \begin{cases} 1 & \text{if } \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n} \\ 0 & \text{otherwise} \end{cases}$$

consider two cases:

- If $\mathbf{h} \cdot \mathbf{z}$ is a multiple of n , i.e., $\mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n}$, then clearly

$$\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i k \mathbf{h} \cdot \mathbf{z} / n} = \frac{1}{n} \sum_{k=0}^{n-1} e^0 = 1.$$

- If $\mathbf{h} \cdot \mathbf{z}$ is not a multiple of n , then by the geometric sum formula

$$\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i k \mathbf{h} \cdot \mathbf{z} / n} = \frac{1}{n} \sum_{k=0}^{n-1} \left(e^{2\pi i \mathbf{h} \cdot \mathbf{z} / n} \right)^k = \frac{1}{n} \frac{1 - (e^{2\pi i \mathbf{h} \cdot \mathbf{z} / n})^n}{1 - e^{2\pi i \mathbf{h} \cdot \mathbf{z} / n}} = 0.$$

This yields the assertion. □

Proof (Rank-1 lattice rule error). Using the Fourier series representation

$$f(\mathbf{x}) = \sum_{\mathbf{h} \in \mathbb{Z}^s} \widehat{f}(\mathbf{h}) e^{2\pi i \mathbf{h} \cdot \mathbf{x}}, \quad \widehat{f}(\mathbf{h}) := \int_{[0,1]^s} f(\mathbf{x}) e^{-2\pi i \mathbf{h} \cdot \mathbf{x}} d\mathbf{x},$$

and noting that $e^{2\pi i \left\{ \frac{kz}{n} \right\} \cdot \mathbf{h}} = e^{2\pi i k \mathbf{z} \cdot \mathbf{h} / n}$, we can change the order of the series (note that the Fourier series is absolutely convergent!) to obtain

$$\begin{aligned} Q_{n,s}(f) - I_s(f) &= \frac{1}{n} \sum_{k=0}^{n-1} f\left(\left\{ \frac{k\mathbf{z}}{n} \right\}\right) - \int_{[0,1]^s} f(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\mathbf{h} \in \mathbb{Z}^s} \widehat{f}(\mathbf{h}) e^{2\pi i \mathbf{h} \cdot \mathbf{z} / n} - \widehat{f}(\mathbf{0}) \\ &= \sum_{\mathbf{h} \in \mathbb{Z}^s} \widehat{f}(\mathbf{h}) \underbrace{\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i \mathbf{h} \cdot \mathbf{z} / n}}_{\begin{array}{l} = 1 \text{ if } \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n} \\ = 0 \text{ otherwise} \end{array}} - \widehat{f}(\mathbf{0}) \\ &= \sum_{\substack{\mathbf{h} \in \mathbb{Z}^s \\ \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n}}} \widehat{f}(\mathbf{h}) - \widehat{f}(\mathbf{0}) = \sum_{\substack{\mathbf{h} \in \mathbb{Z}^s \setminus \{\mathbf{0}\} \\ \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n}}} \widehat{f}(\mathbf{h}). \quad \square \end{aligned}$$

Ultimately, we are interested in applying lattice rules for *non-periodic*, smooth functions. We will need to put in a bit more effort to make this method work in the non-periodic setting...

Worst-case error and reproducing kernel Hilbert space (RKHS)

Worst-case error

In the classical study of quadrature and cubature rules, we usually consider the so-called *worst-case error*. Suppose that $f \in H$, where H is a Hilbert space continuously embedded in $C([0, 1]^s)$. Let $I_s : H \rightarrow \mathbb{R}$ be an integral operator

$$I_s f := \int_{[0,1]^s} f(\mathbf{x}) d\mathbf{x}$$

and let $Q_{n,s} : H \rightarrow \mathbb{R}$ be a QMC rule

$$Q_{n,s} f := \frac{1}{n} \sum_{i=0}^{n-1} f(\mathbf{t}_i),$$

where $P := \{\mathbf{t}_i \in [0, 1]^s \mid 0 \leq i \leq n - 1\}$ is a collection of cubature nodes. The worst-case error of cubature rule $Q_{n,s}$ in H is defined by

$$e_{n,s}(P; H) := \sup_{\substack{f \in H \\ \|f\|_H \leq 1}} |I_s f - Q_{n,s} f|.$$

Note that this is precisely the operator norm of $\|I_s - Q_{n,s}\|_{H \rightarrow \mathbb{R}}$.

Since the worst-case error is just the operator norm of $I_s - Q_{n,s}$, we can express the cubature error as

$$|I_s f - Q_{n,s} f| \leq e_{n,s}(P; H) \|f\|_H.$$

Worst-case errors are in general hard to compute – except for the special case, when H is a *reproducing kernel Hilbert space* (RKHS).

Our strategy will be to *choose* the Hilbert space H (where our integrand f lives) to be such that it is possible to write down the expression for $e_{n,s}(P; H)$ *explicitly* given a family of QMC rules. This allows us to analyze the dependence of the cubature error w.r.t. n and s .

We will end up taking H as an *unanchored, weighted Sobolev space* since this choice turns out to be “compatible” with the family of (randomly shifted) lattice rules!

Reproducing kernel Hilbert space (RKHS)

Let H be a Hilbert space of functions on $D \subseteq \mathbb{R}^s$, with the property that *every point evaluation is a bounded linear functional*. That is, for any $\mathbf{y} \in D$, let

$$T_{\mathbf{y}}(f) := f(\mathbf{y}) \quad \text{for all } f \in H.$$

Then, since $T_{\mathbf{y}}$ is a bounded linear functional, by Riesz representation theorem there exists a unique representer $a_{\mathbf{y}} := K(\cdot, \mathbf{y}) \in H$ such that

$$T_{\mathbf{y}}(f) = \langle f, a_{\mathbf{y}} \rangle = \langle f, K(\cdot, \mathbf{y}) \rangle \quad \text{for all } f \in H.$$

The function $K(\mathbf{x}, \mathbf{y})$ is known as the *reproducing kernel* of H .

Definition (Reproducing kernel)

A *reproducing kernel* of a Hilbert space H of functions on $D \subseteq \mathbb{R}^s$ is a function $K: D \times D \rightarrow \mathbb{R}$ which satisfies

$$K(\cdot, \mathbf{y}) \in H \quad \text{for all } \mathbf{y} \in D$$

$$\text{and } f(\mathbf{y}) = \langle f, K(\cdot, \mathbf{y}) \rangle \quad \text{for all } f \in H \text{ and } \mathbf{y} \in D.$$

The latter property is known as the *reproducing property*.

Remarks

- A *reproducing kernel Hilbert space* (RKHS) is a Hilbert space equipped with a reproducing kernel, or equivalently, it is a Hilbert space in which *every point evaluation is a bounded linear functional.*
- For any other bounded linear functional $A: H \rightarrow \mathbb{R}$, its representer $a \in H$ satisfying $A(f) = \langle f, a \rangle$ for all $f \in H$ is given by

$$a(\mathbf{y}) = \langle a, K(\cdot, \mathbf{y}) \rangle = \langle K(\cdot, \mathbf{y}), a \rangle = A(K(\cdot, \mathbf{y})) \quad \text{for all } \mathbf{y} \in D.$$

- Any reproducing kernel $K(\mathbf{x}, \mathbf{y})$ is symmetric in its arguments:

$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in D.$$

Proof. For fixed $\mathbf{y} \in D$, apply the reproducing property to the function $f = K(\cdot, \mathbf{y})$ to get

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle = \langle K(\cdot, \mathbf{y}), K(\cdot, \mathbf{x}) \rangle \\ &= \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y}) \rangle = K(\mathbf{y}, \mathbf{x}). \quad \square \end{aligned}$$

Example

Suppose that we have a Hilbert space containing continuous functions on $[0, 1]$ with square-integrable first order derivatives, equipped with the inner product

$$\langle f, g \rangle = \left(\int_0^1 f(x) dx \right) \left(\int_0^1 g(x) dx \right) + \int_0^1 f'(x)g'(x) dx.$$

Then this space has the *reproducing kernel*

$$K(x, y) = 1 + \eta(x, y), \quad \eta(x, y) = \frac{1}{2}B_2(|x - y|) + (x - \frac{1}{2})(y - \frac{1}{2}),$$

where $B_2(x) := x^2 - x + \frac{1}{6}$ denotes the *Bernoulli polynomial of degree 2*.

That is, we claim that

$$\langle f, K(\cdot, y) \rangle = f(y) \quad \text{for all } y \in [0, 1].$$

Example (continued)

By observing that

$$\int_0^1 K(x, y) dx = 1 \quad \text{and} \quad \frac{\partial}{\partial x} K(x, y) = x - \frac{1}{2} - \frac{1}{2} \operatorname{sign}(x - y),$$

there holds

$$\begin{aligned}\langle f, K(\cdot, y) \rangle &= \left(\int_0^1 f(x) dx \right) \underbrace{\left(\int_0^1 K(x, y) dx \right)}_{=1} + \int_0^1 f'(x) \left(x - \frac{1}{2} - \frac{1}{2} \operatorname{sign}(x - y) \right) dx \\ &= \int_0^1 f(x) dx + \int_0^1 f'(x)x dx - \frac{1}{2} \int_0^1 f'(x) dx + \frac{1}{2} \int_0^y f'(x) dx - \frac{1}{2} \int_y^1 f'(x) dx \\ &= \cancel{\int_0^1 f(x) dx} + \cancel{f(1)} - \cancel{\int_0^1 f(x) dx} - \frac{1}{2} \cancel{f(1)} + \frac{1}{2} \cancel{f(0)} + \frac{1}{2} f(y) - \frac{1}{2} \cancel{f(0)} - \frac{1}{2} \cancel{f(1)} + \frac{1}{2} f(y) \\ &= f(y)\end{aligned}$$

for all $y \in [0, 1]$, as desired.

Theorem

Let $H := H_s(K)$ be an RKHS and let $K: [0, 1]^s \times [0, 1]^s \rightarrow \mathbb{R}$ be a reproducing kernel that satisfies

$$\int_{[0,1]^s} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} < \infty.$$

Then

$$\begin{aligned} e_{n,s}^2(P; H_s(K)) &= \int_{[0,1]^s} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} - \frac{2}{n} \sum_{i=0}^{n-1} \int_{[0,1]^s} K(\mathbf{t}_i, \mathbf{y}) d\mathbf{y} \\ &\quad + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K(\mathbf{t}_i, \mathbf{t}_j). \end{aligned} \tag{2}$$

Proof. For $f \in H$, we apply the reproducing property $f(\mathbf{t}_k) = \langle f, K(\cdot, \mathbf{t}_k) \rangle_H$ and average the results to obtain

$$Q_{n,s}f = \frac{1}{n} \sum_{k=0}^{n-1} f(\mathbf{t}_k) = \frac{1}{n} \sum_{k=0}^{n-1} \langle f, K(\cdot, \mathbf{t}_k) \rangle_H = \left\langle f, \frac{1}{n} \sum_{k=0}^{n-1} K(\cdot, \mathbf{t}_k) \right\rangle_H. \quad (3)$$

Similarly, we find that

$$I_s f = \int_{[0,1]^s} f(\mathbf{x}) d\mathbf{x} = \int_{[0,1]^s} \langle f, K(\cdot, \mathbf{x}) \rangle_H d\mathbf{x} = \left\langle f, \int_{[0,1]^s} K(\cdot, \mathbf{x}) d\mathbf{x} \right\rangle_H, \quad (4)$$

which holds provided that $\int_{[0,1]^s} K(\cdot, \mathbf{x}) d\mathbf{x} \in H$. However, this is guaranteed by our assumption since

$$\begin{aligned} \left\| \int_{[0,1]^s} K(\cdot, \mathbf{x}) d\mathbf{x} \right\|_H^2 &= \int_{[0,1]^s} \int_{[0,1]^s} \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y}) \rangle_H d\mathbf{x} d\mathbf{y} \\ &= \int_{[0,1]^s} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} < \infty, \end{aligned}$$

which will hold for all the kernels we shall consider.

Taking the difference of (3) and (4) yields

$$I_s f - Q_{n,s} f = \left\langle f, \int_{[0,1]^s} K(\cdot, \mathbf{x}) d\mathbf{x} - \frac{1}{n} \sum_{i=0}^{n-1} K(\cdot, \mathbf{t}_i) \right\rangle_H = \langle f, \xi \rangle_H,$$

where

$$\xi(\mathbf{y}) := \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} - \frac{1}{n} \sum_{i=0}^{n-1} K(\mathbf{y}, \mathbf{t}_i), \quad \mathbf{y} \in [0, 1]^s$$

is called the *representer* of the integration error since

$$e_{n,s}(P; H) = \sup_{\|f\| \leq 1} |\langle f, \xi \rangle_H| = \|\xi\|_H.$$

Especially, the supremum is attained by $f = \xi / \|\xi\| \in H$ and we obtain

$$\begin{aligned} e_{n,s}^2(P; H) &= \left\| \int_{[0,1]^s} K(\cdot, \mathbf{x}) d\mathbf{x} - \frac{1}{n} \sum_{i=0}^{n-1} K(\mathbf{x}, \mathbf{t}_i) \right\|^2 \\ &= \int_{[0,1]^s} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} - \frac{2}{n} \sum_{i=0}^{n-1} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{t}_i) d\mathbf{x} + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K(\mathbf{t}_i, \mathbf{t}_j), \end{aligned}$$

as desired. □



Randomly shifted rank-1 lattice points

In what follows, we will discuss randomly shifted QMC rules.

Consider the rank-1 lattice point set $\mathbf{t}_k := \left\{ \frac{kz}{n} \right\}$ for some generating vector $z \in \mathbb{N}^s$ and fixed $n \in \mathbb{N}$. Given a vector $\Delta \in [0, 1]^s$, known as the *shift*, the Δ -shift of the QMC points $\mathbf{t}_0, \dots, \mathbf{t}_{n-1}$ is defined as the point set

$$\{\mathbf{t}_k + \Delta\}, \quad k = 0, \dots, n - 1.$$

Shifting preserves the lattice structure. In practice, we will generate a number of independent random shifts $\Delta_0, \dots, \Delta_{R-1}$ from $\mathcal{U}([0, 1]^s)$ and take the average of $\Delta_0, \dots, \Delta_{R-1}$ -shifted QMC rules as our approximation of I_s .

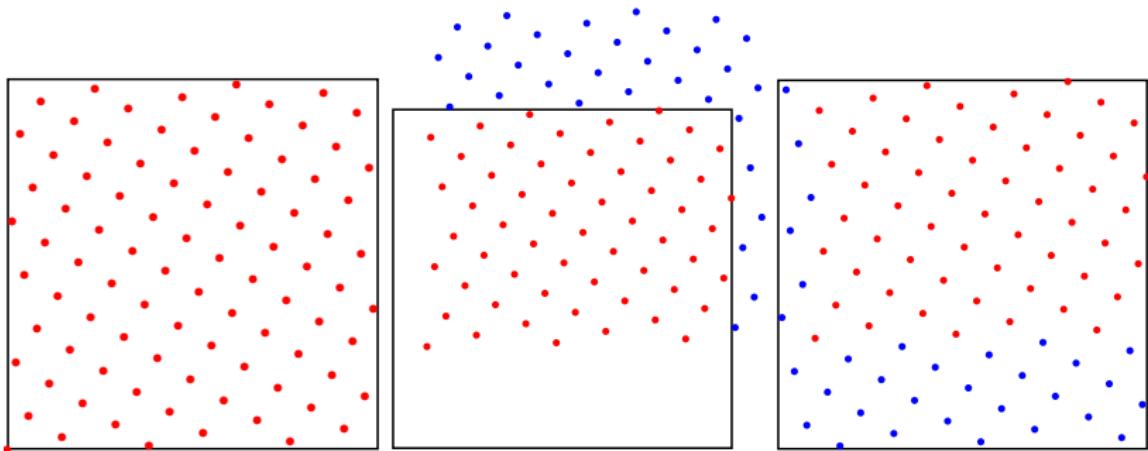
Advantages:

- Leads to a shift-invariant kernel (advantageous for high-dimensional computation).
- Randomization yields an unbiased estimator of the integral.
- Randomization provides a practical error estimate.

Shifted rank-1 lattice rules have points

$$\left\{ \frac{k\mathbf{z}}{n} + \boldsymbol{\Delta} \right\}, \quad k = 0, \dots, n-1.$$

Use a number of random shifts for error estimation.



Lattice rule shifted by $\boldsymbol{\Delta} = (0.1, 0.3)$.

Randomization in practice

- Generate R independent random shifts $\Delta_0, \dots, \Delta_{R-1}$ from $\mathcal{U}([0, 1]^s)$.
- For a given QMC rule with points $(\mathbf{t}_i)_{i=0}^{n-1} \subset [0, 1]^s$, form the approximations $Q_{n,s}^{(0)} f, \dots, Q_{n,s}^{(R-1)} f$, where

$$Q_{n,s}^{\Delta_r} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\{\mathbf{t}_i + \Delta_r\}), \quad r = 0, \dots, R-1,$$

is the approximation of the integral using a Δ_r -shift of the original QMC rule.

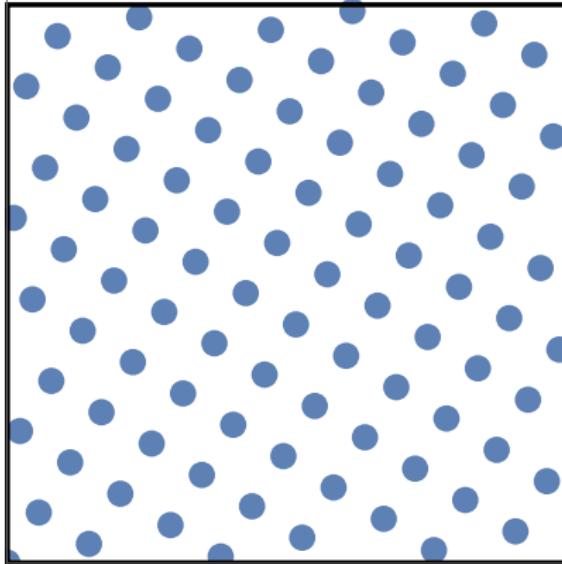
- We take the *average*

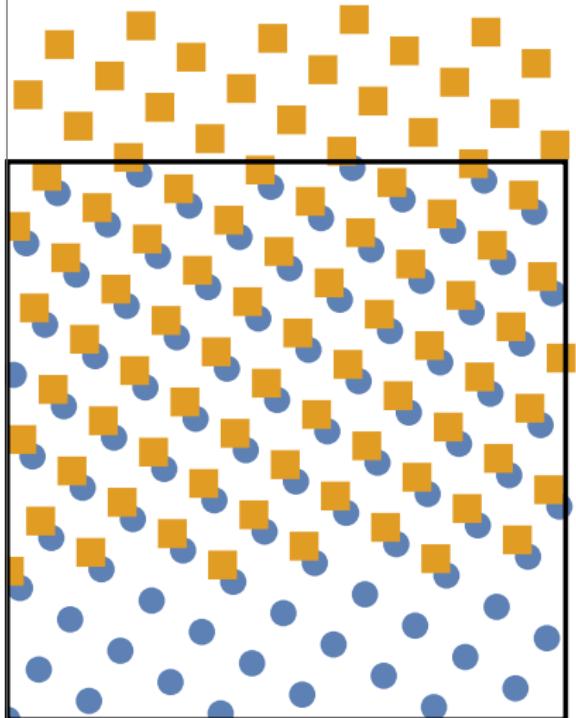
$$\overline{Q}_{n,s,R} f = \frac{1}{R} \sum_{r=0}^{R-1} Q_{n,s}^{\Delta_r} f$$

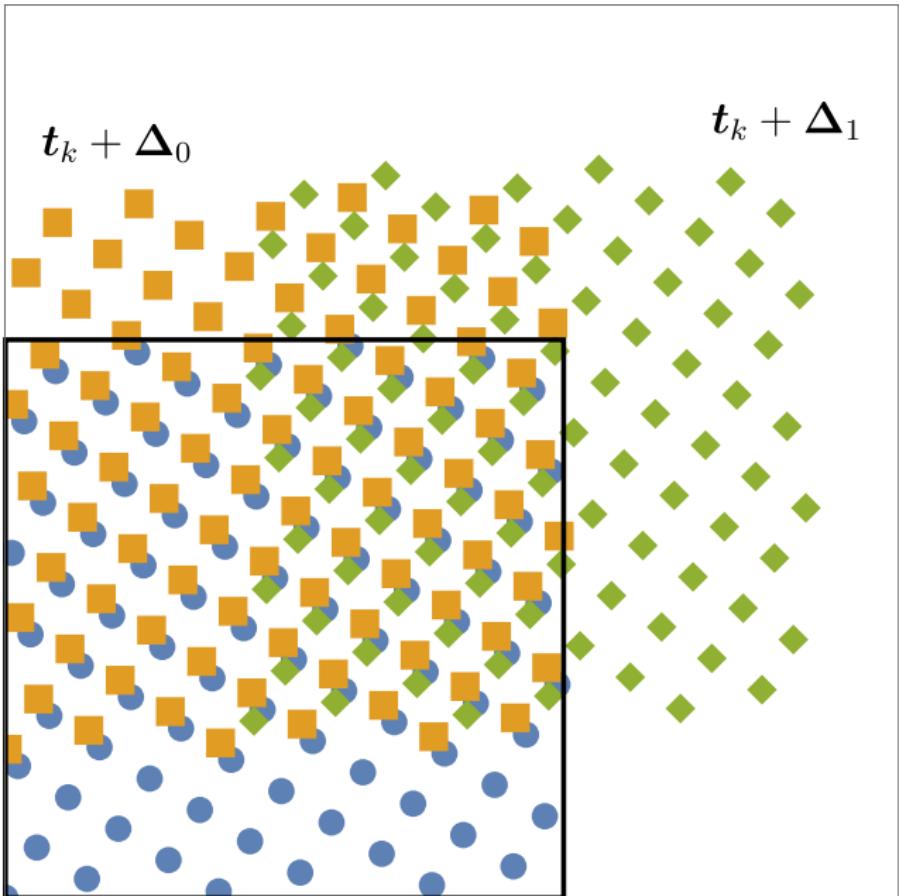
as our *final* approximation of the integral.

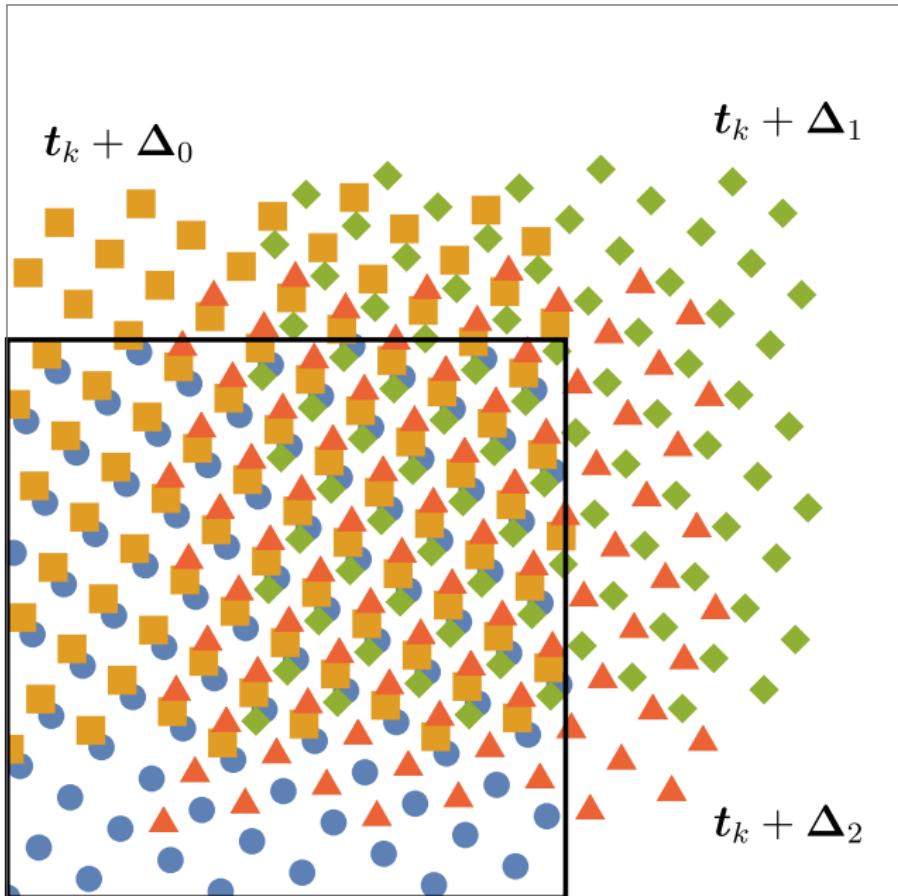
- An *unbiased* estimate for the mean-square error of $\overline{Q}_{n,s,R} f$ is given by

$$\mathbb{E}_{\Delta} |I_s f - Q_{n,s}^{\Delta} f|^2 \approx \frac{1}{R(R-1)} \sum_{r=0}^{R-1} (Q_{n,s}^{\Delta_r} f - \overline{Q}_{n,s,R} f)^2.$$

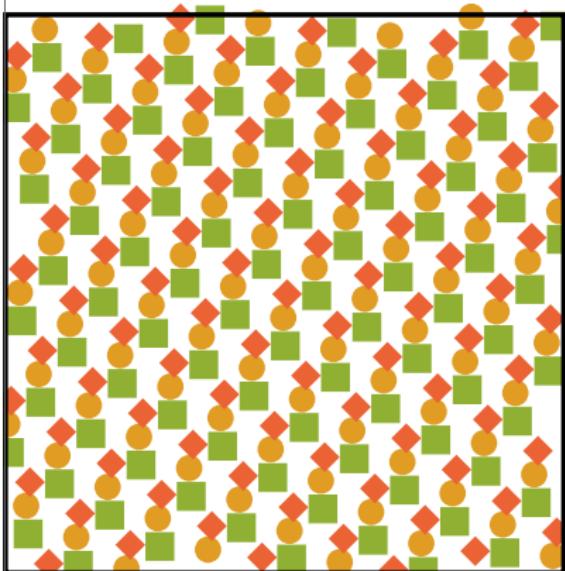


$t_k + \Delta_0$ 





$$\{\{t_k + \Delta_0\}, \{t_k + \Delta_1\}, \{t_k + \Delta_2\}\}$$



$$Q_{n,s}^{\Delta_0} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\{t_i + \Delta_0\}), \quad Q_{n,s}^{\Delta_1} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\{t_i + \Delta_1\}), \quad Q_{n,s}^{\Delta_2} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\{t_i + \Delta_2\})$$

$$\text{QMC approximation with 3 random shifts: } \overline{Q}_{n,s,3} f = \frac{Q_{n,s}^{\Delta_0} f + Q_{n,s}^{\Delta_1} f + Q_{n,s}^{\Delta_2} f}{3}.$$

Shift-averaged worst-case error

For any QMC point set $P = \{\mathbf{t}_0, \dots, \mathbf{t}_{n-1}\}$ and any shift $\Delta \in [0, 1]^s$, let

$$P + \Delta := \{\{\mathbf{t}_i + \Delta\} \mid i = 0, 1, \dots, n-1\}$$

denote the *shifted QMC point set*, and let $Q_{n,s}^\Delta f$ denote a corresponding shifted QMC rule (over the point set $P + \Delta$). For any integrand $f \in H$, it follows from the definition of the worst-case error that

$$|I_s f - Q_{n,s}(\Delta; f)| \leq e_{n,s}(P + \Delta; H) \|f\|_H,$$

where $e_{n,s}(P + \Delta; H) := \sup_{\|f\|_H \leq 1} |I_s(f) - Q_{n,s}^\Delta f|$. We deduce a bound for the *root-mean-square* error

$$\sqrt{\mathbb{E}_\Delta |I_s f - Q_{n,s}^\Delta f|^2} \leq e_{n,s}^{\text{sh}}(P; H) \|f\|_H,$$

where the expected value \mathbb{E}_Δ is taken over the random shift Δ which is uniformly distributed over $[0, 1]^s$ and the quantity

$$e_{n,s}^{\text{sh}}(P; H) := \sqrt{\int_{[0,1]^s} e_{n,s}^2(P + \Delta; H) d\Delta}$$

is called the *shift-averaged worst-case error*.

Theorem (Formula for the shift-averaged worst-case error)

$$[e_{n,s}^{\text{sh}}(P; H_s(K))]^2 = - \int_{[0,1]^s} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K^{\text{sh}}(\mathbf{t}_i, \mathbf{t}_j),$$

where

$$K^{\text{sh}}(\mathbf{x}, \mathbf{y}) := \int_{[0,1]^s} K(\{\mathbf{x} + \boldsymbol{\Delta}\}, \{\mathbf{y} + \boldsymbol{\Delta}\}) d\boldsymbol{\Delta}, \quad \mathbf{x}, \mathbf{y} \in [0, 1]^s.$$

Proof. The definition of shift-averaged WCE and (2) imply

$$\begin{aligned} [e_{n,s}^{\text{sh}}(P; H_s(K))]^2 &= \int_{[0,1]^s} e_{n,s}^2(P + \boldsymbol{\Delta}; H) d\boldsymbol{\Delta} \\ &= \int_{[0,1]^s} \int_{[0,1]^s} K(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} - \frac{2}{n} \sum_{i=0}^{n-1} \int_{[0,1]^s} \int_{[0,1]^s} K(\{\mathbf{t}_i + \boldsymbol{\Delta}\}, \mathbf{y}) d\boldsymbol{\Delta} d\mathbf{y} \\ &\quad + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \int_{[0,1]^s} K(\{\mathbf{t}_i + \boldsymbol{\Delta}\}, \{\mathbf{t}_j + \boldsymbol{\Delta}\}) d\boldsymbol{\Delta}. \end{aligned}$$

The result follows by a change of variables $\mathbf{x} = \{\mathbf{t}_i + \boldsymbol{\Delta}\}$ in the second term. □



Remarks

$$K^{\text{sh}}(\mathbf{x}, \mathbf{y}) := \int_{[0,1]^s} K(\{\mathbf{x} + \boldsymbol{\Delta}\}, \{\mathbf{y} + \boldsymbol{\Delta}\}) d\boldsymbol{\Delta}, \quad \mathbf{x}, \mathbf{y} \in [0, 1]^s.$$

- The function K^{sh} is actually a reproducing kernel, with the *shift-invariant property*

$$K^{\text{sh}}(\mathbf{x}, \mathbf{y}) = K^{\text{sh}}(\{\mathbf{x} + \boldsymbol{\Delta}\}, \{\mathbf{y} + \boldsymbol{\Delta}\}) \quad \text{for all } \mathbf{x}, \mathbf{y}, \boldsymbol{\Delta} \in [0, 1].$$

Equivalently,

$$K^{\text{sh}}(\mathbf{x}, \mathbf{y}) = K^{\text{sh}}(\{\mathbf{x} - \mathbf{y}\}, \mathbf{0}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in [0, 1].$$

- The function K^{sh} is called the *shift-invariant kernel associated with K* .

Weighted Sobolev spaces

Unanchored, weighted Sobolev space

For our purposes, the relevant function space setting will be the *unanchored, weighted Sobolev space*. For any given collection $(\gamma_u)_{u \subseteq \{1:s\}}$ of positive numbers (called *weights*), we associate a space $H_{s,\gamma}$ containing continuous functions on $[0, 1]^s$ whose *mixed first partial derivatives are square-integrable*. It is defined by the reproducing kernel

$$K_{s,\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{u \subseteq \{1:s\}} \gamma_u \prod_{j \in u} \eta(x_j, y_j), \quad \eta(x, y) := \frac{1}{2} B_2(|x-y|) + (x - \frac{1}{2})(y - \frac{1}{2}),$$

where $B_2(x) := x^2 - x + \frac{1}{6}$ is the Bernoulli polynomial of degree 2.

Norm $\|f\|_{s,\gamma} = \sqrt{\langle f, f \rangle_{s,\gamma}}$ induced by the inner product

$$\begin{aligned} \langle f, g \rangle_{s,\gamma} &= \sum_{u \subseteq \{1:s\}} \frac{1}{\gamma_u} \int_{[0,1]^{|u|}} \left(\int_{[0,1]^{s-|u|}} \frac{\partial^{|u|}}{\partial \mathbf{x}_u} f(\mathbf{x}) d\mathbf{x}_{-u} \right) \\ &\quad \times \left(\int_{[0,1]^{s-|u|}} \frac{\partial^{|u|}}{\partial \mathbf{x}_u} g(\mathbf{x}) d\mathbf{x}_{-u} \right) d\mathbf{x}_u, \end{aligned}$$

where $d\mathbf{x}_u := \prod_{j \in u} dx_j$ and $d\mathbf{x}_{-u} := \prod_{j \in \{1:s\} \setminus u} dx_j$.

Remarks

- We sum over all 2^s possible subsets of the indices $\{1 : s\}$. By convention, an empty product is 1.
- Each term of the sum corresponds to a subset of variables $\mathbf{x}_{\mathfrak{u}} = \{x_j \mid j \in \mathfrak{u}\}$. We refer to these as the “active” variables, and denote the remaining “inactive” variables by $\mathbf{x}_{-\mathfrak{u}}$.
- The cardinality $|\mathfrak{u}|$ of the set \mathfrak{u} is referred to as the “order” of the subset of variables $\mathbf{x}_{\mathfrak{u}}$. There is a *weight* parameter $\gamma_{\mathfrak{u}}$ associated with every subset of variables $\mathbf{x}_{\mathfrak{u}}$. The weights together model the relative importance between different subsets of variables. A small weight $\gamma_{\mathfrak{u}}$ means that the L^2 norm of $\frac{\partial^{|\mathfrak{u}|} f}{\partial \mathbf{x}_{\mathfrak{u}}}$ must also be small.
- Note that $\|\cdot\|_{s,\gamma}$ and $\|\cdot\|_{s,c\gamma}$ are equivalent norms for any $c > 0$.[†] Therefore we do not lose any generality by assuming that the weights have been normalized s.t. $\gamma_{\emptyset} = 1$. WLOG, we will always use the convention that $\gamma_{\emptyset} := 1$.

[†]Here, $c\gamma = (c\gamma_{\mathfrak{u}})_{\mathfrak{u} \subseteq \{1:s\}}$.

Special forms of weights

- *Product weights*: we have a sequence of numbers satisfying $\gamma_1 \geq \gamma_2 \geq \dots$ and we take

$$\gamma_{\mathbf{u}} = \prod_{j \in \mathbf{u}} \gamma_j.$$

In this case, the reproducing kernel is given by the product

$$K_{s,\gamma}(\mathbf{x}, \mathbf{y}) = \prod_{j \in \mathbf{u}} \left(1 + \gamma_j \eta(x_j, y_j) \right).$$

- *Finite order weights*: there exists $q \in \mathbb{N}$ s.t. $\gamma_{\mathbf{u}} = 0$ for all $|\mathbf{u}| > q$.
- *Order dependent weights*: we have a sequence of numbers $\Gamma_1, \Gamma_2, \dots$, and take

$$\gamma_{\mathbf{u}} = \Gamma_{|\mathbf{u}|}.$$

- *Product-and-order dependent (POD) weights*: we have two sequences $\gamma_1, \gamma_2, \dots$ and $\Gamma_1, \Gamma_2, \dots$, and take

$$\gamma_{\mathbf{u}} = \Gamma_{|\mathbf{u}|} \prod_{j \in \mathbf{u}} \gamma_j.$$

Why weighted spaces are interesting

Theorem (Sloan and Woźniakowski 1998)

Consider $H_{s,\gamma}$ equipped with product weights $\gamma_u = \prod_{j \in u} \gamma_j$. Then there exist point sets $P_n \subset [0, 1]^s$ for $n = 1, 2, \dots$ such that the worst-case error $e_{n,s}(P_n; H_{s,\gamma})$ is bounded independently of s if and only if

$$\sum_{j=1}^{\infty} \gamma_j < \infty. \tag{5}$$

To be more precise, the result has two parts:

- If condition (5) does *not* hold, then no matter how the points are chosen, the worst-case error is unbounded as $s \rightarrow \infty$.
- However, if (5) holds, then “good points” exist (although the result does not say how to find them).

Recall that $H_{s,\gamma}$ is defined via the reproducing kernel

$$K_{s,\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} \eta(x_j, y_j), \quad \eta(x, y) := \frac{1}{2}B_2(|x-y|) + (x - \frac{1}{2})(y - \frac{1}{2}),$$

where $B_2(x) := x^2 - x + \frac{1}{6}$ is the Bernoulli polynomial of degree 2.

Lemma

$$\int_{[0,1]^s} K_{s,\gamma}(\mathbf{x}, \mathbf{y}) d\mathbf{y} = 1,$$

$$\int_{[0,1]^s} \int_{[0,1]^s} K_{s,\gamma}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = 1,$$

$$\int_{[0,1]^s} K_{s,\gamma}(\mathbf{x}, \mathbf{x}) d\mathbf{x} = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} (\frac{1}{6})^{|\mathfrak{u}|}.$$

Proof. Left as an exercise. □



Recall that $H_{s,\gamma}$ is defined via the reproducing kernel

$$K_{s,\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} \eta(x_j, y_j), \quad \eta(x, y) := \frac{1}{2}B_2(|x-y|) + (x - \frac{1}{2})(y - \frac{1}{2}),$$

where $B_2(x) := x^2 - x + \frac{1}{6}$ is the Bernoulli polynomial of degree 2.

For our analysis, we will need the shift-invariant kernel associated with $K_{s,\gamma}$.

Lemma

$$\begin{aligned} K_{s,\gamma}^{\text{sh}}(\mathbf{x}, \mathbf{y}) &:= \int_{[0,1]^s} K_{s,\gamma}(\{\mathbf{x} + \boldsymbol{\Delta}\}, \{\mathbf{y} + \boldsymbol{\Delta}\}) d\boldsymbol{\Delta} \\ &= \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} B_2(|x_j - y_j|). \end{aligned}$$

Proof. This is an immediate consequence of

$$\int_0^1 \eta(\{x + \Delta\}, \{y + \Delta\}) d\Delta = B_2(|x - y|). \quad \square$$

Let

$$P = \left\{ \left\{ \frac{k\mathbf{z}}{n} \right\} \mid k = 0, \dots, n-1 \right\}$$

be a rank-1 lattice point set corresponding to generating vector $\mathbf{z} \in \mathbb{N}^s$ and $n \in \mathbb{N}$.

When dealing with the shift-invariant kernel corresponding to the unanchored, weighted Sobolev space $H_{s,\gamma}$, we use the shorthand notation

$$e_{n,s}^{\text{sh}}(\mathbf{z}) := e_{n,s}^{\text{sh}}(P; H_{s,\gamma}).$$

Lemma

The shift-averaged worst-case error for a rank-1 lattice rule in the weighted unanchored Sobolev space satisfies

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

Proof. Let $\mathbf{t}_j = \left\{ \frac{j \mathbf{z}}{n} \right\}$. We have the kernel

$$K_{s,\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} \eta(x_j, y_j), \quad \eta(x, y) := \frac{1}{2} B_2(|x-y|) + (x - \frac{1}{2})(y - \frac{1}{2}),$$

which satisfies $\int_{[0,1]^s} \int_{[0,1]^s} K_{s,\gamma}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = 1$. We showed that the shift-invariant kernel related to K is given by

$$K_{s,\gamma}^{\text{sh}}(\mathbf{x}, \mathbf{y}) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{k \in \mathfrak{u}} B_2(|x_k - y_k|).$$

Moreover, we showed that the shift-averaged WCE is given by

$$[e_{n,s}^{\text{sh}}(\mathbf{z})]^2 = - \int_{[0,1]^s} \int_{[0,1]^s} K_{s,\gamma}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K_{s,\gamma}^{\text{sh}}(\mathbf{t}_i, \mathbf{t}_j).$$

Making the obvious substitutions, we arrive at

$$\begin{aligned}[e_{n,s}^{\text{sh}}(\mathbf{z})]^2 &= -1 + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{k \in \mathfrak{u}} B_2 \left(\left\{ \frac{(i-j)z_k}{n} \right\} \right) \quad (\gamma_{\emptyset} := 1) \\ &= \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{k \in \mathfrak{u}} B_2 \left(\left\{ \frac{\text{mod}(i-j, n)z_k}{n} \right\} \right).\end{aligned}$$

As i and j range from 0 to $n - 1$, the values of $\text{mod}(i - j, n)$ are just $0, \dots, n - 1$ in some order (see next slide for illustration), with each value occurring n times. Thus the double sum can be reduced into a single sum:

$$[e_{n,s}^{\text{sh}}(\mathbf{z})]^2 = \frac{1}{n} \sum_{\ell=0}^{n-1} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{k \in \mathfrak{u}} B_2 \left(\left\{ \frac{\ell z_k}{n} \right\} \right),$$

as desired. □



An illustration of the counting argument used on the previous slide

i/j	0	1	2	3	4	\dots	$n-1$
0	0	1	2	3	4	\dots	$n-1$
1	$n-1$	0	1	2	3	\dots	$n-2$
2	$n-2$	$n-1$	0	1	2	\dots	$n-3$
3	$n-3$	$n-2$	$n-1$	0	1	\dots	$n-4$
4	$n-4$	$n-3$	$n-2$	$n-1$	0	\dots	$n-5$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
$n-1$	1	2	3	4	5	\dots	0

Table of the values $\text{mod}(i - j, n)$, when $i, j \in \{0, 1, \dots, n-1\}$.

By a simple counting argument we can write

$$\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} f(\text{mod}(i - j, n)) = n \sum_{\ell=0}^{n-1} f(\ell)$$

for any function $f: \{0, 1, \dots, n-1\} \rightarrow \mathbb{R}$.

Two easy technical results

Lemma (Fourier expansion of the Bernoulli polynomial B_2)

$$B_2(x) = \frac{1}{2\pi^2} \sum_{h \in \mathbb{Z} \setminus \{0\}} \frac{e^{2\pi i h x}}{h^2} \quad \text{for } x \in [0, 1].$$

Proof. Short argument: let $F(x) = \frac{1}{2\pi^2} \sum_{h \in \mathbb{Z} \setminus \{0\}} \frac{e^{2\pi i h x}}{h^2}$. Now[†] $F'(x) = \frac{i}{\pi} \sum_{h \in \mathbb{Z} \setminus \{0\}} \frac{e^{2\pi i h x}}{h} = 2x - 1$, so $F(x) = x^2 - x + c_0$. Moreover, $F(0) = \frac{1}{6}$, so $c_0 = \frac{1}{6}$. Hence $F(x) = x^2 - x + \frac{1}{6} = B_2(x)$. □

Lemma (“Jensen-like” inequality)

$$\sum_{k=1}^{\infty} a_k \leq \left(\sum_{k=1}^{\infty} a_k^{\lambda} \right)^{1/\lambda}, \quad a_k \geq 0, \lambda \in (0, 1].$$

Proof. Suppose that $\sum_{k=1}^{\infty} a_k^{\lambda} = 1$. Then $a_k \leq 1 \Rightarrow a_k \leq a_k^{\lambda}$ $\Rightarrow \sum_{k=1}^{\infty} a_k \leq \sum_{k=1}^{\infty} a_k^{\lambda} = 1$, and hence $\sum_{k=1}^{\infty} a_k \leq (\sum_{k=1}^{\infty} a_k^{\lambda})^{1/\lambda}$. The general case $\sum_{k=1}^{\infty} a_k^{\lambda} = C \in \mathbb{R}_+$ follows by applying the same argument for the scaled sequence $a_k \leftarrow \frac{1}{C^{1/\lambda}} a_k$. □

[†] F is absolutely convergent, so exchanging differentiation and summation is OK.

Component-by-component construction

The components of the generating vector \mathbf{z} can be restricted to the set

$$\mathbb{U}_n := \{z \in \mathbb{Z} \mid 1 \leq z \leq n \text{ and } \gcd(z, n) = 1\},$$

whose cardinality is given by the Euler totient function $\varphi(n) := |\mathbb{U}_n|$. When n is prime, $\varphi(n)$ takes its largest value $n - 1$.

We know that for $f \in H_{s,\gamma}$, there holds

$$\sqrt{\mathbb{E}_{\Delta} |I_s f - Q_{n,s}^{\Delta} f|^2} \leq e_{n,s}^{\text{sh}}(\mathbf{z}) \|f\|_{s,\gamma}.$$

Finding $\mathbf{z}^* = \arg \min_{\mathbf{z} \in \mathbb{U}_n} e_{n,s}^{\text{sh}}(\mathbf{z})$ is not computationally feasible: the search space contains altogether up to $(n - 1)^s$ possible choices for \mathbf{z} . However, the *component-by-component (CBC) construction* provides a feasible way to obtain good lattice generating vectors.

CBC construction

CBC construction. Given n, s , and weights $(\gamma_u)_{u \subseteq \{1:s\}}$.

1. Set $z_1 = 1$.
2. For $k = 2, 3, \dots, s$, choose $z_k \in \mathbb{U}_n$ to minimize $[e_{n,k}^{\text{sh}}(z_1, \dots, z_k)]^2$.

Remarks:

- Note that we have the (in principle computable) expression

$$[e_{n,k}^{\text{sh}}(\mathbf{z})]^2 = \frac{1}{n} \sum_{\emptyset \neq u \subseteq \{1:k\}} \gamma_u \sum_{\ell=0}^{n-1} \prod_{j \in u} B_2 \left(\left\{ \frac{\ell z_j}{n} \right\} \right). \quad (6)$$

- We will show that when the weights $(\gamma_u)_{u \subseteq \{1:s\}}$ are so-called *product-and-order dependent (POD)* weights, i.e., they can be written in the form

$$\gamma_u := \Gamma_{|u|} \prod_{j \in u} \gamma_j, \quad u \subseteq \{1 : s\},$$

where $\gamma_\emptyset := 1$, $(\Gamma_k)_{k=1}^\infty$ and $(\gamma_j)_{j=1}^\infty$ are sequences of positive numbers, then the value of (6) can be obtained in $\mathcal{O}(s n \log n + s^2 n)$ time using the so-called *fast CBC algorithm*. This is quadratic, not exponential, w.r.t. the dimension s .

- The CBC algorithm is a greedy algorithm: in general, it will **not** produce a generating vector which minimizes $e_{n,s}^{\text{sh}}(\mathbf{z})$. Regardless, we **can** produce an error estimate for the *QMC rule based on a generating vector constructed by the CBC algorithm!*

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Sixth lecture, May 19, 2025

Theorem (CBC error bound)

The generating vector $\mathbf{z} \in \mathbb{U}_n^s$ constructed by the CBC algorithm, minimizing the squared shift-averaged worst-case error $[e_{n,s}^{\text{sh}}(\mathbf{z})]^2$ for the weighted unanchored Sobolev space in each step, satisfies

$$[e_{n,s}^{\text{sh}}(\mathbf{z})]^2 \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq \mathbf{u} \subseteq \{1:s\}} \gamma_{\mathbf{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathbf{u}|} \right)^{1/\lambda} \quad \text{for all } \lambda \in (1/2, 1], \quad (1)$$

where $\zeta(x) := \sum_{k=1}^{\infty} k^{-x}$ denotes the Riemann zeta function for $x > 1$.

Proof. Step $s = 1$: by direct calculation, it is easy to see that

$[e_{n,1}^{\text{sh}}(z_1)]^2 = \frac{\gamma_1}{6n^2}$ and this is less than or equal to $\left(\frac{1}{\varphi(n)} \gamma_1^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right) \right)^{1/\lambda}$ for all $n \geq 1$, $\lambda \in (1/2, 1]$, and $\gamma_1 > 0$. Induction step: suppose that we have chosen the first $s - 1$ components z_1, \dots, z_{s-1} , and that (1) holds with s replaced by $s - 1$.

We can write the squared worst-case error in dimension-recursive form as

$$\begin{aligned}[e_{n,s}^{\text{sh}}(z_1, \dots, z_s)]^2 &= \frac{1}{n} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u \sum_{k=0}^{n-1} \prod_{j \in u} B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right) \\ &= [e_{n,s-1}^{\text{sh}}(z_1, \dots, z_{s-1})]^2 + \theta(z_1, \dots, z_{s-1}, z_s),\end{aligned}\tag{2}$$

where (suppressing the dependence of θ on z_1, \dots, z_{s-1})

$$\begin{aligned}\theta(z_s) &:= \sum_{s \in u \subseteq \{1:s\}} \gamma_u \left(\frac{1}{n} \sum_{k=0}^{n-1} \prod_{j \in u} B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right) \right) \quad (\text{use Fourier expansion of } B_2) \\ &= \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u}{(2\pi^2)^{|u|}} \left(\frac{1}{n} \sum_{k=0}^{n-1} \sum_{\mathbf{h}_u \in (\mathbb{Z} \setminus \{0\})^{|u|}} \frac{e^{2\pi i k \mathbf{h}_u \cdot \mathbf{z}_u / n}}{\prod_{j \in u} h_j^2} \right) \\ &= \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u}{(2\pi^2)^{|u|}} \left(\sum_{\substack{\mathbf{h}_u \in (\mathbb{Z} \setminus \{0\})^{|u|} \\ \mathbf{h}_u \cdot \mathbf{z}_u \equiv 0 \pmod{n}}} \frac{1}{\prod_{j \in u} h_j^2} \right),\end{aligned}$$

where we used the character property $\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i k \mathbf{h} \cdot \mathbf{z} / n} = \begin{cases} 1 & \text{if } \mathbf{h} \cdot \mathbf{z} \equiv 0 \pmod{n} \\ 0 & \text{otherwise} \end{cases}$.

Noting that $\mathbf{h}_u \cdot \mathbf{z}_u \equiv 0 \pmod{n}$ can be written equivalently as $\mathbf{h}_{u \setminus \{s\}} \cdot \mathbf{z}_{u \setminus \{s\}} \equiv -h_s z_s \pmod{n}$ for $s \in u \subseteq \{1:s\}$, we arrive at...

$$\theta(z_s) = \sum_{s \in \mathfrak{u} \subseteq \{1:s\}} \frac{\gamma_{\mathfrak{u}}}{(2\pi^2)^{|\mathfrak{u}|}} \left(\sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{h_s^2} \sum_{\substack{\mathbf{h}_{\mathfrak{u} \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|\mathfrak{u}|-1} \\ \mathbf{h}_{\mathfrak{u} \setminus \{s\}} \cdot \mathbf{z}_{\mathfrak{u} \setminus \{s\}} \equiv -h_s z_s \pmod{n}}} \frac{1}{\prod_{j \in \mathfrak{u} \setminus \{s\}} h_j^2} \right)$$

If z_s^* denotes the value chosen by the CBC algorithm in dimension s , then we use the following principle:

Averaging argument: *The minimum is always smaller than or equal to the average.*

In particular, this implies for all $\lambda \in (0, 1]$ that

$$\begin{aligned} [\theta(z_s^*)]^\lambda &\leq \frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} [\theta(z_s)]^\lambda \\ &\leq \frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} \left[\sum_{s \in \mathfrak{u} \subseteq \{1:s\}} \frac{\gamma_{\mathfrak{u}}}{(2\pi^2)^{|\mathfrak{u}|}} \left(\sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{h_s^2} \sum_{\substack{\mathbf{h}_{\mathfrak{u} \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|\mathfrak{u}|-1} \\ \mathbf{h}_{\mathfrak{u} \setminus \{s\}} \cdot \mathbf{z}_{\mathfrak{u} \setminus \{s\}} \equiv -h_s z_s \pmod{n}}} \frac{1}{\prod_{j \in \mathfrak{u} \setminus \{s\}} h_j^2} \right) \right]^\lambda \\ &\leq \frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} \sum_{s \in \mathfrak{u} \subseteq \{1:s\}} \frac{\gamma_{\mathfrak{u}}^\lambda}{(2\pi^2)^{|\mathfrak{u}|\lambda}} \sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{|h_s|^{2\lambda}} \sum_{\substack{\mathbf{h}_{\mathfrak{u} \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|\mathfrak{u}|-1} \\ \mathbf{h}_{\mathfrak{u} \setminus \{s\}} \cdot \mathbf{z}_{\mathfrak{u} \setminus \{s\}} \equiv -h_s z_s \pmod{n}}} \frac{1}{\prod_{j \in \mathfrak{u} \setminus \{s\}} |h_j|^{2\lambda}}, \end{aligned}$$

where we used the inequality $(\sum_k a_k)^\lambda \leq \sum_k a_k^\lambda$, $a_k \geq 0$, $\lambda \in (0, 1]$.

We separate the terms depending on whether or not h_s is a multiple of n . Note that this means

$$\begin{aligned} \sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{|h_s|^{2\lambda}} &= \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \frac{1}{|kn|^{2\lambda}} + \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \not\equiv 0 \pmod{n}}} \frac{1}{|h_s|^{2\lambda}} \\ &= \frac{2\zeta(2\lambda)}{n^{2\lambda}} + \sum_{c=1}^{n-1} \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \equiv c \pmod{n}}} \frac{1}{|h_s|^{2\lambda}}. \end{aligned}$$

It will be convenient to carry out a change of variable to eliminate the dependence on h_s from the innermost sum on the previous slide. Denote by z_s^{-1} the multiplicative inverse of z_s in \mathbb{U}_n , i.e., $z_s z_s^{-1} \equiv 1 \pmod{n}$. Then

$$\begin{aligned} &\frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u^\lambda}{(2\pi^2)^{|u|\lambda}} \sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{|h_s|^{2\lambda}} \sum_{\substack{h_{u \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|u|-1} \\ h_{u \setminus \{s\}} \cdot z_{u \setminus \{s\}} \equiv -h_s z_s \pmod{n}}} \frac{1}{\prod_{j \in u \setminus \{s\}} |h_j|^{2\lambda}} \\ &= \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u^\lambda}{(2\pi^2)^{|u|\lambda}} \frac{2\zeta(2\lambda)}{n^{2\lambda}} \sum_{\substack{h_{u \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|u|-1} \\ h_{u \setminus \{s\}} \cdot z_{u \setminus \{s\}} \equiv 0 \pmod{n}}} \frac{1}{\prod_{j \in u \setminus \{s\}} |h_j|^{2\lambda}} \\ &+ \frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} \sum_{c=1}^{n-1} \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u^\lambda}{(2\pi^2)^{|u|\lambda}} \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \equiv c \pmod{n}}} \frac{1}{|h_s|^{2\lambda}} \sum_{\substack{h_{u \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|u|-1} \\ h_{u \setminus \{s\}} \cdot z_{u \setminus \{s\}} \equiv -cz_s \pmod{n}}} \frac{1}{\prod_{j \in u \setminus \{s\}} |h_j|^{2\lambda}}. \end{aligned}$$

We separate the terms depending on whether or not h_s is a multiple of n . Note that this means

$$\begin{aligned} \sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{|h_s|^{2\lambda}} &= \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \frac{1}{|kn|^{2\lambda}} + \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \not\equiv 0 \pmod{n}}} \frac{1}{|h_s|^{2\lambda}} \\ &= \frac{2\zeta(2\lambda)}{n^{2\lambda}} + \sum_{c=1}^{n-1} \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \equiv c \pmod{n}}} \frac{1}{|h_s|^{2\lambda}}. \end{aligned}$$

It will be convenient to carry out a change of variable to eliminate the dependence on h_s from the innermost sum on the previous slide. Denote by z_s^{-1} the multiplicative inverse of z_s in \mathbb{U}_n , i.e., $z_s z_s^{-1} \equiv 1 \pmod{n}$. Then

$$\begin{aligned} &\frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u^\lambda}{(2\pi^2)^{|u|\lambda}} \sum_{h_s \in \mathbb{Z} \setminus \{0\}} \frac{1}{|h_s|^{2\lambda}} \sum_{\substack{h_{u \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|u|-1} \\ h_{u \setminus \{s\}} \cdot z_{u \setminus \{s\}} \equiv -h_s z_s \pmod{n}}} \frac{1}{\prod_{j \in u \setminus \{s\}} |h_j|^{2\lambda}} \\ &= \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u^\lambda}{(2\pi^2)^{|u|\lambda}} \frac{2\zeta(2\lambda)}{n^{2\lambda}} \sum_{\substack{h_{u \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|u|-1} \\ h_{u \setminus \{s\}} \cdot z_{u \setminus \{s\}} \equiv 0 \pmod{n}}} \frac{1}{\prod_{j \in u \setminus \{s\}} |h_j|^{2\lambda}} \\ &+ \frac{1}{\varphi(n)} \sum_{z_s \in \mathbb{U}_n} \sum_{c=1}^{n-1} \sum_{s \in u \subseteq \{1:s\}} \frac{\gamma_u^\lambda}{(2\pi^2)^{|u|\lambda}} \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \equiv -cz_s^{-1} \pmod{n}}} \frac{1}{|h_s|^{2\lambda}} \sum_{\substack{h_{u \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|u|-1} \\ h_{u \setminus \{s\}} \cdot z_{u \setminus \{s\}} \equiv c \pmod{n}}} \frac{1}{\prod_{j \in u \setminus \{s\}} |h_j|^{2\lambda}}. \end{aligned}$$

For $c \in \{1, \dots, n-1\}$, $\{\text{mod}(cz_s^{-1}, n) : z_s \in \mathbb{U}_n\} = \{\text{mod}(cz, n) : z \in \mathbb{U}_n\}$ and $\gcd(c/g, n/g) = 1$ with $g = \gcd(c, n)$. We obtain

$$\begin{aligned}
& \sum_{z_s \in \mathbb{U}_n} \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \equiv -cz_s^{-1} \pmod{n}}} \frac{1}{|h_s|^{2\lambda}} = \sum_{z \in \mathbb{U}_n} \sum_{\substack{h_s \in \mathbb{Z} \setminus \{0\} \\ h_s \equiv -cz \pmod{n}}} \frac{1}{|h_s|^{2\lambda}} \\
&= \sum_{z \in \mathbb{U}_n} \sum_{m \in \mathbb{Z}} \frac{1}{|mn - cz|^{2\lambda}} \\
&= g^{-2\lambda} \sum_{z \in \mathbb{U}_n} \sum_{m \in \mathbb{Z}} \frac{1}{|m(n/g) - (c/g)z|^{2\lambda}} \\
&= g^{-2\lambda} \sum_{z \in \mathbb{U}_n} \sum_{\substack{h \in \mathbb{Z} \setminus \{0\} \\ h \equiv -(c/g)z \pmod{n/g}}} \frac{1}{|h|^{2\lambda}} \\
&\leq g^{-2\lambda} g \sum_{a=1}^{n/g-1} \sum_{\substack{h \in \mathbb{Z} \setminus \{0\} \\ h \equiv a \pmod{n/g}}} \frac{1}{|h|^{2\lambda}} \leq g^{1-2\lambda} \sum_{h \in \mathbb{Z} \setminus \{0\}} \frac{1}{|h|^{2\lambda}} \leq 2\zeta(2\lambda),
\end{aligned}$$

where the last step holds since $g \geq 1$ and $\lambda > 1/2$. (The condition $\lambda > 1/2$ is needed to ensure that $\zeta(2\lambda) < \infty$.)

Hence

$$\begin{aligned}
[\theta(z_s^*)]^\lambda &\leq \sum_{s \in \mathfrak{u} \subseteq \{1:s\}} \frac{\gamma_{\mathfrak{u}}^\lambda}{(2\pi^2)^{|\mathfrak{u}|\lambda}} \frac{2\zeta(2\lambda)}{n^{2\lambda}} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u} \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|\mathfrak{u}|-1} \\ \boldsymbol{h}_{\mathfrak{u} \setminus \{s\}} \cdot \mathbf{z}_{\mathfrak{u} \setminus \{s\}} \equiv 0 \pmod{n}}} \frac{1}{\prod_{j \in \mathfrak{u} \setminus \{s\}} |h_j|^{2\lambda}} \\
&+ \frac{1}{\varphi(n)} \sum_{s \in \mathfrak{u} \subseteq \{1:s\}} \frac{\gamma_{\mathfrak{u}}^\lambda}{(2\pi^2)^{|\mathfrak{u}|\lambda}} 2\zeta(2\lambda) \sum_{\substack{\boldsymbol{h}_{\mathfrak{u} \setminus \{s\}} \in (\mathbb{Z} \setminus \{0\})^{|\mathfrak{u}|-1} \\ \boldsymbol{h}_{\mathfrak{u} \setminus \{s\}} \cdot \mathbf{z}_{\mathfrak{u} \setminus \{s\}} \not\equiv 0 \pmod{n}}} \frac{1}{\prod_{j \in \mathfrak{u} \setminus \{s\}} |h_j|^{2\lambda}} \\
&\leq \frac{1}{\varphi(n)} \sum_{s \in \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|\mathfrak{u}|},
\end{aligned}$$

where we used $\frac{1}{n^{2\lambda}} \leq \frac{1}{\varphi(n)}$ for $n \geq 1$ and $\lambda \in (1/2, 1]$.[†]

[†] $\varphi(n) \leq n \leq n^{2\lambda} \Rightarrow \frac{1}{n^{2\lambda}} \leq \frac{1}{\varphi(n)}$.

Returning to our original dimension-wise decomposition (2), using the bound on $\theta(z_s^*)$ and the induction hypothesis yield

$$\begin{aligned}
 [e_{n,s}^{\text{sh}}(z_1, \dots, z_s)]^2 &= [e_{n,s-1}^{\text{sh}}(z_1, \dots, z_{s-1})]^2 + \theta(z_1, \dots, z_{s-1}, z_s) \\
 &\leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s-1\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} \right)^{1/\lambda} + \left(\frac{1}{\varphi(n)} \sum_{s \in u \subseteq \{1:s\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} \right)^{1/\lambda} \\
 &\leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s-1\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} + \frac{1}{\varphi(n)} \sum_{s \in u \subseteq \{1:s\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} \right)^{1/\lambda} \\
 &= \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} \right)^{1/\lambda},
 \end{aligned}$$

proving the assertion. □



Significance: Suppose that $f \in H_{s,\gamma}$ for all $\gamma = (\gamma_u)_{u \subseteq \{1:s\}}$. Then for any given sequence of weights γ , we can use the CBC algorithm to obtain a generating vector satisfying the error bound

$$\sqrt{\mathbb{E}_\Delta |I_s f - Q_{n,s}^\Delta f|^2} \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} \right)^{1/(2\lambda)} \|f\|_{s,\gamma} \quad (3)$$

for all $\lambda \in (1/2, 1]$. We can use the following strategy:

- For a given integrand f , estimate the norm $\|f\|_{s,\gamma}$.
- Find weights γ which *minimize* the error bound (3).
- Using the optimized weights γ as input, use the CBC algorithm to find a generating vector which *satisfies* the error bound (3).

Remarks:

- If n is prime, then $\frac{1}{\varphi(n)} = \frac{1}{n-1}$. If $n = 2^k$, then $\frac{1}{\varphi(n)} = \frac{2}{n}$. For general (composite) $n \geq 3$, $\frac{1}{\varphi(n)} \leq \frac{e^\gamma \log \log n + \frac{3}{\log \log n}}{n}$, where $\gamma = 0.57721566\dots$ (Euler–Mascheroni constant).
- The optimal convergence rate close to $\mathcal{O}(n^{-1})$ is obtained with $\lambda \rightarrow 1/2$, but note that $\lambda = 1/2$ is not permitted since $\zeta(2\lambda) \rightarrow \infty$ as $\lambda \rightarrow 1/2$.

Appendix

Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}_+$. Recall that

$$a \equiv b \pmod{m} \Leftrightarrow \frac{a - b}{m} \in \mathbb{Z} \Leftrightarrow a = km + b \text{ for some } k \in \mathbb{Z}.$$

Theorem (Bézout's identity)

Let $a, b \in \mathbb{Z}$. Then there exist $x, y \in \mathbb{Z}$ such that $ax + by = \gcd(a, b)$.

Corollary

Let $a, b \in \mathbb{Z}$ and $m \in \mathbb{Z}_+$.

- The linear congruence $ax \equiv b \pmod{m}$ has a solution if and only if $\gcd(a, m)|b$.
- If $\gcd(a, m)|b$, then there are exactly $\gcd(a, m)$ solutions modulo m to the linear congruence $ax \equiv b \pmod{m}$.

Let $z, n \in \mathbb{N}$ be such that $\gcd(z, n) = 1$. Then the above corollary implies that the linear congruence

$$zx \equiv 1 \pmod{n}$$

has exactly one solution (modulo n). This solution is called the *modular multiplicative inverse* and it is often denoted by $z^{-1} := x$.

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Seventh lecture, May 26, 2025

Let $f \in H_{s,\gamma}$, where we assume that the positive weights $\gamma := (\gamma_u)_{u \subseteq \{1:s\}}$ have the following *product-and-order dependent (POD)* form

$$\gamma_u := \Gamma_{|u|} \prod_{j \in u} \gamma_j, \quad u \subseteq \{1:s\},$$

where $(\Gamma_k)_{k=0}^s$ and $(\gamma_j)_{j=1}^s$ are positive numbers such that $\Gamma_0 = 1$ and the empty product is interpreted as 1.

A randomly shifted rank-1 lattice rule with generating vector $z \in \mathbb{N}^s$ satisfies the error bound

$$\sqrt{\mathbb{E}_\Delta |I_s f - Q_{n,s}^\Delta f|^2} \leq e_{n,s}^{\text{sh}}(z) \|f\|_{s,\gamma},$$

where the squared shift-averaged worst-case error in the weighted unanchored Sobolev space is given by the formula

$$[e_{n,s}^{\text{sh}}(z)]^2 = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u \prod_{j \in u} B_2\left(\left\{\frac{kz_j}{n}\right\}\right),$$

with $B_2(x) = x^2 - x + \frac{1}{6}$ denoting the Bernoulli polynomial of degree 2.

The components of the generating vector \mathbf{z} can be restricted to the set

$$\mathbb{U}_n := \{z \in \mathbb{Z} \mid 1 \leq z \leq n \text{ and } \gcd(z, n) = 1\},$$

whose cardinality is given by the Euler totient function $\varphi(n) := |\mathbb{U}_n|$.

Component-by-component (CBC) construction.

Given n , s , and weights $(\gamma_u)_{u \subseteq \{1:s\}}$, do

1. Set $z_1 = 1$.
2. With z_1 fixed, choose $z_2 \in \mathbb{U}_n$ to minimize $[e_{n,2}^{\text{sh}}(z_1, z_2)]^2$.
3. With z_1 and z_2 fixed, choose $z_3 \in \mathbb{U}_n$ to minimize $[e_{n,3}^{\text{sh}}(z_1, z_2, z_3)]^2$.
- ⋮

From the previous lecture, we know that the generating vector obtained using the CBC algorithm satisfies a certain *a priori* cubature error bound.

This week's lecture: *How to implement the CBC algorithm efficiently for POD weights and prime n ?*

Remark: The so-called POD weights arise in the context of elliptic PDEs with random coefficients (next week's lecture), hence our interest in weights having this abstract form.

Our strategy will be as follows:

- First, we will describe a computationally inefficient version of the CBC algorithm. This will serve as a basis for a more efficient implementation.
- We will address the computational bottlenecks inherent in the naïve implementation of the CBC algorithm in order to construct an implementation of the so-called *fast CBC algorithm*.

For the fast CBC algorithm, we will require some sophisticated mathematical machinery (specifically, an algorithm for computing a primitive root modulo n and carrying out circulant matrix-vector multiplication using the fast Fourier transform), which will be discussed later on.

Naïve CBC construction

We write the error criterion as

$$\begin{aligned}
 [e_{n,d}^{\text{sh}}(z_1, \dots, z_d)]^2 &= \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\emptyset \neq u \subseteq \{1:d\}} \gamma_u \prod_{j \in u} B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right) \\
 &= \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\ell=1}^d \underbrace{\sum_{\substack{|u|=\ell \\ u \subseteq \{1:d\}}} \gamma_u \prod_{j \in u} B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right)}_{=: p_{d,\ell}(k)}.
 \end{aligned}$$

By plugging in the POD weights $\gamma_u := \Gamma_{|u|} \prod_{j \in u} \gamma_j$, note that we have the following recursion (we split the sum over u in two parts depending on whether $d \in u$):

$$\begin{aligned}
 p_{d,\ell}(k) &= \sum_{\substack{|u|=\ell \\ u \subseteq \{1:d\}}} \Gamma_\ell \left(\prod_{j \in u} \gamma_j B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right) \right) \\
 &= \sum_{\substack{|u|=\ell \\ u \subseteq \{1:d-1\}}} \Gamma_\ell \left(\prod_{j \in u} \gamma_j B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right) \right) \\
 &\quad + \sum_{\substack{|u|=\ell-1 \\ u \subseteq \{1:d-1\}}} \Gamma_\ell \gamma_d B_2 \left(\left\{ \frac{kz_d}{n} \right\} \right) \left(\prod_{j \in u} \gamma_j B_2 \left(\left\{ \frac{kz_j}{n} \right\} \right) \right) \\
 &= p_{d-1,\ell}(k) + \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d B_2 \left(\left\{ \frac{kz_d}{n} \right\} \right) p_{d-1,\ell-1}(k).
 \end{aligned}$$

Plugging the recurrence

$$p_{d,\ell}(k) = p_{d-1,\ell}(k) + \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d B_2\left(\left\{ \frac{kz_d}{n} \right\}\right) p_{d-1,\ell-1}(k)$$

into the expression for the squared shift-averaged WCE yields

$$\begin{aligned} [e_{n,d}^{\text{sh}}(z_1, \dots, z_d)]^2 &= \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\ell=1}^d p_{d,\ell}(k) \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\ell=1}^d p_{d-1,\ell}(k) + \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\ell=1}^d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d B_2\left(\left\{ \frac{kz_d}{n} \right\}\right) p_{d-1,\ell-1}(k) \\ &= [e_{n,d-1}^{\text{sh}}(z_1, \dots, z_{d-1})]^2 + \frac{1}{n} \sum_{k=0}^{n-1} B_2\left(\left\{ \frac{kz_d}{n} \right\}\right) \sum_{\ell=1}^d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d p_{d-1,\ell-1}(k). \end{aligned}$$

Recall that in the d^{th} step of the CBC algorithm, the components z_1, \dots, z_{d-1} are fixed and it is therefore sufficient to find $z_d \in \mathbb{U}_n$ which minimizes the expression $\sum_{k=0}^{n-1} B_2\left(\left\{ \frac{kz_d}{n} \right\}\right) \sum_{\ell=1}^d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d p_{d-1,\ell-1}(k)$.

Let us introduce the matrix $\Omega_n := \left[B_2\left(\left\{ \frac{kz}{n} \right\} \right) \right]_{k \in \{0, \dots, n-1\}, z \in \mathbb{U}_n}$ and define a set of n -vectors recursively via

$$\mathbf{p}_{d,\ell} = \mathbf{p}_{d-1,\ell} + \gamma_d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \Omega_n(z_d, :) * \mathbf{p}_{d-1,\ell-1}$$

starting from the initial values

$$\mathbf{p}_{d,0} = \mathbf{1}_n \quad \text{for all } d \geq 1,$$

$$\mathbf{p}_{d,\ell} = \mathbf{0}_n \quad \text{for all } d \geq 1 \text{ and } \ell > d,$$

with $.*$ denoting the componentwise product between two vectors.

Then the value of $\sum_{k=0}^{n-1} B_2\left(\left\{ \frac{kz_d}{n} \right\} \right) \sum_{\ell=1}^d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d \mathbf{p}_{d-1,\ell-1}(k)$ in the d^{th} step of the CBC algorithm can be obtained for all $z_d \in \mathbb{U}_n$ via

$$\Omega_n \mathbf{x}, \quad \text{where } \mathbf{x} = \sum_{\ell=1}^d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d \mathbf{p}_{d-1,\ell-1}.$$

CBC algorithm – naïve version

1. Define the matrix $\Omega_n := \left[B_2\left(\left\{\frac{kz}{n}\right\}\right) \right]_{k \in \{0, \dots, n-1\}, z \in \mathbb{U}_n}$ and initialize the n -vectors

$$\mathbf{p}_{d,0} = \mathbf{1}_n \quad \text{for all } d \geq 1,$$

$$\mathbf{p}_{d,\ell} = \mathbf{0}_n \quad \text{for all } d \geq 1 \text{ and } \ell > d.$$

for $d = 1, \dots, s$, **do**

2. Pick the value $z_d \in \{1, \dots, n - 1\}$ corresponding to the smallest entry in the matrix-vector product

$$\Omega_n \mathbf{x}, \quad \text{where } \mathbf{x} = \sum_{\ell=1}^d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d \mathbf{p}_{d-1,\ell-1}. \quad (1)$$

3. Update $\mathbf{p}_{d,\ell} = \mathbf{p}_{d-1,\ell} + \gamma_d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \Omega_n(z_d, :) * \mathbf{p}_{d-1,\ell-1}$.

end for

Remarks: We only need the ratio $a_\ell := \frac{\Gamma_\ell}{\Gamma_{\ell-1}}$ for the implementation, e.g., for $\Gamma_\ell = \ell!$ this is $a_\ell = \ell$. The computational bottleneck is the dense matrix-vector product $\Omega_n \mathbf{x}$ in (1), which has complexity $\mathcal{O}(n^2)$. The *fast CBC algorithm* reduces this product down to $\mathcal{O}(n \log n)$ complexity.

Fast CBC algorithm

What makes fast CBC fast?

The matrix-vector product $\Omega_n \mathbf{x}$ has time complexity $\mathcal{O}(n^2)$, which is too slow if n is, say, of the order of a million or more. (Not to mention the problem of storing a dense matrix of such size!)

However, the matrix Ω_n has a lot of structure. It turns out that we can implement the matrix-vector product $\Omega_n \mathbf{x}$ in $\mathcal{O}(n \log n)$ time using some sophisticated mathematical tools.

In a nutshell, we let $n \geq 3$ be *prime* and do the following:

- Using some natural symmetries of Ω_n , we can ignore the first column (since it corresponds to shifting the objective functional in the CBC minimization step by a constant value) and it will be sufficient to consider only the top-left block $\Omega'_n := \Omega_n(1 : m, 2 : m + 1)$, where $m := (n - 1)/2$.
- For *prime* n , we can find a *generator* g (primitive root modulo n) and use this to permute Ω'_n into a circulant matrix.
- A circulant matrix implements a circular convolution, so a matrix-vector product (in the permuted indexing) can be implemented in $\mathcal{O}(n \log n)$ time using the fast Fourier transform (FFT).

Before getting to the implementational details of fast CBC, we will need to

- discuss an algorithm to find a primitive root modulo n ;
- discuss how to compute a circulant matrix-vector product using FFT.

Primitive root modulo n

Definition

Let $g, n \in \mathbb{N}$. The number g is called a *primitive root modulo n* if for any integer $a \in \mathbb{N}$ such that $\gcd(a, n) = 1$, there exists an integer k (called the *index*) such that

$$g^k \equiv a \pmod{n}.$$

Such a number g is the *generator* of the multiplicative group of integers modulo n , i.e., $(\mathbb{Z}/n\mathbb{Z})^\times$.

Theorem (Gauss 1801)

A primitive root modulo n exists if and only if

- n is 1, 2, 4, or
- $n = p^k$, where $p \geq 3$ is a prime and $k \in \mathbb{N}$, or
- $n = 2p^k$, where $p \geq 3$ is a prime and $k \in \mathbb{N}$.

Note especially that a primitive root modulo n exists whenever n is prime.

Recall that the Euler totient function is defined by
 $\varphi(n) := |\{k \in \mathbb{N} \mid 1 \leq k \leq n, \gcd(k, n) = 1\}|$. We have the following.

Proposition

The number g is a primitive root modulo n if and only if the smallest positive integer k for which $g^k \equiv 1 \pmod{n}$ is precisely $k = \varphi(n)$.

Lagrange's theorem: the smallest k satisfying $g^k \equiv 1 \pmod{n}$ divides $\varphi(n)$. Therefore, it is enough to check for all proper divisors $d|\varphi(n)$ that $g^d \not\equiv 1 \pmod{n}$.

However, we can do even better!

Find the prime number factorization $\varphi(n) = p_1^{a_1} \cdots p_\ell^{a_\ell}$. It turns out that it is enough to check that $g^d \not\equiv 1 \pmod{n}$ for all $d \in \left\{ \frac{\varphi(n)}{p_1}, \dots, \frac{\varphi(n)}{p_\ell} \right\}$. To see this, let d be any proper divisor of $\varphi(n)$. Then there exists j such that $d \mid \frac{\varphi(n)}{p_j}$, meaning that $dk = \frac{\varphi(n)}{p_j}$ for some $k \in \mathbb{N}$. However, if $g^d \equiv 1 \pmod{n}$, we would get

$$g^{\frac{\varphi(n)}{p_j}} \equiv g^{dk} \equiv (g^d)^k \equiv 1^k \equiv 1 \pmod{n}.$$

That is, if g was not a primitive root, then one could find a number of the form $\frac{\varphi(n)}{p_j}$ for which $g^{\frac{\varphi(n)}{p_j}} \equiv 1 \pmod{n}$.

\therefore It is enough to check that $g^{\frac{\phi(n)}{p_j}} \not\equiv 1 \pmod{n}$ for all $j \in \{1, \dots, \ell\}$.

Algorithm for finding a primitive root modulo n

1. Find the prime number factorization $\varphi(n) = p_1^{a_1} \cdots p_\ell^{a_\ell}$.

Iterate through all numbers $g = 1, 2, \dots, n - 1$ and, for each number, check whether it is a primitive root by doing the following:

2. Calculate $\text{mod}(g^{\frac{\varphi(n)}{p_j}}, n)$ for all $j \in \{1, \dots, \ell\}$.
3. If all the calculated values are different from 1, then g is a primitive root.

Remark: In Python, the quantities in step 2 can be computed, e.g., via `pow(g, sympy.totient(n)/pj, n)`

Discrete and fast Fourier transform

The *discrete Fourier transform* of (complex) vector $\mathbf{x} := (x_j)_{j=1}^n$ is defined as the vector $\mathbf{y} := (y_j)_{j=1}^n$ with

$$y_j = \sum_{k=1}^n x_k e^{-2\pi i(j-1)(k-1)/n}, \quad j \in \{1, \dots, n\},$$

and the *inverse discrete Fourier transform* is given by

$$x_j = \frac{1}{n} \sum_{k=1}^n y_k e^{2\pi i(j-1)(k-1)/n}, \quad j \in \{1, \dots, n\}.$$

The *fast Fourier transform (FFT)* can be used to carry out these operations in $\mathcal{O}(n \log n)$ time. In Python, one has $\mathbf{y} = \text{numpy.fft.fft}(\mathbf{x})$ and $\mathbf{x} = \text{numpy.fft.ifft}(\mathbf{y})$.

Circular convolution

Let $\mathbf{x} := (x_i)_{i=1}^n$ and $\mathbf{y} := (y_i)_{i=1}^n$ be (complex) vectors. Then the sequence $\mathbf{z} := (z_i)_{i=1}^n$ defined by

$$z_i = \sum_{k=1}^n x_k y_{\text{mod}(i-k, n)+1}, \quad i \in \{1, \dots, n\},$$

is called the *circular convolution* of \mathbf{x} and \mathbf{y} and we denote it by $\mathbf{z} := \mathbf{x} \star \mathbf{y}$.

Similarly to the continuous convolution, we have the following identity using discrete/fast Fourier transform:

$$\text{fft}(\mathbf{x} \star \mathbf{y}) = \text{fft}(\mathbf{x}) \cdot \text{fft}(\mathbf{y}),$$

where $\mathbf{x} \cdot \mathbf{y} := (x_i y_i)_{i=1}^n$ is the pointwise product of two vectors.

Circular convolution and circulant matrices

A matrix $A \in \mathbb{R}^{n \times n}$ is called *circulant* if it has the form

$$A = \begin{bmatrix} a_0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_1 & a_0 & a_{n-1} & & a_2 \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & & \ddots & \ddots & a_{n-1} \\ a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 \end{bmatrix}.$$

- Each row is equal to the row above shifted to the right by one (wrapping around the edge in a periodic way).
- The first column/row contains all information about the matrix.
- A circulant matrix implements a circular convolution:

$$Ax = \mathbf{a} \star \mathbf{x}, \tag{2}$$

where $\mathbf{a} := [a_0, a_1, \dots, a_{n-1}]^T$ is the first column of matrix A .

- The identity (2) implies that a circulant matrix-vector product can be implemented in $\mathcal{O}(n \log n)$ time as $A\mathbf{x} = \text{ifft}(\text{fft}(\mathbf{a}) \cdot \text{fft}(\mathbf{x}))$.

Putting it all together

The matrix-vector product $\Omega_n \mathbf{x}$ in the CBC loop costs $\mathcal{O}(n^2)$ operations. However, it was shown by Kuo, Nuyens, and Cools (2006) that the blocks of Ω_n can be permuted into circulant form → the matrix-vector product can be implemented in $\mathcal{O}(n \log n)$ operations using FFT.

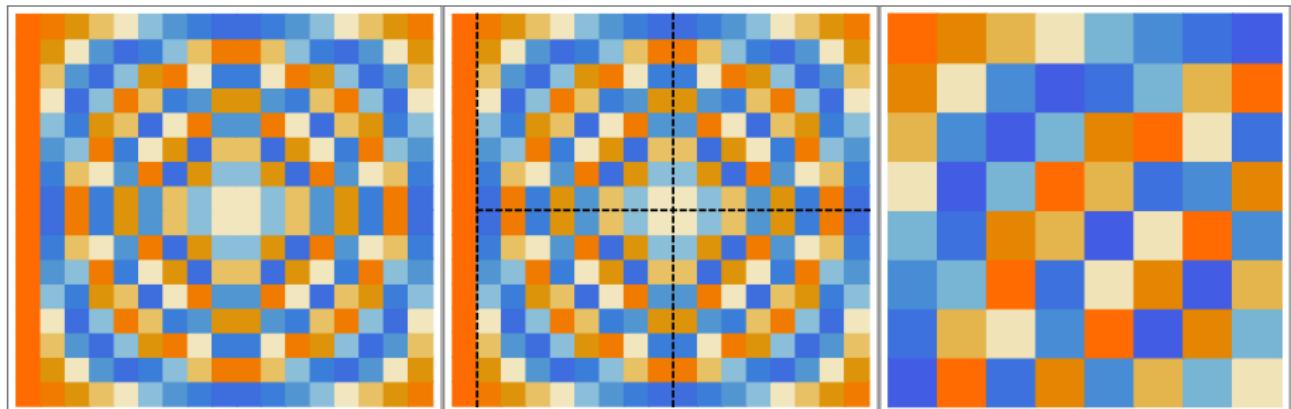


Figure: Example with Ω_{17} . Note that the first column is a constant and can be left out (the components of $\Omega_n \mathbf{x}$ are shifted by a constant → the smallest component stays invariant). Noting the obvious symmetries in the remaining four blocks, we can focus on the top left block.

When n is prime, it is possible to use the so-called Rader transformation to permute the block matrices into circulant form. The permutation matrices can be easily obtained by computing the generator, i.e., primitive root modulo n .

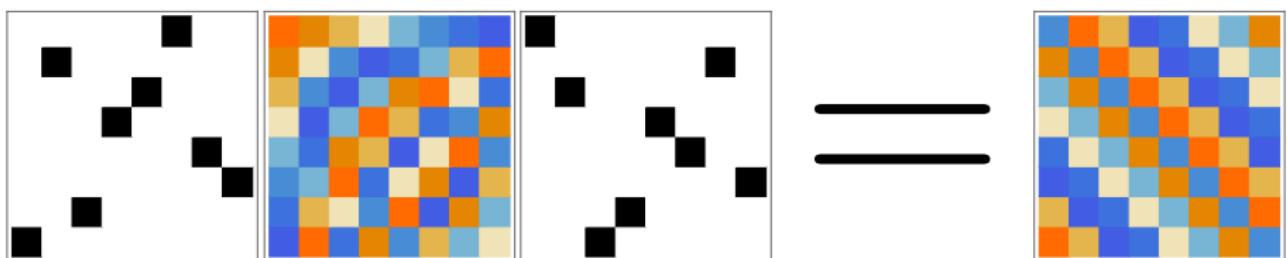


Figure: The original block matrix is multiplied from both sides by Rader permutation matrices (the black elements indicate the value 1 and white elements indicate the value 0) to obtain a circulant matrix.

Example with $n = 1009$

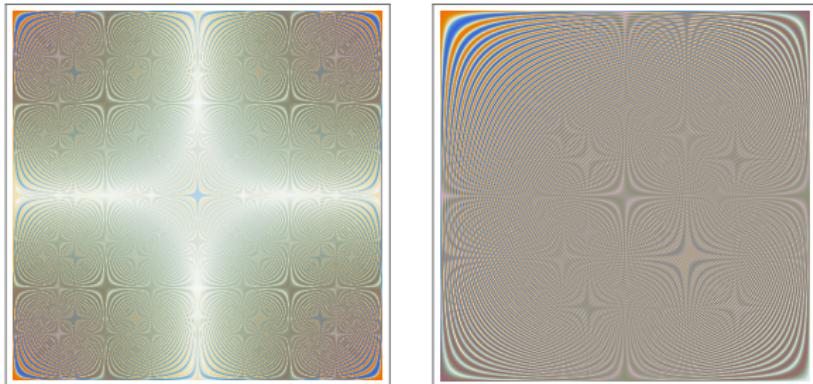


Figure: LHS: Original Ω_{1009} . RHS: top left block of Ω_{1009} (sans first column).

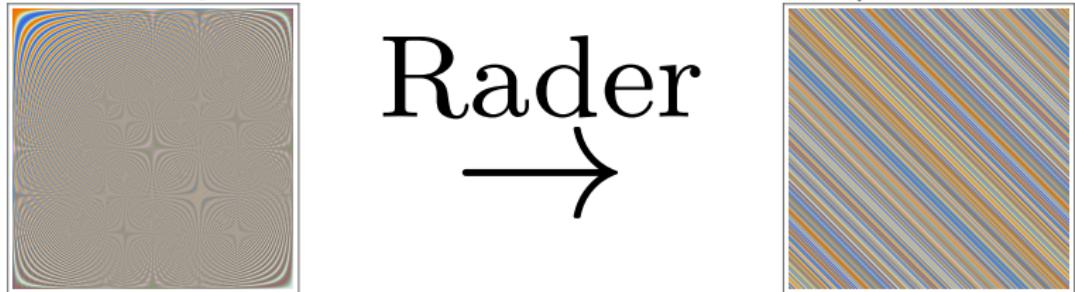


Figure: Rader transformation turns the top left block matrix circulant.

Python implementation given in the file `fastcbc.py` available on the course webpage!

- The overall cost of the CBC algorithm with POD weights is $\mathcal{O}(s n \log n + s^2 n)$.
- For simplicity, we considered only the case where n is prime. An extension for composite n was discussed by Nuyens and Cools (J. Complexity 2006). The idea for composite n is that the complete matrix Ω_n can be partitioned in blocks which have a circulant or block-circulant structure. The special case of n being a power of 2 has been discussed by Cools, Kuo, and Nuyens (SIAM J. Sci. Comput. 2006).
- There also exist freely available software implementing the fast CBC construction, cf., e.g.,

<https://people.cs.kuleuven.be/~dirk.nuyens/qmc4pde/>,
<https://people.cs.kuleuven.be/~dirk.nuyens/fast-cbc/>,
<https://qmcpy.org/>, ...

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eighth lecture, June 16, 2025

Recap: Suppose that $f \in H_{s,\gamma}$ for all $\gamma = (\gamma_u)_{u \subseteq \{1:s\}}$. The unanchored, weighted Sobolev space $H_{s,\gamma}$ is equipped with the norm

$$\|f\|_{s,\gamma}^2 := \sum_{u \subseteq \{1:s\}} \frac{1}{\gamma_u} \int_{[0,1]^{|u|}} \left(\int_{[0,1]^{s-|u|}} \frac{\partial^{|u|}}{\partial \mathbf{y}_u} f(\mathbf{y}) d\mathbf{y}_{-u} \right)^2 d\mathbf{y}_u.$$

For any given sequence of weights γ , we can use the CBC algorithm (*implementational details were considered during the 7th lecture*) to obtain a generating vector for a randomly shifted rank-1 lattice QMC rule satisfying the error bound

$$\sqrt{\mathbb{E}_{\Delta} |I_s f - Q_{n,s}^{\Delta} f|^2} \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|u|} \right)^{1/(2\lambda)} \|f\|_{s,\gamma} \quad (1)$$

for all $\lambda \in (1/2, 1]$. We can use the following strategy:

- For a given integrand f , estimate the norm $\|f\|_{s,\gamma}$.
- Find weights γ which *minimize* the error bound (1).
- Using the optimized weights γ as input, use the CBC algorithm to find a generating vector which *satisfies* the error bound (1).

Application to parametric PDE problems

For the application of QMC methods to parametric PDE problems, we follow the survey papers

-  F. Y. Kuo and D. Nuyens. Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients - a survey of analysis and implementation. *Found. Comput. Math.* **16**:1631–1696, 2016. arXiv version: <https://arxiv.org/abs/1606.06613>
-  F. Y. Kuo and D. Nuyens. Application of quasi-Monte Carlo methods to PDEs with random coefficients – an overview and tutorial. In: A. Owen and P. Glynn (eds), *Monte Carlo and Quasi-Monte Carlo Methods 2016*, pp. 53–71. arXiv version:
<https://arxiv.org/abs/1710.10984>

Let us first consider applying QMC for the uniform and affine model problem discussed during week 4.

Recall 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 $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient has the parameterization

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

where $a_0 \in L^\infty(D)$, there exist $a_{\min}, a_{\max} > 0$

s.t. $0 < a_{\min} \leq a(\mathbf{x}, \mathbf{y}) \leq a_{\max} < \infty$ for all $\mathbf{x} \in D$ and $\mathbf{y} \in U$, and the *stochastic fluctuations* $\psi_j: D \rightarrow \mathbb{R}$ are functions of the spatial variable such that

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

Total error decomposition

In practice, we need to truncate the infinite-dimensional parametric vector $\mathbf{y} \in [-1/2, 1/2]^{\mathbb{N}}$ to a finite number of terms. Moreover, the PDE needs to be discretized spatially using, e.g., the finite element method.

Let $u_s(\mathbf{y}) := u_s(y_1, \dots, y_s, 0, 0, \dots)$ denote the dimensionally-truncated PDE solution for $\mathbf{y} \in [-1/2, 1/2]^{\mathbb{N}}$ (we often abuse notation and also write $u_s(\mathbf{y})$ for $\mathbf{y} \in [-1/2, 1/2]^s$), and let $u_{s,h}(\cdot, \mathbf{y}) \in V_h$ denote the dimensionally-truncated FE solution in the FE subspace spanned by piecewise linear FE basis functions. Furthermore, let $\{\mathbf{t}_i\}_{i=1}^n$ be a QMC point set in $[-1/2, 1/2]^s$.

Total error decomposition

For simplicity, let us consider the problem of computing $\mathbb{E}[G(u)]$, where $u(\cdot, \mathbf{y}) \in H_0^1(D)$ is the PDE solution for $\mathbf{y} \in U$ and $G: H_0^1(D) \rightarrow \mathbb{R}$ is a linear functional (quantity of interest). We decompose the total error as

$$\begin{aligned}& \int_{[-1/2,1/2]^{\mathbb{N}}} G(u(\cdot, \mathbf{y})) d\mathbf{y} - \frac{1}{n} \sum_{i=1}^n G(u_{s,h}(\cdot, \mathbf{t}_i)) \\&= \int_{[-1/2,1/2]^{\mathbb{N}}} (G(u(\cdot, \mathbf{y}) - u_s(\cdot, \mathbf{y}))) d\mathbf{y} \\&+ \int_{[-1/2,1/2]^s} G(u_s(\cdot, \mathbf{y}) - u_{s,h}(\cdot, \mathbf{y})) d\mathbf{y} \\&+ \int_{[-1/2,1/2]^s} G(u_{s,h}(\cdot, \mathbf{y})) d\mathbf{y} - \frac{1}{n} \sum_{i=1}^n G(u_{s,h}(\cdot, \mathbf{t}_i)).\end{aligned}$$

Using the triangle inequality, we are left with the total error decomposition

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

Let us focus today on the cubature error.

Remarks:

- We'll discuss the other error contributions (dimension truncation and finite element errors) later. Furthermore, we'll see how the analysis differs in the lognormal setting.
- It turns out that if we can control the error for all linear quantities of interest $G : H_0^1(D) \rightarrow \mathbb{R}$, we can control the error for the full PDE solution with respect to the $\|\cdot\|_{H_0^1(D)}$ norm using a duality argument.

Multi-index notation

We introduce the set of *finitely-supported multi-indices*

$$\mathcal{F} := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\text{supp}(\boldsymbol{\nu})| < \infty\},$$

where the *support* of a multi-index $\boldsymbol{\nu}$ is defined as the set

$$\text{supp}(\boldsymbol{\nu}) := \{i \in \mathbb{N} : \nu_i \neq 0\}.$$

As before, the *order* of a multi-index is defined as

$$|\boldsymbol{\nu}| := \sum_{j \geq 1} \nu_j$$

and we use the special multi-index notations

$$\partial^{\boldsymbol{\nu}} := \partial_{\mathbf{y}}^{\boldsymbol{\nu}} := \prod_{j \in \text{supp}(\boldsymbol{\nu})} \frac{\partial^{\nu_j}}{\partial y_j^{\nu_j}}, \quad \mathbf{x}^{\boldsymbol{\nu}} := \prod_{j \in \text{supp}(\boldsymbol{\nu})} x_j^{\nu_j}, \quad \binom{\boldsymbol{\nu}}{\mathbf{m}} := \prod_{j \in \text{supp}(\boldsymbol{\nu})} \binom{\nu_j}{m_j}.$$

Recursive bound

Consider the weak formulation

$$\int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}. \quad (2)$$

Noting that

$$\partial^\nu a(\mathbf{x}, \mathbf{y}) = \begin{cases} a(\mathbf{x}, \mathbf{y}) & \text{if } \nu = \mathbf{0}, \\ \psi_j(\mathbf{x}) & \text{if } \nu = \mathbf{e}_j, \\ 0 & \text{otherwise,} \end{cases}$$

then by differentiating (2) on both sides with ∂^ν and using the Leibniz product rule[†] yields

$$\partial^\nu \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} = 0$$

$$\Leftrightarrow \sum_{\mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \int_D \partial^\mathbf{m} a(\mathbf{x}) \nabla \partial^{\nu-\mathbf{m}} u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} = 0$$

$$\Leftrightarrow \int_D a(\mathbf{x}, \mathbf{y}) \nabla \partial^\nu u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} = - \sum_{j \in \text{supp}(\nu)} \nu_j \int_D \psi_j(\mathbf{x}) \nabla \partial^{\nu - \mathbf{e}_j} u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x}.$$

[†] $\partial^\nu(fg) = \sum_{\mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \partial^\mathbf{m} f \partial^{\nu-\mathbf{m}} g$ (exercise)

Testing this against $v = \partial^\nu u(\mathbf{x}, \mathbf{y})$ yields

$$\begin{aligned} & a_{\min} \|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)}^2 \\ & \leq \int_D a(\mathbf{x}, \mathbf{y}) \|\nabla \partial^\nu u(\mathbf{x}, \mathbf{y})\|^2 d\mathbf{x} \\ & \leq \sum_{j \in \text{supp}(\nu)} \nu_j \|\psi_j\|_{L^\infty(D)} \|\partial^{\nu - e_j} u(\cdot, \mathbf{y})\|_{H_0^1(D)} \|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} \end{aligned}$$

Thus we obtain the recursive relation

$$\|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} \leq \sum_{j \in \text{supp}(\nu)} \nu_j \underbrace{\frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}}_{=: b_j} \|\partial^{\nu - e_j} u(\cdot, \mathbf{y})\|_{H_0^1(D)}.$$

For later convenience, we introduce here the sequence $\mathbf{b} := (b_j)_{j \geq 1}$ defined by $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$. Recall that by the assumptions we placed on the uniform and affine model, there holds $\mathbf{b} \in \ell^p$ for some $p \in (0, 1)$.

Parametric regularity

Proposition

For all $\mathbf{y} \in [-1/2, 1/2]^{\mathbb{N}}$ and $\nu \in \mathcal{F}$, there holds

$$\|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} \leq \frac{C_P \|f\|_{L^2(D)}}{a_{\min}} \mathbf{b}^\nu |\nu|!,$$

where C_P is the Poincaré constant satisfying $\|v\|_{L^2(D)} \leq C_P \|v\|_{H_0^1(D)}$ for all $v \in H_0^1(D)$.

Proof. By induction w.r.t. the order of the multi-index $\nu \in \mathcal{F}$. If $\nu = \mathbf{0}$, then this is the ordinary Lax–Milgram *a priori* bound

$$\begin{aligned} a_{\min} \underbrace{\int_D |\nabla u(x, \mathbf{y})|^2 dx}_{=\|u(\cdot, \mathbf{y})\|_{H_0^1(D)}^2} &\leq \int_D a(x, \mathbf{y}) \nabla u(x, \mathbf{y}) \cdot \nabla u(x, \mathbf{y}) dx = \int_D f(x) u(x, \mathbf{y}) dx \\ &\leq \|f\|_{L^2(D)} \|u(\cdot, \mathbf{y})\|_{L^2(D)} \leq C_P \|f\|_{L^2(D)} \|u(\cdot, \mathbf{y})\|_{H_0^1(D)} \end{aligned}$$

whence

$$\|u(\cdot, \mathbf{y})\|_{H_0^1(D)} \leq \frac{C_P \|f\|_{L^2(D)}}{a_{\min}}.$$

Next, let $\nu \in \mathcal{F}$ and suppose that the claim has been proved for all multi-indices with order less than $|\nu|$. Then using the recursive relation we derived previously, we obtain

$$\begin{aligned} \|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} &\leq \sum_{j \in \text{supp}(\nu)} \nu_j b_j \|\partial^{\nu - e_j} u(\cdot, \mathbf{y})\|_{H_0^1(D)} \\ &\leq \frac{C_P \|f\|_{L^2(D)}}{a_{\min}} \sum_{j \in \text{supp}(\nu)} \nu_j b_j |\nu - e_j|! \mathbf{b}^{\nu - e_j} \\ &= \frac{C_P \|f\|_{L^2(D)}}{a_{\min}} \mathbf{b}^\nu (|\nu| - 1)! \sum_{j \geq 1} \nu_j \\ &= \frac{C_P \|f\|_{L^2(D)}}{a_{\min}} \mathbf{b}^\nu |\nu|!, \end{aligned}$$

as desired. □

Remark. Note that the same regularity bound holds for the dimensionally-truncated FE solution $u_{s,h}$ as long as a (conforming) Galerkin FE discretization has been used to construct the FE approximation. This is due to the fact that the weak formulation of the Galerkin discretization is exactly the same (only the function space differs).

Now that we know the regularity of the PDE problem, we can analyze the QMC cubature error! Let $G: H_0^1(D) \rightarrow \mathbb{R}$ be a linear and bounded functional, $u_{s,h}$ the dimensionally-truncated FE solution, and define $F(\mathbf{y}) := G(u_{s,h}(\cdot, \mathbf{y} - \frac{1}{2}))$ for $\mathbf{y} \in [0, 1]^s$. Let $\gamma = (\gamma_u)_{u \subseteq \{1:s\}}$ be a sequence of positive weights. Then we know that the generating vector obtained by the CBC algorithm satisfies the error bound

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|u|} \right)^{1/(2\lambda)} \|F\|_{s,\gamma}$$

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

$$\begin{aligned} \|F\|_{s,\gamma}^2 &= \sum_{u \subseteq \{1:s\}} \frac{1}{\gamma_u} \int_{[0,1]^{|u|}} \left(\int_{[0,1]^{s-|u|}} \frac{\partial^{|u|}}{\partial \mathbf{x}_u} F(\mathbf{y}) d\mathbf{y}_{-u} \right)^2 d\mathbf{y}_u \\ &\leq \left(\frac{C_P \|G\|_{H_0^1(D) \rightarrow \mathbb{R}} \|f\|_{L^2(D)}}{a_{\min}} \right)^2 \sum_{u \subseteq \{1:s\}} \frac{1}{\gamma_u} (|u|!)^2 \prod_{j \in u} b_j^2. \end{aligned}$$

Plugging this norm bound back into the QMC error bound yields...

$$\begin{aligned} \sqrt{\mathbb{E}_\Delta |I_s F - Q_{n,s}^\Delta F|^2} &\lesssim \left(\frac{1}{\varphi(n)} \right)^{1/(2\lambda)} \left(\sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|\mathfrak{u}|} \right)^{1/(2\lambda)} \\ &\quad \times \left(\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} (|\mathfrak{u}|!)^2 \prod_{j \in \mathfrak{u}} b_j^2 \right)^{1/2}. \end{aligned}$$

The upper bound can be *minimized* by choosing the *POD weights*

$$\gamma_{\mathfrak{u}} := \left(|\mathfrak{u}|! \prod_{j \in \mathfrak{u}} \frac{b_j}{\sqrt{\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}}} \right)^{2/(1+\lambda)},$$

as explained by the following lemma.

Lemma

Let (α_i) and (β_i) be sequences of positive real numbers. The expression

$$g(\gamma) := \left(\sum_i \alpha_i \gamma_i^\lambda \right)^{1/\lambda} \left(\sum_i \beta_i \gamma_i^{-1} \right)$$

is minimized by $\gamma_i = c \left(\frac{\beta_i}{\alpha_i} \right)^{1/(1+\lambda)}$ for arbitrary $c > 0$.

Proof. Let us find out when the gradient vanishes:

$$0 = \partial_j g(\boldsymbol{\gamma}) = \frac{1}{\lambda} \left(\sum_i \alpha_i \gamma_i^\lambda \right)^{1/\lambda - 1} \lambda \alpha_j \gamma_j^{\lambda-1} \left(\sum_i \beta_i \gamma_i^{-1} \right) \\ - \beta_j \gamma_j^{-2} \left(\sum_i \alpha_i \gamma_i^\lambda \right)^{1/\lambda}.$$

After some trivial simplifications, we can see that this is equivalent to

$$\gamma_j^{\lambda+1} = \frac{\beta_j}{\alpha_j} \frac{\sum_i \alpha_i \gamma_i^\lambda}{\sum_i \beta_i \gamma_i^{-1}}.$$

Furthermore, this condition is satisfied if

$$\gamma_j = c \left(\frac{\beta_j}{\alpha_j} \right)^{1/(1+\lambda)},$$

where $c > 0$ is arbitrary. □

Note that plugging $\gamma_i = c \left(\frac{\beta_i}{\alpha_i} \right)^{1/(1+\lambda)}$ into $(\sum_i \alpha_i \gamma_i^\lambda)^{1/(2\lambda)} (\sum_i \beta_i \gamma_i^{-1})^{1/2}$ yields the expression $(\sum_i \alpha_i^{1/(1+\lambda)} \beta_i^{\lambda/(1+\lambda)})^{(1+\lambda)/(2\lambda)}$. Thus, plugging the optimal POD weights into the QMC error bound results in

$$\sqrt{\mathbb{E}_\Delta |I_s F - Q_{n,s}^\Delta F|^2} \lesssim \left(\frac{1}{\varphi(n)} \right)^{1/(2\lambda)} C(s, \gamma, \lambda)^{(1+\lambda)/(2\lambda)},$$

where

$$C(s, \gamma, \lambda) := \sum_{\mathfrak{u} \subseteq \{1:s\}} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|\mathfrak{u}|/(1+\lambda)} (|\mathfrak{u}|!)^{2\lambda/(1+\lambda)} \prod_{j \in \mathfrak{u}} b_j^{2\lambda/(1+\lambda)}.$$

This is the punchline:

Lemma

By choosing

$$\lambda = \begin{cases} \frac{p}{2-p} & \text{when } p \in (2/3, 1) \\ \frac{1}{2-2\delta} \text{ for arbitrary } \delta \in (0, 1/2) & \text{when } p \in (0, 2/3], \end{cases}$$

there exists a constant $C(\gamma, \lambda) < \infty$ independently of s s.t. $C(s, \gamma, \lambda) \leq C(\gamma, \lambda) < \infty$.

Proof. First observe that

$$\begin{aligned}
 C(s, \gamma, \lambda) &= \sum_{\mathfrak{u} \subseteq \{1:s\}} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|\mathfrak{u}|/(1+\lambda)} (|\mathfrak{u}|!)^{2\lambda/(1+\lambda)} \prod_{j \in \mathfrak{u}} b_j^{2\lambda/(1+\lambda)} \\
 &= \sum_{\ell=0}^s \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{\ell/(1+\lambda)} (\ell!)^{2\lambda/(1+\lambda)} \sum_{\substack{|\mathfrak{u}|=\ell \\ \mathfrak{u} \subseteq \{1:s\}}} \prod_{j \in \mathfrak{u}} b_j^{2\lambda/(1+\lambda)} \\
 &\leq \sum_{\ell=0}^{\infty} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{\ell/(1+\lambda)} (\ell!)^{2\lambda/(1+\lambda)-1} \left(\sum_{j \geq 1} b_j^{2\lambda/(1+\lambda)} \right)^\ell
 \end{aligned}$$

where we used the inequality $\sum_{|\mathfrak{u}|=\ell, \mathfrak{u} \subseteq \mathbb{Z}_+} \prod_{j \in \mathfrak{u}} c_j \leq \frac{1}{\ell!} \left(\sum_{j \geq 1} c_j \right)^\ell$.

Case 1: $p \in (2/3, 1)$. We choose $p = \frac{2\lambda}{1+\lambda} \Leftrightarrow \lambda = \frac{p}{2-p} \in (1/2, 1)$, and

$$C(s, \gamma, \lambda) \leq \underbrace{\sum_{\ell=0}^{\infty} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{\ell/(1+\lambda)} (\ell!)^{p-1} \left(\sum_{j \geq 1} b_j^p \right)^\ell}_{=: a_\ell}$$

It is easy to see that $\frac{a_{\ell+1}}{a_\ell} \xrightarrow{\ell \rightarrow \infty} 0$. By the ratio test, this upper bound is finite independently of s .

Case 2: $p \in (0, 2/3]$. Let $\delta \in (0, 1/2)$ be arbitrary. We choose $\lambda = \frac{1}{2-2\delta} \in (1/2, 1)$. Now $\frac{2\lambda}{1+\lambda} = \frac{2}{3-2\delta} \in (2/3, 1)$. Especially, $\|\mathbf{b}\|_{\ell^{2\lambda/(1+\lambda)}} \leq \|\mathbf{b}\|_{\ell^p}$, and we obtain from the estimate on the previous slide that

$$\begin{aligned} C(s, \gamma, \lambda) &\leq \sum_{\ell=0}^{\infty} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{\ell/(1+\lambda)} (\ell!)^{2\lambda/(1+\lambda)-1} \left(\sum_{j \geq 1} b_j^{2\lambda/(1+\lambda)} \right)^\ell \\ &\leq \underbrace{\sum_{\ell=0}^{\infty} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{\ell/(1+\lambda)} (\ell!)^{2/(3-2\delta)-1} \left(\sum_{j \geq 1} b_j^p \right)^{2\ell/((3-2\delta)p)}}_{=: a_\ell} \end{aligned}$$

Again, $\frac{a_{\ell+1}}{a_\ell} \xrightarrow{\ell \rightarrow \infty} 0$, so by the ratio test this upper bound is finite independently of s . □

Theorem

Let $\delta \in (0, 1/2)$ be arbitrary. By choosing the POD weights

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

then the QMC approximation for the expected value of the PDE problem satisfies

$$\text{R.M.S. error} \lesssim \begin{cases} \left(\frac{1}{\varphi(n)}\right)^{1/p-1/2} & \text{if } p \in (2/3, 1), \\ \left(\frac{1}{\varphi(n)}\right)^{1-\delta} & \text{if } p \in (0, 2/3], \end{cases}$$

where the implied coefficient is independent of the dimension s .

Remark: We have the following dimension-independent convergence rates:

- n is prime $\Rightarrow \frac{1}{\varphi(n)} = \frac{1}{n-1} \Rightarrow$ QMC rate $\mathcal{O}(n^{\max\{-1/p+1/2, -1+\delta\}})$.
- $n = 2^k \Rightarrow \frac{1}{\varphi(n)} = \frac{2}{n} \Rightarrow$ QMC rate $\mathcal{O}(n^{\max\{-1/p+1/2, -1+\delta\}})$.
- For general composite n , the dimension-independent QMC rate is at best essentially linear up to a double logarithmic factor of n .

Remarks on implementation

Let $G: H_0^1(D) \rightarrow \mathbb{R}$ be a bounded linear functional. Consider the problem of approximating

$$\mathbb{E}[G(u_{s,h})] = \int_{[-1/2,1/2]^s} G(u_{s,h}(\cdot, \mathbf{y})) d\mathbf{y},$$

where $u_{s,h}$ is the dimensionally-truncated FE approximation to the elliptic PDE with a uniform and affine diffusion coefficient.

Our QMC approximation is guaranteed to satisfy the R.M.S. error bound from the previous slide if we plug the theoretically derived weights as input to the fast CBC algorithm. This produces a generating vector $\mathbf{z} \in \mathbb{N}^s$. The generating vector is designed to be used to compute the estimate

$$\overline{Q}_{n,s,R} G(u_{s,h}) := \frac{1}{R} \sum_{r=0}^{R-1} Q_{n,s}^{\Delta_r} G(u_{s,h}),$$

where $Q_{n,s}^{\Delta_r} F := \frac{1}{n} \sum_{i=0}^{n-1} f(\{\mathbf{t}_i + \Delta_r\} - \frac{1}{2})$, $\mathbf{t}_k := \{\frac{k\mathbf{z}}{n}\}$, and $\Delta_0, \dots, \Delta_{R-1}$ are independent random shifts drawn from $\mathcal{U}([0, 1]^s)$.

- Typically, the number of random shifts is taken to be rather small, e.g., $8 \leq R \leq 64$.
- A practical estimate for the R.M.S. error is given by the formula

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \approx \sqrt{\frac{1}{R(R-1)} \sum_{r=0}^{R-1} (Q_{n,s,R}^{\Delta_r} F - \bar{Q}_{n,s,R} F)^2}.$$

- For the computation of the variance, note that

$$\text{Var}[G(u_{s,h})] = \mathbb{E}[G(u_{s,h})^2] - \mathbb{E}[G(u_{s,h})]^2.$$

We already know how to approximate $\mathbb{E}[G(u_{s,h})]$ using QMC, but the weights need to be updated if we wish to construct a QMC rule with a dimension-independent convergence rate for $\mathbb{E}[G(u_{s,h})^2]$ (exercise).

- If a QMC rule converges independently of s for the approximation of $\mathbb{E}[G(u_{s,h})^2]$, then the same rule will have dimension-independent convergence for $\mathbb{E}[G(u_{s,h})]$ as well.
- If we instead wish to estimate $\mathbb{E}[u_{s,h}(\mathbf{x}, \cdot)]$ or $\text{Var}[u_{s,h}(\mathbf{x}, \cdot)]$ (i.e., leave out the *quantity of interest* $G : H_0^1(D) \rightarrow \mathbb{R}$), the same weights can be used as input to the CBC algorithm (but we still need to prove this).

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 $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient has the parameterization

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

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

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

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

and found that it is possible to construct a QMC point set $\mathbf{t}_i := \left\{ \frac{i\mathbf{z}}{n} \right\}$ 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 19th 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

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

∴ 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

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

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

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

and

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

Multinomial theorem

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

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

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

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

and we will later require the following special case:

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

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 $0 < p \leq q < \infty$, 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

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

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

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

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

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

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

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

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

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

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

$$\int_U f(\mathbf{y}) \, d\mathbf{y},$$

where we have the infinite tensor product measure

$$d\mathbf{y} := \bigotimes_{j=1}^{\infty} dy_j.$$

The σ -algebra \mathcal{F} for $d\mathbf{y}$ 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 $(U, \mathcal{F}, d\mathbf{y})$ 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:

$$\int_U f(\mathbf{y}) d\mathbf{y} = \lim_{s \rightarrow \infty} \int_{[-1/2, 1/2]^s} f(y_1, \dots, y_s, 0, 0, \dots) dy_1 \cdots dy_s. \quad (2)$$

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 $f: U \rightarrow \mathbb{R}$ be integrable w.r.t. the measure $d\mathbf{y} := \bigotimes_{j=1}^{\infty} dy_j$ which satisfies

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

$$|f(y_1, \dots, y_s, 0, 0, \dots)| \leq |g(\mathbf{y})| \quad \text{for a.e. } \mathbf{y} \in U$$

for some integrable function $g: U \rightarrow \mathbb{R}$ w.r.t. the measure $d\mathbf{y}$. 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 (A1)–(A3) hold and

$\|\psi_1\|_{L^\infty(D)} \geq \|\psi_2\|_{L^\infty(D)} \geq \|\psi_3\|_{L^\infty(D)} \geq \dots$. Then for every $f \in L^2(D)$ and every bounded linear functional $G: H_0^1(D) \rightarrow \mathbb{R}$, there holds

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

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 $u(\cdot, \mathbf{y})$ denotes the parametric PDE solution and f the source term, we require for the analysis the (linear) *forward operator*

$$A(\mathbf{y}): u(\cdot, \mathbf{y}) \mapsto f$$

and the *solution operator*

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

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

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

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

(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.,

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

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

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

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

Now

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

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

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

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

and

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

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

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

[†]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

$$A(\mathbf{y}), A^s(\mathbf{y}) : H_0^1(D) \rightarrow H^{-1}(D),$$

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

where $B_j : H_0^1(D) \rightarrow H^{-1}(D)$ 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

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

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

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

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

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

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

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

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 $\mathbf{y} \in U$ that

$$\begin{aligned}\|A(\mathbf{y})\|_{H_0^1(D) \rightarrow H^{-1}(D)} &\leq a_{\max}, \quad \|A^s(\mathbf{y})\|_{H_0^1(D) \rightarrow H^{-1}(D)} \leq a_{\max} \\ \|A(\mathbf{y})^{-1}\|_{H^{-1}(D) \rightarrow H_0^1(D)} &\leq \frac{1}{a_{\min}}, \quad \|A^s(\mathbf{y})^{-1}\|_{H^{-1}(D) \rightarrow H_0^1(D)} \leq \frac{1}{a_{\min}}.\end{aligned}$$

For brevity, let us denote

$$\begin{aligned}u(\mathbf{y}) &:= u(\cdot, \mathbf{y}), \quad \mathbf{y} \in U, \\ u_s(\mathbf{y}) &:= u_s(\cdot, \mathbf{y}), \quad \mathbf{y} \in U.\end{aligned}$$

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

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

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

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

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

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

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

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

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

is well-defined. Moreover, we have the representation

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

The integrand can be expanded as

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

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

$$\begin{aligned} & \int_U G \left(\left(\sum_{j=s+1}^{\infty} y_j A^s(\mathbf{y})^{-1} B_j \right)^k u_s(\mathbf{y}) \right) d\mathbf{y} \\ &= \sum_{\eta_1, \dots, \eta_k = s+1}^{\infty} \underbrace{\left(\int_U \prod_{i=1}^k y_{\eta_i} d\mathbf{y} \right)}_{=: I_1} \underbrace{\left(\int_{U_s} G \left(\left(\prod_{i=1}^k A^s(\mathbf{y})^{-1} B_{\eta_i} \right) u_s(\mathbf{y}) \right) d\mathbf{y}_{\{1:s\}} \right)}_{=: I_2}. \end{aligned}$$

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

Earlier we arrived at

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

We can estimate the summands as

$$\begin{aligned} & \left| (-1)^k \int_U G \left(\left(\sum_{j=s+1}^{\infty} y_j A^s(\mathbf{y})^{-1} B_j \right)^k u_s(\mathbf{y}) \right) d\mathbf{y} \right| \\ & \leq \frac{\|G\|_{H^{-1}(D)} \|F\|_{H^{-1}(D)}}{a_{\min}} \sum_{\eta_1, \dots, \eta_k=s+1}^{\infty} \left(\int_U \prod_{k=1}^k y_{\eta_i} d\mathbf{y} \right) \left(\prod_{i=1}^k b_{\eta_i} \right) \\ & = \frac{\|G\|_{H^{-1}(D)} \|F\|_{H^{-1}(D)}}{a_{\min}} \int_U \sum_{\eta_1, \dots, \eta_k=s+1}^{\infty} \left(\prod_{k=1}^k y_{\eta_i} \right) \left(\prod_{i=1}^k b_{\eta_i} \right) d\mathbf{y} \\ & = \frac{\|G\|_{H^{-1}(D)} \|F\|_{H^{-1}(D)}}{a_{\min}} \int_U \left(\sum_{j=s+1}^{\infty} y_j b_j \right)^k d\mathbf{y} \\ & \stackrel{(1)}{=} \frac{\|G\|_{H^{-1}(D)} \|F\|_{H^{-1}(D)}}{a_{\min}} \int_U \sum_{\substack{|\nu|=k \\ \nu_j=0 \ \forall j \leq s}} \frac{k!}{\nu!} \left(\prod_{j=s+1}^{\infty} y_j^{\nu_j} \right) \left(\prod_{j=s+1}^{\infty} b_j^{\nu_j} \right) d\mathbf{y}. \end{aligned}$$

The integrals vanish whenever ν contains an element equal to 1, hence

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

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

$$\begin{aligned} & \left| \int_U G(u(\mathbf{y}) - u_s(\mathbf{y})) d\mathbf{y} \right| \leq \frac{\|G\|_{H^{-1}(D)} \|F\|_{H^{-1}(D)}}{a_{\min}} \sum_{k=1}^{\infty} \sum_{\substack{|\nu|=k \\ \nu_j=0 \ \forall j \leq s \\ \nu_j \neq 1 \ \forall j > s}} \frac{k!}{\nu!} \mathbf{b}^\nu \\ & = \frac{\|G\|_{H^{-1}(D)} \|F\|_{H^{-1}(D)}}{a_{\min}} \left[\sum_{k=k'}^{\infty} \sum_{\substack{|\nu|=k \\ \nu_j=0 \ \forall j \leq s \\ \nu_j \neq 1 \ \forall j > s}} \frac{k!}{\nu!} \mathbf{b}^\nu + \sum_{k=2}^{k'-1} \sum_{\substack{|\nu|=k \\ \nu_j=0 \ \forall j \leq s \\ \nu_j \neq 1 \ \forall j > s}} \frac{k!}{\nu!} \mathbf{b}^\nu \right], \end{aligned}$$

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

$$\begin{aligned} \sum_{k=k'}^{\infty} \sum_{\substack{|\nu|=k \\ \nu_j=0 \ \forall j \leq s \\ \nu_j \neq 1 \ \forall j > s}} \frac{k!}{\nu!} \mathbf{b}^{\nu} &\leq \sum_{k=k'}^{\infty} \left(\sum_{j=s+1}^{\infty} b_j \right)^k \\ &\leq \left(\sum_{j=s+1}^{\infty} b_j \right)^{k'} \frac{1}{1 - \sum_{j=s+1}^{\infty} b_j} \leq C s^{k'(-1/p+1)}, \end{aligned}$$

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

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

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

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

$$\begin{aligned} \sum_{\substack{|\nu|=k \\ \nu_j=0 \ \forall j \leq s \\ \nu_j \neq 1 \ \forall j > s}} \mathbf{b}^\nu &\leq \sum_{\substack{0 \neq |\nu|_\infty \leq k \\ \nu_j=0 \ \forall j \leq s \\ \nu_j \neq 1 \ \forall j > s}} \mathbf{b}^\nu = \prod_{j=s+1}^{\infty} \left(1 + \sum_{\ell=2}^k b_j^\ell\right) - 1 \\ &= \prod_{j=s+1}^{\infty} \left(1 + b_j^2 \frac{1 - b_j^{j-1}}{1 - b_j}\right) - 1 \leq \prod_{j=s+1}^{\infty} (1 + 2b_j^2) - 1 \\ &\leq \exp \left(2 \sum_{j=s+1}^{\infty} b_j^2\right) - 1 \leq C_2 s^{-2/p+1}, \quad C_2 := 2(e-1) \left(\sum_{j=1}^{\infty} b_j^p\right)^{1/p}, \end{aligned}$$

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

Putting everything together, we conclude that

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

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

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

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

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

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 $h \rightarrow 0$. 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 \mathbf{y} and f such that the solution $u(\cdot, \mathbf{y}) \in H_0^1(D)$ of the parametric PDE problem satisfies

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

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

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

Dimensionally-truncated finite element solution

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

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

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:

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

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

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

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 \mathbf{y} such that

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

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

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

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

It follows from Galerkin orthogonality of the finite element solution that

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

In consequence,

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

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...

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

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

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

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

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

Overall error

Let $I(F) := \int_U F(\mathbf{y}) d\mathbf{y}$.

Theorem

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded polyhedron, assume (A1)–(A3), $\|\psi_1\|_{L^\infty(D)} \geq \|\psi_2\|_{L^\infty(D)} \geq \|\psi_3\|_{L^\infty(D)} \geq \dots$, 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 $G: L^2(D) \rightarrow \mathbb{R}$ 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

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

as inputs to construct a randomly shifted rank-1 lattice rule

$Q_{n,s}^{\Delta}(F) := \sum_{k=0}^{n-1} F(\{\frac{k\mathbf{z}}{n} + \Delta\} - \frac{1}{2})$, $\Delta \in [0, 1]^s$, we have the overall error

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

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

Proof. We have the total error decomposition[†]

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

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

$$|(I - I_s)(G(u))| = \mathcal{O}(s^{-2/p+1}),$$

$$|I_s(G(u_s - u_{s,h}))| = \mathcal{O}(h^2),$$

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

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

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

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

But what if we wanted to approximate

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

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.

By Fubini's theorem, we have that

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

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

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

The *shift-averaged worst-case error* $e_{n,s}^{\text{sh}}(z)$ is precisely the same object that we have considered in the past, i.e.,

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

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

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

The generating vector obtained using the CBC algorithm satisfies the estimate

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

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

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

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).

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Tenth lecture, June 30, 2025

Today's lecture follows the survey article

-  F. Y. Kuo and D. Nuyens. Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients - a survey of analysis and implementation. *Found. Comput. Math.* **16**:1631–1696, 2016. arXiv version: <https://arxiv.org/abs/1606.06613>

Introduction: transformation to the unit cube

Consider the (univariate) integral

$$\int_{-\infty}^{\infty} g(y)\phi(y) dy,$$

where $\phi: \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is a univariate probability density function, i.e., $\int_{-\infty}^{\infty} \phi(y) dy = 1$. How do we transform the integral into $[0, 1]$?

Let $\Phi: \mathbb{R} \rightarrow [0, 1]$ denote the cumulative distribution function of ϕ , defined by $\Phi(y) := \int_{-\infty}^y \phi(t) dt$ and let $\Phi^{-1}: [0, 1] \rightarrow \mathbb{R}$ denote its inverse. Then we use the change of variables

$$x = \Phi(y) \Leftrightarrow y = \Phi^{-1}(x)$$

to obtain

$$\int_{-\infty}^{\infty} g(y)\phi(y) dy = \int_0^1 g(\Phi^{-1}(x)) dx = \int_0^1 f(x) dx,$$

where $f := g \circ \Phi^{-1}$ is the transformed integrand.

Actually, we can multiply and divide by any other probability density function $\tilde{\phi}$ and then map to $[0, 1]$ using its inverse cumulative distribution function $\tilde{\Phi}^{-1}$:

$$\begin{aligned}\int_{-\infty}^{\infty} g(y)\phi(y) dy &= \int_{-\infty}^{\infty} \frac{g(y)\phi(y)}{\tilde{\phi}(y)} \tilde{\phi}(y) dy \\ &= \int_{-\infty}^{\infty} \tilde{g}(y)\tilde{\phi}(y) dy && (\tilde{g}(y) := \frac{g(y)\phi(y)}{\tilde{\phi}(y)}) \\ &= \int_0^1 \tilde{g}(\tilde{\Phi}^{-1}(x)) dx = \int_0^1 \tilde{f}(x) dx. && (\tilde{f} := \tilde{g} \circ \tilde{\Phi}^{-1})\end{aligned}$$

Ideally we would like to use a density function which leads to an easy integrand in the unit cube. (Compare this with *importance sampling* for the Monte Carlo method.)

This transformation can be generalized to s dimensions in the following way. If we have a product of univariate densities, then we can apply the mapping Φ^{-1} *componentwise*

$$\mathbf{y} = \Phi^{-1}(\mathbf{x}) = [\Phi^{-1}(x_1), \dots, \Phi^{-1}(x_s)]^T$$

to obtain

$$\int_{\mathbb{R}^s} g(\mathbf{y}) \prod_{j=1}^s \phi(y_j) d\mathbf{y} = \int_{(0,1)^s} g(\Phi^{-1}(\mathbf{x})) d\mathbf{x} = \int_{(0,1)^s} f(\mathbf{x}) d\mathbf{x}.$$

(Of course, dividing and multiplying by a product of arbitrary probability density functions would work here as well!)

Lognormal model

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded Lipschitz domain. In the “lognormal” case, we assume that the parameter \mathbf{y} is distributed in $\mathbb{R}^{\mathbb{N}}$ according to the product Gaussian measure $\mu_G = \bigotimes_{j=1}^{\infty} \mathcal{N}(0, 1)$. The parametric coefficient $a(\mathbf{x}, \mathbf{y})$ now takes the form

$$a(\mathbf{x}, \mathbf{y}) := a_0(\mathbf{x}) \exp \left(\sum_{j=1}^{\infty} y_j \psi_j(\mathbf{x}) \right), \quad \mathbf{x} \in D, \quad \mathbf{y} \in \mathbb{R}^{\mathbb{N}}, \quad (1)$$

where $a_0 \in L^\infty(D)$ with $a_0(\mathbf{x}) > 0$, $\mathbf{x} \in D$.

A coefficient of the form (1) can arise from the Karhunen–Loève (KL) expansion in the case where $\log(a)$ is a stationary Gaussian random field with a specified mean and a covariance function.

Example

Consider a Gaussian random field with an isotropic *Matérn covariance* $\text{Cov}(\mathbf{x}, \mathbf{x}') := \rho_\nu(|\mathbf{x} - \mathbf{x}'|)$, with

$$\rho_\nu(r) := \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(2\sqrt{\nu} \frac{r}{\lambda_C}\right)^\nu K_\nu\left(2\sqrt{\nu} \frac{r}{\lambda_C}\right),$$

where Γ is the gamma function and K_ν is the modified Bessel function of the second kind. The parameter $\nu > 1/2$ is a smoothness parameter, σ^2 is the variance, and λ_C is the correlation length scale.

If $\{(\lambda_j, \xi_j)\}_{j=1}^\infty$ is the sequence of eigenvalues and eigenfunctions of the covariance operator $(\mathcal{C}f)(\mathbf{x}) := \int_D \rho_\nu(|\mathbf{x} - \mathbf{x}'|) f(\mathbf{x}') d\mathbf{x}'$, i.e., $\mathcal{C}\xi_j = \lambda_j \xi_j$, where we assume that $\lambda_1 \geq \lambda_2 \geq \dots$ and the eigenfunctions are normalized s.t. $\|\xi_j\|_{L^2(D)} = 1$, then we can set $\psi_j(\mathbf{x}) := \sqrt{\lambda_j} \xi_j(\mathbf{x})$ in (1) to obtain the KL expansion for this Gaussian random field.

Lognormal model: let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded Lipschitz domain, and let $f \in H^{-1}(D)$. Let $\psi_j \in L^\infty(D)$ and $b_j := \|\psi_j\|_{L^\infty}$ for $j \in \mathbb{N}$ such that $\sum_{j=1}^{\infty} b_j < \infty$, and set

$$U_b := \left\{ \mathbf{y} \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} b_j |y_j| < \infty \right\}.$$

Consider the problem of finding, for all $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient is assumed to have the parameterization

$$a(\mathbf{x}, \mathbf{y}) := a_0(\mathbf{x}) \exp \left(\sum_{j=1}^{\infty} y_j \psi_j(\mathbf{x}) \right), \quad \mathbf{x} \in D, \quad \mathbf{y} \in U_b,$$

where $a_0 \in L^\infty(D)$ is such that $a_0(\mathbf{x}) > 0$, $\mathbf{x} \in D$.

Standing assumptions for the lognormal model

- (B1) We have $a_0 \in L^\infty(D)$ and $\sum_{j=1}^{\infty} b_j < \infty$.
- (B2) For every $\mathbf{y} \in U_b$, the expressions $a_{\max}(\mathbf{y}) := \max_{\mathbf{x} \in \bar{D}} a(\mathbf{x}, \mathbf{y})$ and $a_{\min}(\mathbf{y}) := \min_{\mathbf{x} \in \bar{D}} a(\mathbf{x}, \mathbf{y})$ are well-defined and satisfy $0 < a_{\min}(\mathbf{y}) \leq a(\mathbf{x}, \mathbf{y}) \leq a_{\max}(\mathbf{y}) < \infty$.
- (B3) $\sum_{j=1}^{\infty} b_j^p < \infty$ for some $p \in (0, 1)$.

Remark: Note that in the lognormal case, $a(\mathbf{x}, \mathbf{y})$ can take values which are arbitrarily close to 0 or arbitrarily large. Thus, the best we can do is to find \mathbf{y} -dependent lower and upper bounds $a_{\min}(\mathbf{y})$ and $a_{\max}(\mathbf{y})$. This will lead to a \mathbf{y} -dependent *a priori* bound and, consequently, \mathbf{y} -dependent parametric regularity bounds. This will make the QMC analysis more involved, leading one to consider “special” weighted, unanchored Sobolev spaces.

Clearly, the diffusion coefficient $a(\mathbf{x}, \mathbf{y})$ blows up for certain values of $\mathbf{y} \in \mathbb{R}^N$ (think of $y_j = b_j^{-1}$), but the PDE problem is well-defined in the parameter set U_b which turns out to be of full measure in $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), \mu_G)$.

Lemma

There holds $U_b \in \mathcal{B}(\mathbb{R}^N)$, where \mathcal{B} denotes the Borel σ -algebra and $\mu_G(U_b) = 1$.

Proof. See Lemma 2.28 in “Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs” by Ch. Schwab and C. J. Gittelson (2011). □

The previous lemma implies that

$$I(F) := \int_{\mathbb{R}^N} F(\mathbf{y}) \mu_G(d\mathbf{y}) = \int_{U_b} F(\mathbf{y}) \mu_G(d\mathbf{y}).$$

Thus, it is sufficient to restrict our parametric regularity analysis to $\mathbf{y} \in U_b$, for which $a(\mathbf{x}, \mathbf{y})$ (and hence $u(\mathbf{x}, \mathbf{y})$) are well-defined.

Let $G \in H^{-1}(D)$, our (dimensionally-truncated) integral quantity of interest can thus be written as

$$\begin{aligned} I_s(G(u_s)) &:= \int_{\mathbb{R}^s} G(u_s(\cdot, \mathbf{y})) \prod_{j=1}^s \phi(y_j) d\mathbf{y} = \int_{(0,1)^s} G(u(\Phi^{-1}(\mathbf{w}))) d\mathbf{w} \\ &\approx \frac{1}{n} \sum_{i=1}^n G(u(\Phi^{-1}(\mathbf{t}_i))) \\ &=: Q_{n,s}(G(u(\cdot, \Phi^{-1}(\cdot)))), \end{aligned}$$

where $Q_{n,s}$ represents a QMC rule over an s -dimensional point set $\{\mathbf{t}_i\}_{i=1}^n \subset (0,1)^s$.

Akin to the uniform case, we have a total error decomposition of the form

$$\begin{aligned}|I(G(u)) - Q_{n,s}(G(u_{s,h}))| &\leq |I(G(u - u_h))| \\&\quad + |I(G(u_h) - G(u_{s,h}))| \\&\quad + |I_s(G(u_{s,h})) - Q_{n,s}(G(u_{s,h}))|.\end{aligned}$$

We focus on the QMC error, but briefly mention the corresponding dimension truncation and finite element error results below. For further details, see Graham, Kuo, Nichols, Scheichl, Schwab, Sloan (2015).

- If $D \subset \mathbb{R}^2$ is a bounded convex polyhedron, $f \in L^2(D)$, $G \in L^2(D)'$, and $a(\cdot, \mathbf{y})$ is Lipschitz for all $\mathbf{y} \in U_b$, then the finite element error satisfies $\mathbb{E}[G(u - u_h)] = \mathcal{O}(h^2)$. (Similar result holds for $D \subset \mathbb{R}^3$.)
- For the Matérn covariance with $\nu > d/2$, there holds

$$|I(G(u_h)) - I(G(u_{s,h}))| = \mathcal{O}(s^{-\chi}), \quad 0 < \chi < \frac{\nu}{d} - \frac{1}{2}.$$

There has been some recent work on generalizing this result, cf., e.g., Guth and Kaarnioja (2024): <https://arxiv.org/abs/2209.06176>

Let us focus on the QMC error

$$\int_{\mathbb{R}^s} G(u_{s,h}(\cdot, \mathbf{y})) d\mathbf{y} - \frac{1}{n} \sum_{k=1}^n G(u_{s,h}(\cdot, \Phi^{-1}(\mathbf{t}_k))).$$

In this setting, we have

$$I_s(F) := \int_{\mathbb{R}^s} F(\mathbf{y}) \prod_{j=1}^s \phi(y_j) d\mathbf{y} = \int_{(0,1)^s} F(\Phi^{-1}(\mathbf{w})) d\mathbf{w}$$

and the randomly shifted QMC rules

$$Q_{n,s}^{(r)}(F) = \frac{1}{n} \sum_{k=1}^n F(\Phi^{-1}(\{\mathbf{t}_k + \boldsymbol{\Delta}_r\})),$$

$$\overline{Q}_{n,R}(F) := \frac{1}{R} \sum_{r=1}^R Q_{n,s}^{(r)}(F),$$

where we have R independent random shifts $\boldsymbol{\Delta}_1, \dots, \boldsymbol{\Delta}_R$ drawn from $\mathcal{U}([0, 1]^s)$, $\mathbf{t}_k := \{\frac{k\mathbf{z}}{n}\}$, with generating vector $\mathbf{z} \in \mathbb{N}^s$.

Function space setting

Kuo, Sloan, Wasilkowski, Waterhouse (2010): It turns out that the appropriate function space for unbounded integrands is a “special” weighted, unanchored Sobolev space equipped with the norm

$$\|F\|_{s,\gamma} = \left[\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \left(\int_{\mathbb{R}^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \mathbf{y}_{\mathfrak{u}}} F(\mathbf{y}) \left(\prod_{j \in \{1:s\} \setminus \mathfrak{u}} \phi(y_j) \right) d\mathbf{y}_{-\mathfrak{u}} \right)^2 \times \left(\prod_{j \in \mathfrak{u}} \varpi_j^2(y_j) \right) d\mathbf{y}_{\mathfrak{u}} \right]^{1/2}$$

where we have the weights

$$\varpi_j^2(y) := \exp(-2\alpha_j|y_j|), \quad \alpha_j > 0.$$

Brief idea: We’re interested in functions of the form $g(\mathbf{y}) = f(\Phi^{-1}(\mathbf{y}))$, where $f \in \mathcal{F}$. Now there exists an isometric space \mathcal{G} of functions s.t.

$$f \in \mathcal{F} \Leftrightarrow g = f(\Phi^{-1}(\cdot)) \in \mathcal{G} \text{ and } \|f\|_{\mathcal{F}} = \|g\|_{\mathcal{G}}.$$

If \mathcal{F} is a RKHS with kernel $K_{\mathcal{F}}$, then \mathcal{G} is a RKHS with kernel $K_{\mathcal{G}}(\mathbf{x}, \mathbf{y}) = K_{\mathcal{F}}(\Phi^{-1}(\mathbf{x}), \Phi^{-1}(\mathbf{y}))$. Thus the core idea is to investigate Sobolev spaces over unbounded domains which can be mapped isomorphically onto weighted Sobolev spaces over $(0, 1)^s$.

Theorem (Graham, Kuo, Nichols, Scheichl, Schwab, Sloan (2015))

Let F belong to the special weighted space over \mathbb{R}^s with weights γ , with ϕ being the standard normal density, and the weight functions ϖ_j defined as above. A randomly shifted lattice rule in s dimensions with n being a prime power can be constructed by a CBC algorithm such that

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \leq \left(\frac{2}{n} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^\lambda \prod_{j \in u} \varrho_j(\lambda) \right)^{1/(2\lambda)} \|F\|_{s,\gamma},$$

where $\lambda \in (1/2, 1]$ and

$$\varrho_j(\lambda) = 2 \left(\frac{\sqrt{2\pi} \exp(\alpha_j^2/\eta_*)}{\pi^{2-2\eta_*} (1-\eta_*) \eta_*} \right)^\lambda \zeta(\lambda + \tfrac{1}{2}) \quad \text{and} \quad \eta_* = \frac{2\lambda - 1}{4\lambda},$$

with $\zeta(x) := \sum_{k=1}^{\infty} k^{-x}$ denoting the Riemann zeta function for $x > 1$.

The steps for QMC analysis are the same as in the uniform case: (1) estimate $\|\cdot\|_{s,\gamma}$ for a given integrand (2) find weights γ which minimize the upper bound (3) plug the weights into the new error bound and estimate the constant (which ideally can be bounded independently of s). 304

Applying the theory in practice

Let us consider the parametric regularity of

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

where $a(\mathbf{x}, \mathbf{y}) := a_0(\mathbf{x}) \exp\left(\sum_{j=1}^{\infty} y_j \psi_j(\mathbf{x})\right)$ and $f \in H^{-1}(D)$.

Our strategy will be to obtain a parametric regularity bound for

$$\|\sqrt{a(\cdot, \mathbf{y})} \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)},$$

that is, we find a *sharp* estimate $\partial^\nu u(\cdot, \mathbf{y})$ in the *energy norm*, and then use the coercivity of the problem to bound this from below by

$$\begin{aligned} \|\sqrt{a(\cdot, \mathbf{y})} \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} &\geq \sqrt{a_{\min}(\mathbf{y})} \|\nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} \\ &= \sqrt{a_{\min}(\mathbf{y})} \|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)}. \end{aligned}$$

(Compare with task 1 of Exercise 2, where we used a similar technique to obtain a better constant for Céa's lemma!)

Lemma

$$\|\sqrt{a(\cdot, \mathbf{y})} \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} \leq \Lambda_{|\nu|} \mathbf{b}^\nu \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}},$$

where $(\Lambda_k)_{k=0}^\infty$ are the ordered Bell numbers defined by the recursion

$$\Lambda_0 := 1 \quad \text{and} \quad \Lambda_k := \sum_{\ell=1}^k \binom{k}{\ell} \Lambda_{k-\ell}, \quad k \geq 1.$$

Proof. By induction with respect to the order of the multi-indices. The case $|\nu| = 0$ is resolved by observing that

$$\begin{aligned} \|a(\cdot, \mathbf{y})^{1/2} \nabla u(\cdot, \mathbf{y})\|_{L^2(D)}^2 &= \int_D a(\mathbf{x}, \mathbf{y}) |\nabla u(\mathbf{x}, \mathbf{y})|^2 d\mathbf{x} = \int_D f(\mathbf{x}) u(\mathbf{x}, \mathbf{y}) d\mathbf{x} \\ &\leq \|f\|_{H^{-1}(D)} \|u(\cdot, \mathbf{y})\|_{H_0^1(D)} \\ &\leq \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \|a(\cdot, \mathbf{y})^{1/2} \nabla u(\cdot, \mathbf{y})\|_{L^2(D)} \end{aligned}$$

Next, let $\nu \in \mathcal{F} \setminus \{\mathbf{0}\}$ be such that the claim has been proved for all multi-indices with order $< |\nu|$. By exploiting the fact that

$$\left\| \frac{\partial^{\mathbf{m}} a(\cdot, \mathbf{y})}{a(\cdot, \mathbf{y})} \right\|_{L^\infty(D)} = \left\| \prod_{j \geq 1} \psi_j(\cdot)^{\nu_j} \right\|_{L^\infty(D)} \leq \mathbf{b}^\nu,$$

we obtain (using the Leibniz product rule)

$$\begin{aligned} & \sum_{\mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \int_D \partial^{\mathbf{m}} a(\mathbf{x}, \mathbf{y}) \nabla \partial^{\nu - \mathbf{m}} u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = 0 \\ \Leftrightarrow & \int_D a(\mathbf{x}, \mathbf{y}) \nabla \partial^\nu u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} \\ &= - \sum_{\mathbf{0} \neq \mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \int_D \underbrace{\frac{\partial^{\mathbf{m}} a(\mathbf{x}, \mathbf{y})}{a(\mathbf{x}, \mathbf{y})} a(\mathbf{x}, \mathbf{y})}_{= \frac{\partial^{\mathbf{m}} a(\mathbf{x}, \mathbf{y})}{a(\mathbf{x}, \mathbf{y})} a(\mathbf{x}, \mathbf{y})} \nabla \partial^{\nu - \mathbf{m}} u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Testing against $v = \partial^\nu u$ yields...

$$\begin{aligned}
& \|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)}^2 = \int_D a(\mathbf{x}, \mathbf{y}) |\nabla u(\mathbf{x}, \mathbf{y})|^2 d\mathbf{x} \\
& \leq \sum_{\mathbf{0} \neq \mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \int_D \left| \frac{\partial^{\mathbf{m}} a(\mathbf{x}, \mathbf{y})}{a(\mathbf{x}, \mathbf{y})} \right| a(\mathbf{x}, \mathbf{y}) \nabla \partial^{\nu-\mathbf{m}} u(\mathbf{x}, \mathbf{y}) \cdot \nabla \partial^\nu u(\mathbf{x}, \mathbf{y}) d\mathbf{x} \\
& \leq \sum_{\mathbf{0} \neq \mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \mathbf{b}^{\mathbf{m}} \|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^{\nu-\mathbf{m}} u(\cdot, \mathbf{y})\|_{L^2(D)} \|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)}
\end{aligned}$$

leading to the recurrence relation

$$\|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} \leq \sum_{\mathbf{0} \neq \mathbf{m} \leq \nu} \binom{\nu}{\mathbf{m}} \mathbf{b}^{\mathbf{m}} \|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^{\nu-\mathbf{m}} u(\cdot, \mathbf{y})\|_{L^2(D)}$$

By our induction hypothesis,

$$\|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^{\nu-\mathbf{m}} u(\cdot, \mathbf{y})\|_{L^2(D)} \leq \Lambda_{|\nu|-|\mathbf{m}|} \mathbf{b}^{\nu-\mathbf{m}} \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}}. \text{ This results in...}$$

$$\begin{aligned}
& \|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} \leq \sum_{\mathbf{0} \neq \mathbf{m} \leq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\mathbf{m}} \mathbf{b}^{\mathbf{m}} \|a^{1/2}(\cdot, \mathbf{y}) \nabla \partial^{\boldsymbol{\nu}-\mathbf{m}} u(\cdot, \mathbf{y})\|_{L^2(D)} \\
& \leq \mathbf{b}^\boldsymbol{\nu} \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \sum_{\mathbf{0} \neq \mathbf{m} \leq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\mathbf{m}} \Lambda_{|\boldsymbol{\nu}| - |\mathbf{m}|} \\
& = \mathbf{b}^\boldsymbol{\nu} \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \sum_{\ell=1}^{|\boldsymbol{\nu}|} \Lambda_{|\boldsymbol{\nu}| - \ell} \sum_{\substack{|\mathbf{m}| = \ell \\ \mathbf{m} \leq \boldsymbol{\nu}}} \binom{\boldsymbol{\nu}}{\mathbf{m}} \\
& = \mathbf{b}^\boldsymbol{\nu} \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \sum_{\ell=1}^{|\boldsymbol{\nu}|} \Lambda_{|\boldsymbol{\nu}| - \ell} \binom{|\boldsymbol{\nu}|}{\ell} \\
& = \mathbf{b}^\boldsymbol{\nu} \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \Lambda_{|\boldsymbol{\nu}|}. \quad \square
\end{aligned}$$

A bound for Λ_k

The ordered Bell numbers have the following simple upper bound.

Lemma (Beck, Tempone, Nobile, Tamellini (2012))

$$\Lambda_k \leq \frac{k!}{(\log 2)^k}$$

Proof. By definition $\Lambda_k = \sum_{\ell=1}^k \binom{k}{\ell} \Lambda_{k-\ell} = \sum_{\ell=1}^k \frac{k!}{\ell!} \frac{\Lambda_{k-\ell}}{(k-\ell)!}$, $\Lambda_0 = 1$. Define $f_k := \frac{\Lambda_k}{k!}$; then clearly

$$f_k = \sum_{\ell=1}^k \frac{f_{k-\ell}}{\ell!}, \quad f_0 = f_1 = 1.$$

We prove by induction that $f_k \leq \alpha^k$ for some $\alpha \geq 1$. The base steps $k = 0, 1$ hold for all $\alpha \geq 1$ due to $f_0 = f_1 = 1$. Thus we assume that the claim holds for f_1, \dots, f_{k-1} .

$$f_k = \sum_{\ell=1}^k \frac{f_{k-\ell}}{\ell!} \leq \sum_{\ell=1}^k \frac{\alpha^{k-\ell}}{\ell!} = \alpha^k \sum_{\ell=1}^k \frac{\alpha^{-\ell}}{\ell!} \leq \alpha^k (\mathrm{e}^{\frac{1}{\alpha}} - 1) \leq \alpha^k,$$

where the last step holds provided that

$$\begin{aligned} \mathrm{e}^{\frac{1}{\alpha}} - 1 \leq 1 &\Leftrightarrow \mathrm{e}^{\frac{1}{\alpha}} \leq 2 \\ &\Leftrightarrow \frac{1}{\alpha} \leq \log 2 \\ &\Leftrightarrow \alpha \geq \frac{1}{\log 2}. \end{aligned}$$

Thus $f_k \leq \alpha^k$ for all $\alpha \geq \frac{1}{\log 2} (> 1)$. We get the sharpest bound by taking $\alpha = \frac{1}{\log 2}$, which yields

$$\Lambda_k = k! f_k \leq \frac{k!}{(\log 2)^k}$$

as desired. □



Proposition

$$\|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} \leq \frac{\|f\|_{H^{-1}(D)}}{\min_{\mathbf{x} \in \bar{D}} a_0(\mathbf{x})} \frac{|\nu|!}{(\log 2)^{|\nu|}} \mathbf{b}^\nu \prod_{j \geq 1} \exp(b_j |y_j|)$$

Proof. From the previous discussion, we have that

$$\begin{aligned} \sqrt{a_{\min}(\mathbf{y})} \|\nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} &\leq \|\sqrt{a(\cdot, \mathbf{y})} \nabla \partial^\nu u(\cdot, \mathbf{y})\|_{L^2(D)} \\ &\leq \Lambda_{|\nu|} \mathbf{b}^\nu \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \\ &\leq \frac{|\nu|!}{(\log 2)^{|\nu|}} \mathbf{b}^\nu \frac{\|f\|_{H^{-1}(D)}}{\sqrt{a_{\min}(\mathbf{y})}} \\ \Rightarrow \quad \|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} &\leq \frac{\|f\|_{H^{-1}(D)}}{a_{\min}(\mathbf{y})} \frac{|\nu|!}{(\log 2)^{|\nu|}} \mathbf{b}^\nu. \end{aligned}$$

The claim follows by observing that

$$\frac{1}{a_{\min}(\mathbf{y})} = \frac{1}{\min_{\mathbf{x} \in \bar{D}} (a_0(\mathbf{x}) \exp(\sum_{j \geq 1} y_j \psi_j(\mathbf{x})))} \leq \frac{\exp(\sum_{j \geq 1} |y_j| \|\psi_j\|_{L^\infty(D)})}{\min_{\mathbf{x} \in \bar{D}} a_0(\mathbf{x})}.$$



Estimating the special weighted Sobolev norm

Let $G \in H^{-1}(D)$. Then

$$\begin{aligned} & \|G(u)\|_{s,\gamma}^2 \\ &= \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \left(\int_{\mathbb{R}^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \mathbf{y}_{\mathfrak{u}}} G(u(\cdot, \mathbf{y})) \prod_{j \notin \mathfrak{u}} \phi(y_j) d\mathbf{y}_{-\mathfrak{u}} \right)^2 \prod_{j \in \mathfrak{u}} \varpi_j^2(y_j) d\mathbf{y}_{\mathfrak{u}} \\ &\lesssim \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \left(\prod_{j \in \mathfrak{u}} \frac{b_j}{\log 2} \right)^2 \int_{\mathbb{R}^s} \prod_{j=1}^s \exp(2b_j|y_j|) \prod_{j \notin \mathfrak{u}} \phi(y_j) \prod_{j \in \mathfrak{u}} \varpi_j^2(y_j) d\mathbf{y} \\ &= \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \left(\prod_{j \in \mathfrak{u}} \frac{b_j}{\log 2} \right)^2 \left(\prod_{j \notin \mathfrak{u}} \underbrace{\int_{\mathbb{R}} \exp(2b_j|y_j|) \phi(y_j) dy_j}_{=2 \exp(2b_j^2) \Phi(2b_j)} \right) \\ &\quad \times \left(\prod_{j \in \mathfrak{u}} \int_{\mathbb{R}} \exp(2b_j|y_j|) \varpi_j^2(y_j) dy_j \right) \end{aligned}$$

Multiplying and dividing the summand by $\prod_{j \in \mathfrak{u}} 2 \exp(2b_j^2) \Phi(2b_j)$ yields...

$$\begin{aligned}
& \|G(u)\|_{s,\gamma}^2 \\
& \leq \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \left(\prod_{j=1}^s 2 \exp(2b_j^2) \Phi(2b_j) \right) \\
& \quad \times \left(\prod_{j \in \mathfrak{u}} \frac{b_j^2}{2(\log 2)^2 \exp(2b_j^2) \Phi(2b_j)} \int_{\mathbb{R}} \exp(2b_j |y_j|) \varpi_j^2(y_j) dy_j \right).
\end{aligned}$$

Recall that $\varpi_j^2(y_j) = \exp(-2\alpha_j |y_j|)$. If $\alpha_j > b_j$, then

$$\int_{\mathbb{R}} \exp(2b_j |y_j|) \varpi_j^2(y_j) dy_j = \frac{1}{\alpha_j - b_j}$$

and we obtain

$$\begin{aligned}
& \|G(u)\|_{s,\gamma}^2 \\
& \leq \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \left(\prod_{j=1}^s 2 \exp(2b_j^2) \Phi(2b_j) \right) \\
& \quad \times \left(\prod_{j \in \mathfrak{u}} \frac{b_j^2}{2(\log 2)^2 \exp(2b_j^2) \Phi(2b_j) (\alpha_j - b_j)} \right).
\end{aligned}$$

The remainder of the argument follows by similar reasoning as the uniform setting: the error criterion is minimized by choosing the weights

$$\gamma_{\mathfrak{u}} = \left(|\mathfrak{u}|! \prod_{j \in \mathfrak{u}} \frac{b_j}{\sqrt{2}(\log 2) \exp(b_j^2) \sqrt{\Phi(2b_j)(\alpha_j - b_j)\varrho_j(\lambda)}} \right)^{2/(1+\lambda)} \quad (2)$$

for $\mathfrak{u} \subseteq \{1 : s\}$, with

$$\lambda = \begin{cases} \frac{1}{2-2\delta} & \text{for arbitrary } \delta \in (0, 1/2) \\ \frac{p}{2-p} & \text{if } p \in (2/3, 1). \end{cases}$$

The resulting bound can be minimized with respect to the parameters α_j . This corresponds to minimizing $\varrho_j(\lambda)^{1/\lambda}/(\alpha_j - b_j)$ with respect to α_j , which yields

$$\alpha_j = \frac{1}{2} \left(b_j + \sqrt{b_j^2 + 1 - \frac{1}{2\lambda}} \right).$$

We obtain the overall cubature error rate $\mathcal{O}(n^{\max\{-1/p+1/2, -1+\delta\}})$ independently of the dimension s . Thus using the weights (2) as inputs to a (fast) CBC algorithm produces a QMC rule with a dimension independent convergence rate in the lognormal setting!

Uncertainty Quantification and Quasi-Monte Carlo

Sommersemester 2025

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eleventh lecture, July 7, 2025
[Summary](#)

Elliptic PDE

Many physical phenomena can be modeled using elliptic partial differential equations of the form

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) = f(\mathbf{x}), & \mathbf{x} \in D, \\ +\text{boundary conditions} \end{cases}$$

Uncertainties can appear in the material parameter a , source term f , boundary conditions, or the domain D .

- For the purposes of analysis, we consider the weak formulation of the PDE. Under certain conditions, the solution to the weak formulation can be shown to exist and be uniquely defined.
- When we solve the PDE numerically using the finite element method, we are actually approximating the solution to the *the weak formulation* of the PDE problem.
- Under suitably strong regularity assumptions (D convex Lipschitz domain, $f \in L^2(D)$, and a Lipschitz), the weak solution satisfies $-\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) = f(\mathbf{x})$ for a.e. $\mathbf{x} \in D$ with $u|_{\partial D} = 0$.

Let $D \subset \mathbb{R}^d$ be a nonempty open set.

$$L^2(D) := \{v: D \rightarrow \mathbb{R} \mid v \text{ is measurable}, \|v\|_{L^2(D)} := \left(\int_D |v(x)|^2 dx \right)^{1/2} < \infty\},$$

$$H^1(D) := \{v \in L^2(D) \mid \partial_j v \in L^2(D) \text{ for all } j \in \{1, \dots, d\}\},$$

$$\text{with } \|v\|_{H^1(D)} := (\|v\|_{L^2(D)}^2 + \|\nabla v\|_{L^2(D)}^2)^{1/2},$$

$$C_0^\infty(D) := \{v \in C^\infty(D) \mid \text{supp}(v) \subset D \text{ is a compact set}\},$$

$$\text{where } \text{supp}(v) := \overline{\{x \in D \mid v(x) \neq 0\}},$$

$$H_0^1(D) := \text{cl}_{H^1(D)}(C_0^\infty(D)).$$

The spaces $L^2(D)$, $H^1(D)$, and $H_0^1(D)$ are Hilbert spaces.

Poincaré's inequality: if $D \subset \mathbb{R}^d$ is a bounded domain, then there exists a constant $C_P > 0$ (depending on the domain D) such that

$$\|v\|_{L^2(D)} \leq C_P \|\nabla v\|_{L^2(D)} \quad \text{for all } v \in H_0^1(D).$$

Therefore, we can define an equivalent norm in $H_0^1(D)$ by setting

$$\|v\|_{H_0^1(D)} := \|\nabla v\|_{L^2(D)}.$$

This induces exactly the same topology in $H_0^1(D)$ as the usual Sobolev norm $\|\cdot\|_{H^1(D)}$.

Trace theorem and boundary values

Trace theorem: Let D be a bounded Lipschitz domain. Then the trace operator

$$\gamma: C^\infty(\overline{D}) \rightarrow C^\infty(\partial D), \quad \gamma u = u|_{\partial D},$$

has a unique extension to a bounded linear operator $\gamma: H^1(D) \rightarrow L^2(\partial D)$.

This means that even though $u \in H^1(D)$ is not well-defined over a set of measure zero, we can interpret its restriction to the boundary of the domain D as the trace $\gamma u \in L^2(\partial D)$.

Especially, Sobolev functions $u \in H^1(D)$ with zero trace are precisely the elements of $H_0^1(D)$:

$$u \in H_0^1(D) \quad \Leftrightarrow \quad \gamma u = 0: \partial D \rightarrow \mathbb{R}.$$

Q: How to solve such PDE problems in practice?

A: We consider the weak formulation of the PDE problem: given $f \in L^2(D)$, find $u \in H_0^1(D)$ such that

$$\underbrace{\int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x}}_{=:B(u,v)} = \underbrace{\int_D f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x}}_{=:F(v)} \quad \text{for all } v \in H_0^1(D), \quad (1)$$

where $F: H_0^1(D) \rightarrow \mathbb{R}$ is a bounded linear functional. If there exist $a_{\min}, a_{\max} > 0$ s.t. $0 < a_{\min} \leq a(\mathbf{x}) \leq a_{\max} < \infty$ for all $\mathbf{x} \in D$, then the bilinear form $B: H_0^1(D) \times H_0^1(D) \rightarrow \mathbb{R}$ is bounded, i.e.,

$$|B(u, v)| = \left| \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} \right| \leq a_{\max} \|u\|_{H_0^1(D)} \|v\|_{H_0^1(D)}$$

for all $u, v \in H_0^1(D)$, and coercive, i.e.,

$$B(u, u) = \left| \int_D a(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla u(\mathbf{x}) \, d\mathbf{x} \right| \geq a_{\min} \|u\|_{H_0^1(D)}^2 \quad \text{for all } u \in H_0^1(D),$$

the *Lax–Milgram lemma* ensures that there exists a unique solution $u \in H_0^1(D)$ to (1).

Galerkin method

To solve the system approximately, let $V_m \subset H_0^1(D)$ be a finite-dimensional subspace of the solution space $H_0^1(D)$.

The *Galerkin solution* $u_m \in V_m$ of the system (1) is the unique solution such that

$$\int_D a(\mathbf{x}) \nabla u_m(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \, d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } v \in V_m.$$

Let V_m be spanned by ψ_1, \dots, ψ_m . We can write the solution as $u_m = \sum_{i=1}^m c_i \psi_i$. The above system reduces to the linear system of equations

$$\begin{bmatrix} \int_D \nabla \psi_1(\mathbf{x}) \cdot \nabla \psi_1(\mathbf{x}) \, d\mathbf{x} & \cdots & \int_D \nabla \psi_1(\mathbf{x}) \cdot \nabla \psi_m(\mathbf{x}) \, d\mathbf{x} \\ \vdots & \ddots & \vdots \\ \int_D \nabla \psi_m(\mathbf{x}) \cdot \nabla \psi_1(\mathbf{x}) \, d\mathbf{x} & \cdots & \int_D \nabla \psi_m(\mathbf{x}) \cdot \nabla \psi_m(\mathbf{x}) \, d\mathbf{x} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \int_D f(\mathbf{x}) \psi_1(\mathbf{x}) \, d\mathbf{x} \\ \vdots \\ \int_D f(\mathbf{x}) \psi_m(\mathbf{x}) \, d\mathbf{x} \end{bmatrix}.$$

Solving this system and plugging the expansion coefficients back into the expression for u_m yields the Galerkin solution.

Céa's lemma

The solution to the Galerkin system is quasi-optimal in the following sense:

$$\|u - u_m\|_{H_0^1(D)} \leq \frac{a_{\max}}{a_{\min}} \inf_{v_m \in V_m} \|u - v_m\|_{H_0^1(D)}.$$

That is, the $H_0^1(D)$ error between the true PDE solution u and the Galerkin approximation u_m differs from the *optimal approximation* in V_m up to a constant factor.

Finite element method

The finite element method is a particular method of constructing the finite-dimensional subspaces V_m of the solution space $H_0^1(D)$.

- Construct a triangulation for the computational domain D .
- The space V_m is spanned by piecewise linear functions ψ_1, \dots, ψ_m which are constructed to satisfy

$$\psi_i(\mathbf{n}_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{n}_1, \dots, \mathbf{n}_m$ are the *interior* nodes of the triangulation.

- The finite element solution can be written as $u_h(\mathbf{x}) = \sum_{i=1}^m c_i \psi_i(\mathbf{x}) \in V_h$, where the expansion coefficients are solved from the Galerkin system. Note that $u_h(\mathbf{n}_j) = c_j$.
- If $v_h(\mathbf{x}) = \sum_{i=1}^m c_i \psi_i(\mathbf{x}) \in V_h$, then, e.g., $\|v_h\|_{L^2(D)} = \sqrt{\mathbf{c}^T M \mathbf{c}}$, where $\mathbf{c} := [c_1, \dots, c_m]^T$ and $M = [M_{i,j}]_{i,j=1}^m$ is the mass matrix defined elementwise by $M_{i,j} := \int_D \psi_i(\mathbf{x}) \psi_j(\mathbf{x}) d\mathbf{x}$, $i, j \in \{1, \dots, m\}$.

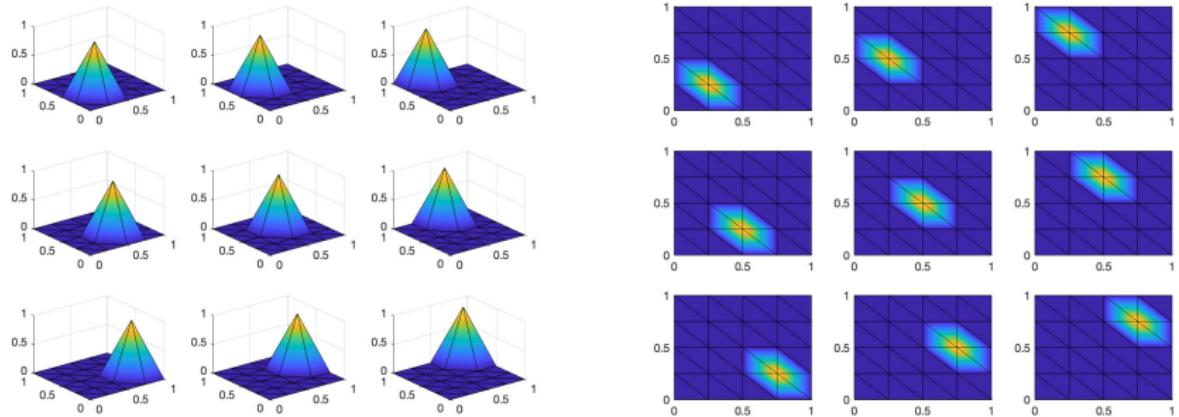


Figure: Left: An illustration of global, piecewise linear FE basis functions spanning V_h over a regular, uniform triangulation of $(0, 1)^2$. Right: Bird's-eye view of the same global FE basis functions.

Random field

Definition

Let $D \subset \mathbb{R}^d$ and let $(\Omega, \mathcal{F}, \mu)$ be a probability space. A function $A: D \times \Omega \rightarrow X$ is called a *random field* if $A(x, \cdot)$ is an X -valued random variable for all $x \in D$.

Definition

We call a random field $A: D \times \Omega \rightarrow X$ square-integrable if

$$\int_{\Omega} |A(x, \omega)|^2 \mu(d\omega) < \infty \quad \text{for all } x \in D.$$

Our goal will be to model (infinite-dimensional) input random fields using finite-dimensional expansions with s variables.

Comment on notation: In what follows, s will always refer to the “stochastic dimension” (dimension of the stochastic/parametric space) while d will refer to the “spatial dimension” (dimension of the spatial Lipschitz domain $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$).

Mercer's theorem

Let $a(x, \omega)$ be a square-integrable random field with mean

$$\bar{a}(x) = \int_{\Omega} a(x, \omega) \mu(d\omega), \quad x \in D,$$

and a continuous, symmetric, positive definite[†] covariance

$$K(x, x') = \int_{\Omega} (a(x, \omega) - \bar{a}(x))(a(x', \omega) - \bar{a}(x')) \mu(d\omega).$$

Mercer's theorem: the covariance operator $\mathcal{C}: L^2(D) \rightarrow L^2(D)$,

$$(\mathcal{C}u)(x) = \int_D K(x, x') u(x') dx', \quad x \in D,$$

has a countable sequence of eigenvalues $\{\lambda_k\}_{k \geq 1}$ and corresponding eigenfunctions $\{\psi_k\}_{k \geq 1}$ satisfying $\mathcal{C}\psi_k = \lambda_k \psi_k$ such that $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ and $\lambda_k \rightarrow 0$ and the eigenfunctions form an orthonormal basis for $L^2(D)$.

Note that this means:

$$\int_D K(x, x') \psi_k(x') dx' = \lambda_k \psi_k(x), \quad \int_D \psi_k(x) \psi_\ell(x) dx = \delta_{k,\ell}.$$

[†]In this context, positive definite means: for all choices of finitely many points $x_1, \dots, x_k \in D$, $k \in \mathbb{N}$, the Gram matrix $G := [K(x_i, x_j)]_{i,j=1}^k$ is positive semidefinite.

The Karhunen–Loève (KL) expansion of a random field

Theorem

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, let $D \subset \mathbb{R}^d$ be closed and bounded, and let $a: D \times \Omega \rightarrow \mathbb{R}$ be a square-integrable random field with continuous, symmetric, positive definite covariance

$K(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(a(\mathbf{x}, \cdot) - \bar{a}(\mathbf{x}))(a(\mathbf{x}', \cdot) - \bar{a}(\mathbf{x}'))]$. Then the eigensystem $(\lambda_k, \psi_k) \in \mathbb{R}_+ \times L^2(D)$ of the covariance operator $\mathcal{C}: L^2(D) \rightarrow L^2(D)$, as described on the previous slide, can be used to write

$$a(\mathbf{x}, \omega) = \bar{a}(\mathbf{x}) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \psi_k(\mathbf{x}),$$

$$\text{where } \xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_D (a(\mathbf{x}, \omega) - \bar{a}(\mathbf{x})) \psi_k(\mathbf{x}) d\mathbf{x},$$

where the convergence is in L^2 w.r.t. the stochastic parameter and uniform in \mathbf{x} . Furthermore, the random variables ξ_k are zero-mean uncorrelated random variables with unit variance, i.e.,

$$\mathbb{E}[\xi_k] = 0 \quad \text{and} \quad \mathbb{E}[\xi_k \xi_\ell] = \delta_{k,\ell}.$$

The Karhunen–Loève (KL) expansion of random field $a(\mathbf{x}, \omega)$ can be written as

$$a(\mathbf{x}, \omega) = \bar{a}(\mathbf{x}) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \psi_k(\mathbf{x}).$$

- The KL expansion minimizes the mean-square truncation error:
$$\|a(\mathbf{x}, \omega) - \bar{a}(\mathbf{x}) - \sum_{k=1}^s \sqrt{\lambda_k} \xi_k(\omega) \psi_k(\mathbf{x})\|_{L^2(\Omega, \mu; L^2(D))} = \left(\sum_{k=s+1}^{\infty} \lambda_k \right)^{1/2}.$$
- The random variables ξ_k are centered and uncorrelated, but not necessarily independent.
- If the random field $a(\mathbf{x}, \omega)$ is Gaussian – by definition, this means that $(a(\mathbf{x}_1, \omega), \dots, a(\mathbf{x}_k, \omega))$ is a multivariate Gaussian random variable for all $\mathbf{x}_1, \dots, \mathbf{x}_k \in D$, $k \in \mathbb{N}$ – then the random variables ξ_k are independent.
- The KL expansion is an effective method of representing *input* random fields when their covariance structure is known. If the (infinite-dimensional) input random field has a known covariance (which satisfies the conditions of Mercer's theorem), then we can use the KL expansion to find a finite-dimensional approximation, optimal in the mean-square error sense.

Modeling assumptions

In engineering and practical applications, the idea is that we have some *a priori* knowledge/belief that the unknown input random field is distributed according to some known probability distribution with a known covariance.

- If the input random field is Gaussian with a known, nice covariance function[†], then we use the KL expansion to find a reasonable finite-dimensional approximation of true input. Since the KL expansion decorrelates the stochastic variables, and uncorrelated jointly Gaussian random variables are independent, we can exploit the independence of the stochastic variables to parameterize the model problem.
- If the input random field is *not Gaussian*, then the stochastic variables in the KL expansion are uncorrelated *but not necessarily independent*. For the purposes of mathematical analysis, we typically assume that the random variables in the input random field are independent so that we can parameterize the model problem. (Transforming dependent random variables into independent random variables can be done using, e.g., the Rosenblatt transformation, but this is computationally expensive.)

Note especially that in the Gaussian setting we do not need to make any “extra” effort to ensure the independence of the stochastic variables in the KL expansion.

[†]Matérn covariance is an especially popular choice.

Example (Lognormal input random field)

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a Lipschitz domain and consider the PDE problem

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}) & \text{for } \mathbf{x} \in D, \\ u(\cdot, \omega)|_{\partial D} = 0, \end{cases}$$

where $f: D \rightarrow \mathbb{R}$ is a fixed (deterministic) source term. We can model a lognormally distributed random diffusion coefficient $a: D \times \Omega \rightarrow \mathbb{R}$ using the KL expansion, e.g., as

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) \exp \left(\sum_{k=1}^{\infty} y_k(\omega) \psi_k(\mathbf{x}) \right), \quad y_k \sim \mathcal{N}(0, 1),$$

where $a_0 \in L^\infty(D)$ is such that $a_0(\mathbf{x}) > 0$ and the random variables y_k are uncorrelated (and thus independent in the Gaussian case).

Due to the independence, we can consider the above as a *parametric PDE* with $a(\mathbf{x}, \mathbf{y}) \equiv a(\mathbf{x}, \mathbf{y}(\omega))$ and $u(\mathbf{x}, \mathbf{y}) \equiv u(\mathbf{x}, \mathbf{y}(\omega))$, where (formally) $\mathbf{y} \in \mathbb{R}^N$ is a *parametric vector* endowed with a product Gaussian measure.

Example (Uniform and affine input random field)

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a Lipschitz domain, $f: D \rightarrow \mathbb{R}$ is a fixed (deterministic) source term, and consider the PDE problem

$$\begin{cases} -\nabla \cdot (a(\mathbf{x}, \omega) \nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}) & \text{for } \mathbf{x} \in D, \\ u(\cdot, \omega)|_{\partial D} = 0. \end{cases}$$

We can model a uniformly distributed random diffusion coefficient $a: D \times \Omega \rightarrow \mathbb{R}$ using the KL expansion, e.g., as

$$a(\mathbf{x}, \omega) = a_0(\mathbf{x}) + \sum_{k=1}^{\infty} y_k(\omega) \psi_k(\mathbf{x}), \quad y_k \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2}),$$

where the random variables y_k are uncorrelated. *Unlike the Gaussian setting, the random variables y_k are generally not independent!*

In numerical analysis, the random variables y_k are often **assumed** to be independent – this allows us to consider the above as a parametric PDE with $a(\mathbf{x}, \mathbf{y}) \equiv a(\mathbf{x}, \mathbf{y}(\omega))$ and $u(\mathbf{x}, \mathbf{y}) \equiv u(\mathbf{x}, \mathbf{y}(\omega))$, where $\mathbf{y} \in [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ is a *parametric vector* endowed with a uniform probability measure.

To estimate the statistical response, note that in the *lognormal model* the expected value of the PDE solution is given by

$$\mathbb{E}[u(\mathbf{x}, \cdot)] = \lim_{s \rightarrow \infty} \int_{\mathbb{R}^s} u(\mathbf{x}, \mathbf{y}) \prod_{j=1}^s \frac{e^{-\frac{1}{2}y_j^2}}{\sqrt{2\pi}} d\mathbf{y}$$

while in the *affine and uniform model* the expected value of the PDE solution is given by

$$\mathbb{E}[u(\mathbf{x}, \cdot)] = \lim_{s \rightarrow \infty} \int_{[-1/2, 1/2]^s} u(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

- In practice, we need to truncate these infinite-dimensional integrals into finite-dimensional ones, incurring the so-called *dimension truncation error*. Since the PDE is solved numerically using the finite element method, this also incurs a *finite element discretization error*.
- To compute the resulting high-dimensional integrals for the dimensionally-truncated, finite element discretized PDE solution we use a *quasi-Monte Carlo (QMC) method*.

Quasi-Monte Carlo (QMC) methods are a class of *equal weight* cubature rules

$$\int_{[0,1]^s} f(\mathbf{y}) d\mathbf{y} \approx \frac{1}{n} \sum_{i=1}^n f(\mathbf{t}_i),$$

where $(\mathbf{t}_i)_{i=1}^n$ is an ensemble of *deterministic* nodes in $[0, 1]^s$.

The nodes $(\mathbf{t}_i)_{i=1}^n$ are chosen deterministically.

QMC methods exploit the smoothness and anisotropy of an integrand in order to achieve better-than-Monte Carlo rates.

Lattice rules

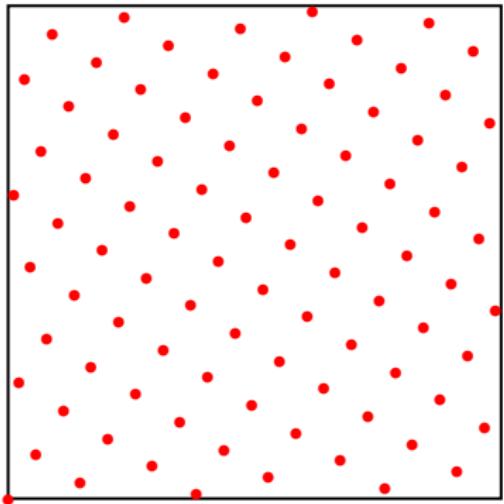
Rank-1 lattice rules

$$Q_{n,s}(f) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{t}_i)$$

have the points

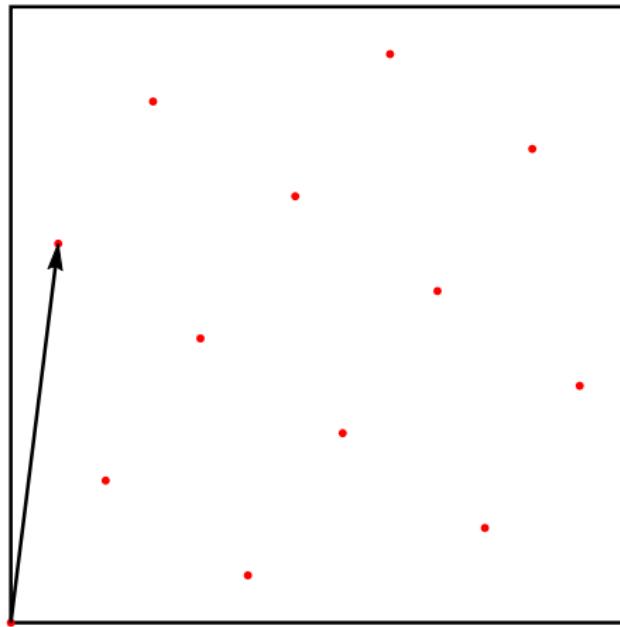
$$\mathbf{t}_i = \text{mod}\left(\frac{i\mathbf{z}}{n}, 1\right), \quad i \in \{1, \dots, n\},$$

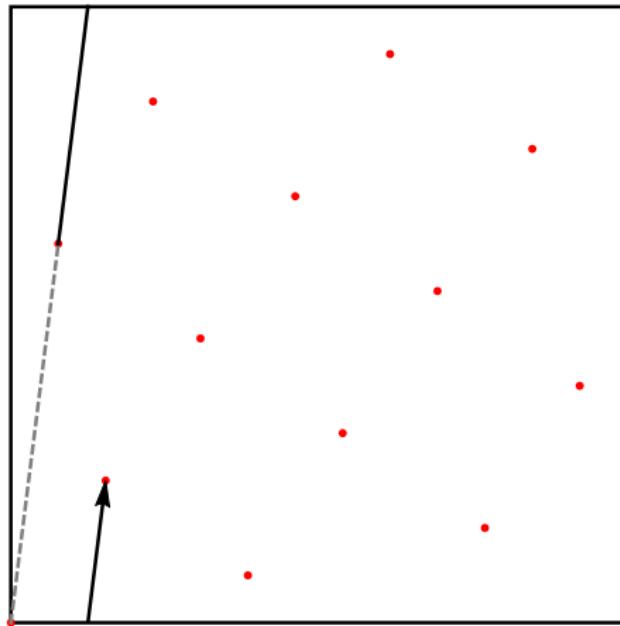
where the entire point set is determined by the *generating vector* $\mathbf{z} \in \mathbb{N}^s$, with all components *coprime* to n .

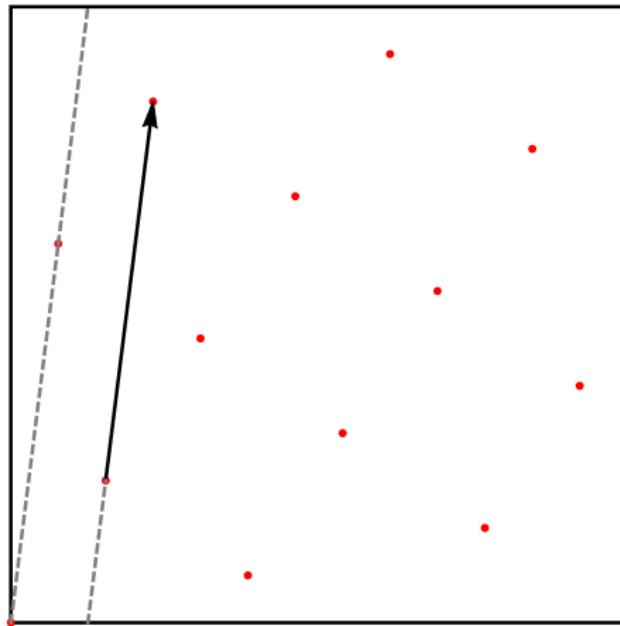


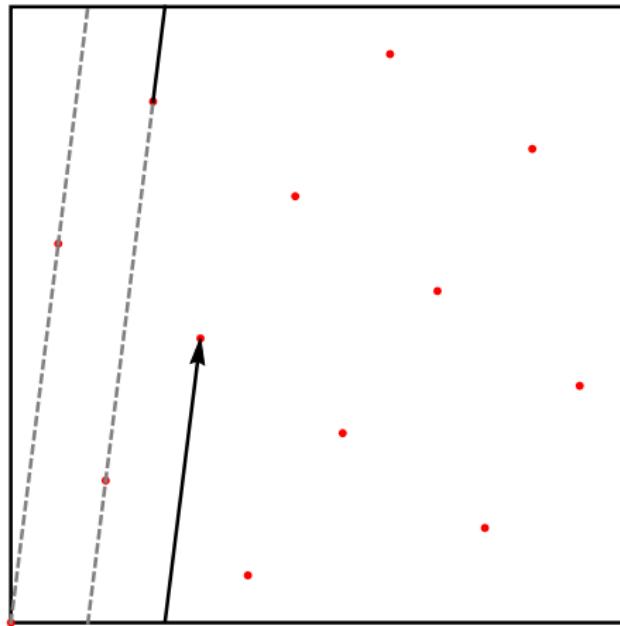
Lattice rule with $\mathbf{z} = (1, 55)$ and $n = 89$
nodes in $[0, 1]^2$

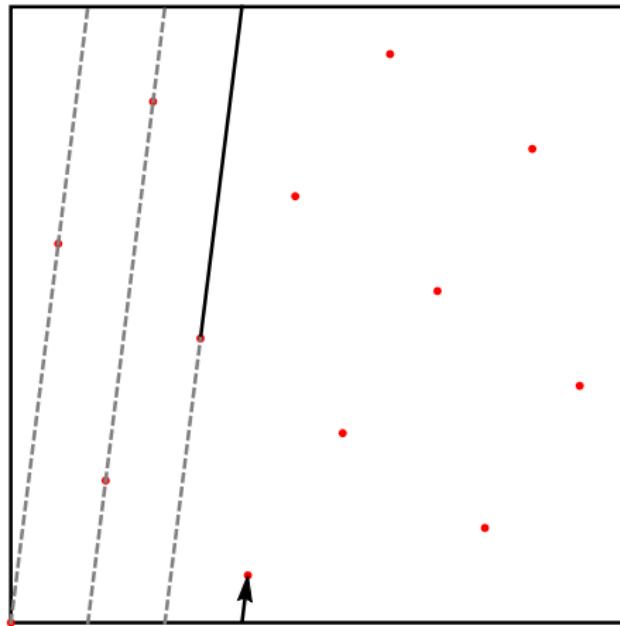
The quality of the lattice rule is determined by the choice of \mathbf{z} .

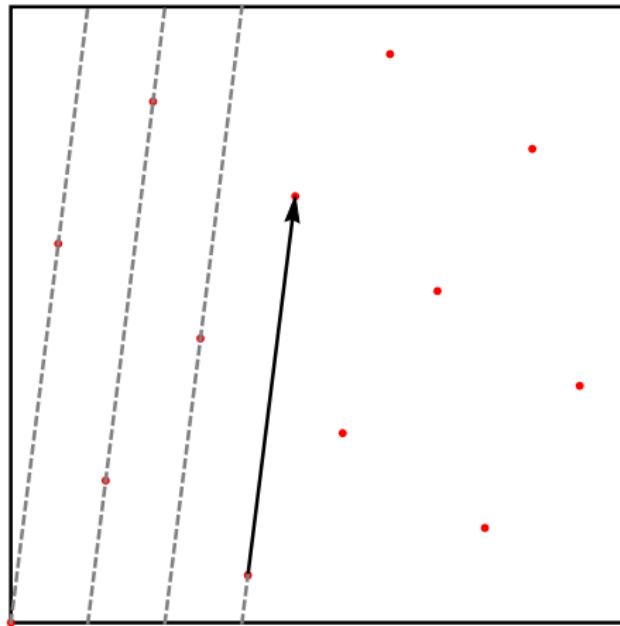


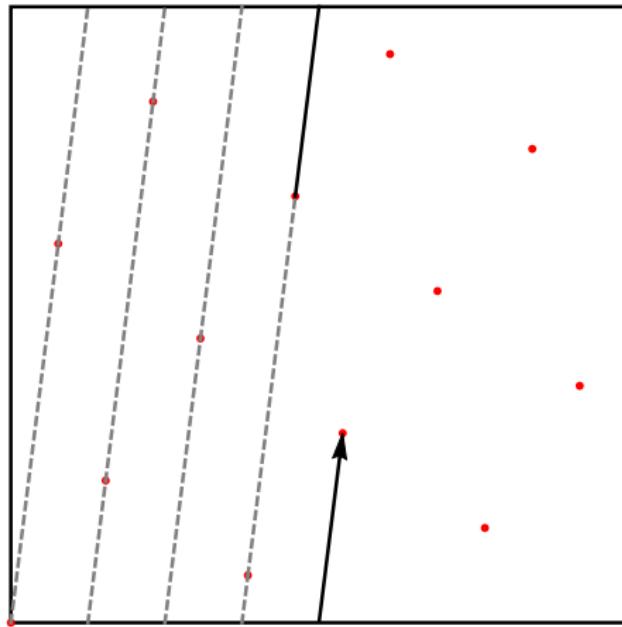


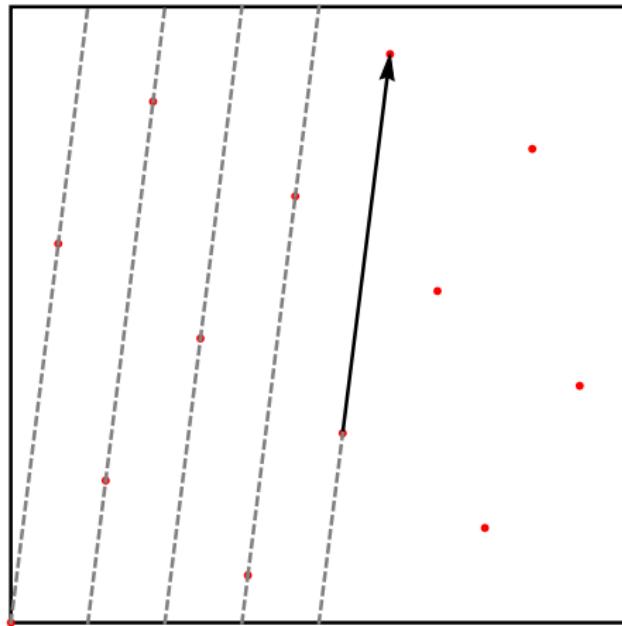


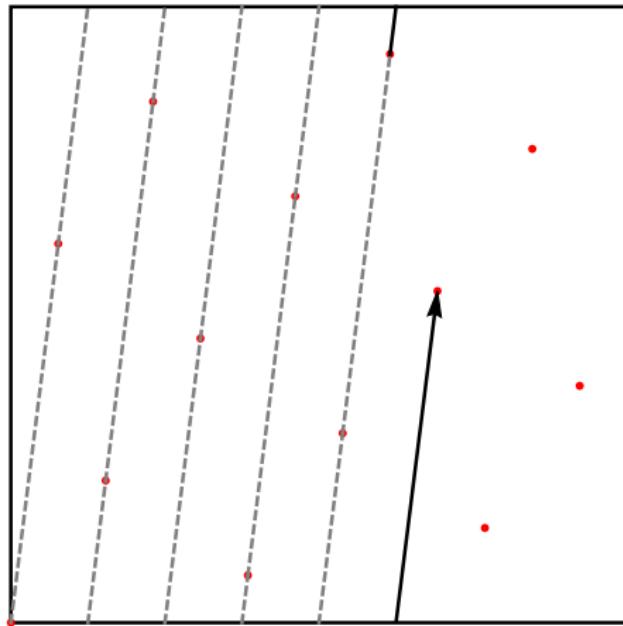


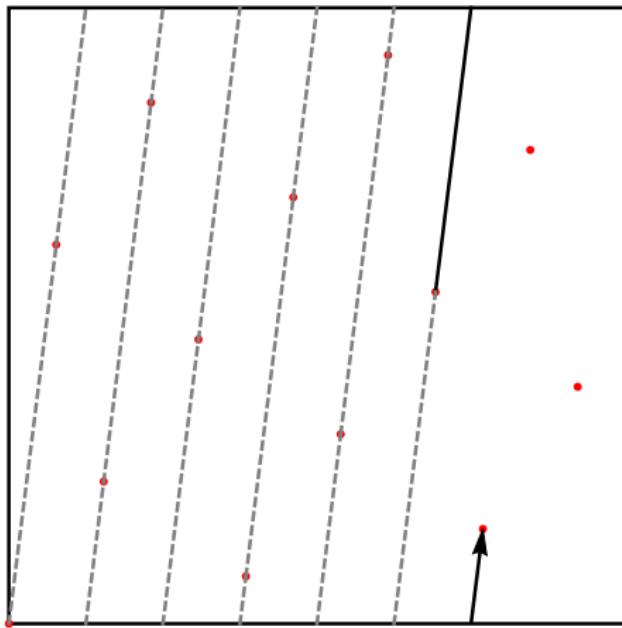


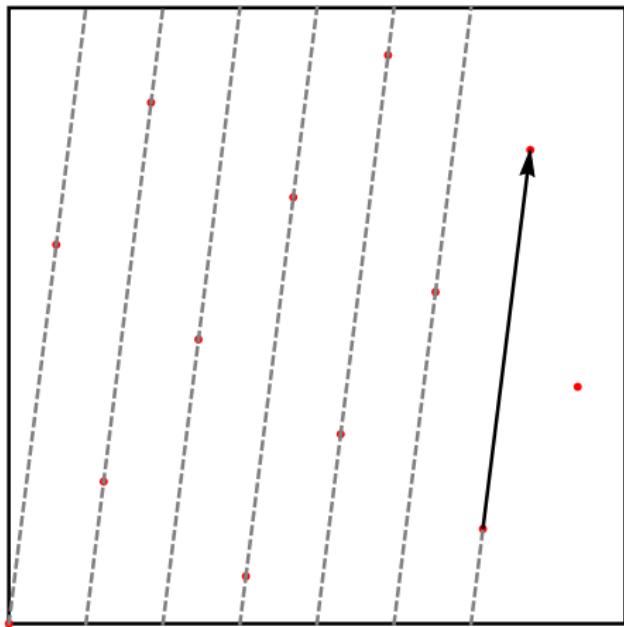


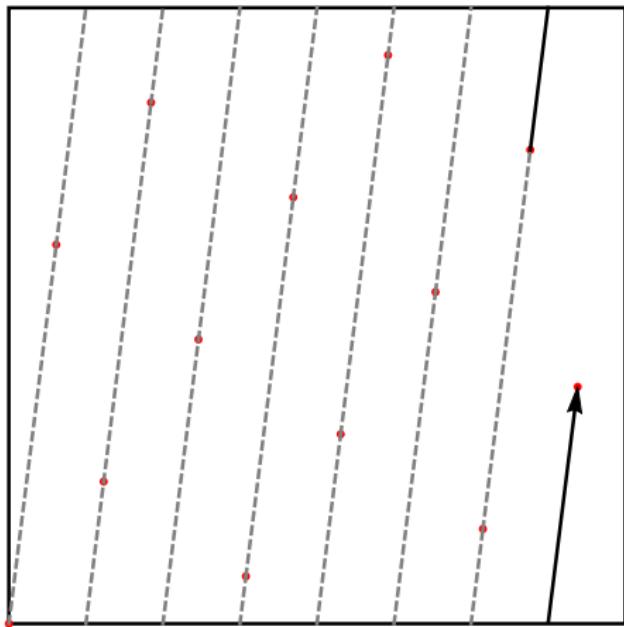












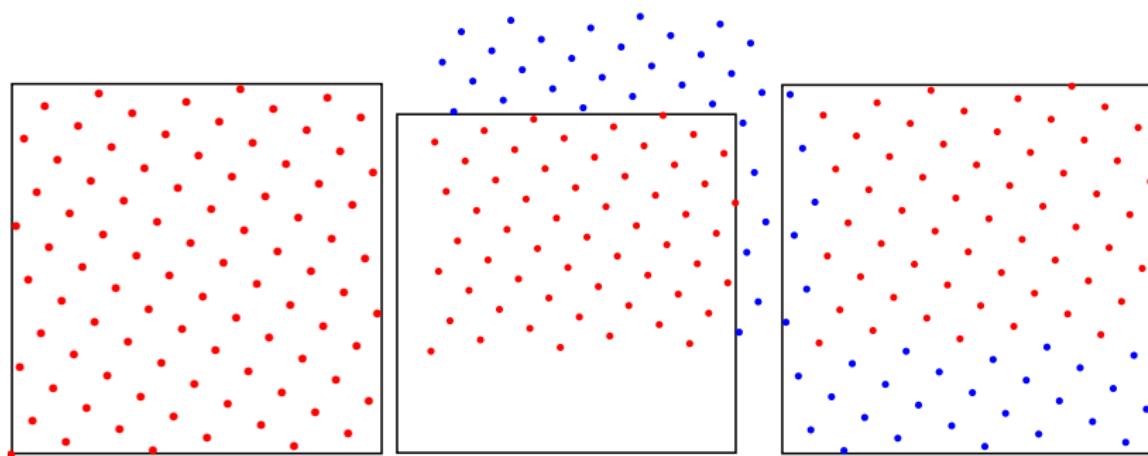
Randomly shifted lattice rules

Shifted rank-1 lattice rules have points

$$\mathbf{t}_i = \text{mod}\left(\frac{i\mathbf{z}}{n} + \boldsymbol{\Delta}, 1\right), \quad i \in \{1, \dots, n\}.$$

$\boldsymbol{\Delta} \in [0, 1)^s$ is the *shift*

Use a number of random shifts for error estimation.



Lattice rule shifted by $\boldsymbol{\Delta} = (0.1, 0.3)$.

Let Δ_r , $r = 1, \dots, R$, be independent random shifts drawn from $U([0, 1]^s)$ and define

$$Q_{n,s}^{\Delta_r}(f) := \frac{1}{n} \sum_{i=1}^n f(\text{mod}(\mathbf{t}_i + \Delta_r, 1)). \quad (\text{QMC rule with 1 random shift})$$

Then

$$\overline{Q}_{n,s}(f) = \frac{1}{R} \sum_{r=1}^R Q_{n,s}^{\Delta_r} f \quad (\text{QMC rule with } R \text{ random shifts})$$

is an unbiased estimator of $I_s(f)$.

Let $f: [0, 1]^s \rightarrow \mathbb{R}$ be sufficiently smooth.

Error bound (one random shift):

$$|I_s(f) - Q_{n,s}^{\Delta}(f)| \leq e_{n,s,\gamma}^{\Delta}(z) \|f\|_{\gamma}.$$

R.M.S. error bound (shift-averaged):

$$\sqrt{\mathbb{E}_{\Delta}[|I_s(f) - \bar{Q}_{n,s}(f)|^2]} \leq e_{n,s,\gamma}^{\text{sh}}(z) \|f\|_{\gamma}.$$

We consider weighted Sobolev spaces with dominating mixed smoothness, equipped with norm

$$\|f\|_{\gamma}^2 = \sum_{\mathbf{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathbf{u}}} \int_{[0,1]^{|\mathbf{u}|}} \left(\int_{[0,1]^{s-|\mathbf{u}|}} \frac{\partial^{|\mathbf{u}|} f}{\partial \mathbf{y}_{\mathbf{u}}}(\mathbf{y}) d\mathbf{y}_{-\mathbf{u}} \right)^2 d\mathbf{y}_{\mathbf{u}}$$

and (squared) worst case error

$$P(z) := e_{n,s,\gamma}^{\text{sh}}(z)^2 = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\emptyset \neq \mathbf{u} \subseteq \{1:s\}} \gamma_{\mathbf{u}} \prod_{j \in \mathbf{u}} \omega\left(\left\{ \frac{kz_j}{n} \right\}\right)$$

where $\omega(x) = x^2 - x + \frac{1}{6}$.

CBC algorithm (Sloan, Kuo, Joe 2002)

The idea of the *component-by-component* (CBC) algorithm is to find a good generating vector $\mathbf{z} = (z_1, \dots, z_s)$ by proceeding as follows:

1. Set $z_1 = 1$ (this is a freebie since $P(1) = P(z)$ for all $z \in \mathbb{N}$);
2. With z_1 fixed, choose z_2 to minimize error criterion $P(z_1, z_2)$;
3. With z_1 and z_2 fixed, choose z_3 to minimize error criterion $P(z_1, z_2, z_3)$
- ⋮
- The CBC algorithm is a *greedy algorithm*: in general, it will not find the generating vector \mathbf{z} that minimizes $P(\mathbf{z})$. However, it can be shown that the generating vector obtained by the CBC algorithm satisfies an error bound (see next slide).
- For generic $\boldsymbol{\gamma} = (\gamma_u)_{u \subseteq \{1:s\}}$, evaluating $P(\mathbf{z}) = P(\boldsymbol{\gamma}, \mathbf{z})$ takes $\mathcal{O}(2^s)$ operations. For an efficient implementation, it is desirable that the weights $\boldsymbol{\gamma}$ can be characterized by an expression that does not contain too many degrees of freedom.
- Efficient implementation using FFT! (QMC4PDE, QMCPy, etc.)

Theorem (CBC error bound)

The generating vector $\mathbf{z} \in \mathbb{U}_n^s$ constructed by the CBC algorithm, minimizing the squared shift-averaged worst-case error $[e_{n,s,\gamma}^{\text{sh}}(\mathbf{z})]^2$ for the weighted unanchored Sobolev space in each step, satisfies

$$[e_{n,s,\gamma}^{\text{sh}}(\mathbf{z})]^2 \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq \mathbf{u} \subseteq \{1:s\}} \gamma_{\mathbf{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathbf{u}|} \right)^{1/\lambda} \quad \text{for all } \lambda \in (1/2, 1],$$

where $\zeta(x) := \sum_{k=1}^{\infty} k^{-x}$ denotes the Riemann zeta function for $x > 1$.

Remarks:

- Optimal rate of convergence $\mathcal{O}(n^{-1+\delta})$ in weighted Sobolev spaces, independently of s under an appropriate condition on the weights.
- Cost of algorithm for POD weights is $\mathcal{O}(s n \log n + s^2 n)$ using FFT.
- Fast CBC works for any (composite) number $n \geq 2$, but the implementation is more involved when n is not prime.

Significance: Suppose that $f \in H_{s,\gamma}$ for all $\gamma = (\gamma_u)_{u \subseteq \{1:s\}}$. Then for any given sequence of weights γ , we can use the CBC algorithm to obtain a generating vector satisfying the error bound

$$\sqrt{\mathbb{E}_\Delta |I_s f - Q_{n,s}^\Delta f|^2} \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq u \subseteq \{1:s\}} \gamma_u^\lambda \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|u|} \right)^{1/(2\lambda)} \|f\|_{s,\gamma} \quad (2)$$

for all $\lambda \in (1/2, 1]$. We can use the following strategy:

- For a given integrand f , estimate the norm $\|f\|_{s,\gamma}$.
- Find weights γ which *minimize* the error bound (2).
- Using the optimized weights γ as input, use the CBC algorithm to find a generating vector which *satisfies* the error bound (2).

Remarks:

- If n is prime, then $\frac{1}{\varphi(n)} = \frac{1}{n-1}$. If $n = 2^k$, then $\frac{1}{\varphi(n)} = \frac{2}{n}$. For general (composite) $n \geq 3$, $\frac{1}{\varphi(n)} \leq \frac{e^\gamma \log \log n + \frac{3}{\log \log n}}{n}$, where $\gamma = 0.57721566\dots$ (Euler–Mascheroni constant).
- The optimal convergence rate close to $\mathcal{O}(n^{-1})$ is obtained with $\lambda \rightarrow 1/2$, but $\lambda = 1/2$ is not permitted since $\zeta(2\lambda) \xrightarrow{\lambda \rightarrow 1/2^+} \infty$.

Example: applying QMC theory for a simplified parametric PDE

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a convex, bounded Lipschitz domain and consider the following (simplified!) elliptic PDE

$$\begin{cases} -\nabla \cdot (a(\mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) = f(\mathbf{x}), & \mathbf{x} \in D, \quad \mathbf{y} \in [-1/2, 1/2]^s, \\ u(\mathbf{x}, \mathbf{y}) = 0, & \mathbf{x} \in \partial D, \quad \mathbf{y} \in [-1/2, 1/2]^s, \end{cases}$$

where the source term $f \in L^2(D)$ is fixed and

$$a(\mathbf{y}) := 1 + \sum_{j=1}^s \beta_j y_j, \quad y_j \in [-1/2, 1/2],$$

where $\beta_j \geq 0$ are assumed to be *constants* for all $j \geq 1$ (i.e., independent of \mathbf{x}) s.t. $a(\mathbf{y}) \geq a_{\min} > 0$ for all $\mathbf{y} \in [-1/2, 1/2]^s$ and $\sum_{j=1}^{\infty} \beta_j^p < \infty$ for some $p \in (0, 1)$. Due to the linearity of the PDE problem, we can write

$$u(\mathbf{x}, \mathbf{y}) = \frac{g(\mathbf{x})}{1 + \sum_{j=1}^s \beta_j y_j}, \quad \text{where } \begin{cases} -\Delta g(\mathbf{x}) = f(\mathbf{x}), & \mathbf{x} \in D, \\ g|_{\partial D} = 0. \end{cases}$$

Note that the Poisson problem has a continuous solution $g \in C(D)$.

Clearly,

$$\mathbb{E}[u(\mathbf{x}, \cdot)] = g(\mathbf{x}) \int_{[-1/2, 1/2]^s} \underbrace{\frac{1}{1 + \sum_{j=1}^s \beta_j y_j}}_{=: F(\mathbf{y})} d\mathbf{y}.$$

(Note the similarity to exercise 2 of week 8!)

Steps of QMC analysis:

- Estimate the (parametric) derivatives $\partial^\nu F(\mathbf{y})$.
- Using the above, estimate $\|F(\cdot - \frac{1}{2})\|_{s, \gamma}$.
- Plug the weighted Sobolev norm into QMC error bound and choose the weights $\gamma = (\gamma_u)_{u \subseteq \{1:s\}}$ to minimize the resulting error bound.
 - The resulting weights are used as inputs to the CBC algorithm. The generating vector (and the resulting randomly shifted QMC point set) are guaranteed to satisfy the rigorous CBC error bound.
- Analysis: is the coefficient of the CBC error bound independent of the dimension s with the chosen weights?

Step 1: *Parametric regularity.* It is not difficult to see that

$$\frac{\partial^{|\mathbf{u}|}}{\partial \mathbf{y}_{\mathbf{u}}} F(\mathbf{y}) = |\mathbf{u}|! F(\mathbf{y})^{|\mathbf{u}|+1} \prod_{j \in \mathbf{u}} (-\beta_j) \quad \text{for all } \mathbf{u} \subseteq \{1 : s\}.$$

Exploiting the fact that we assumed before that $1 + \sum_{j=1}^s \beta_j y_j \geq a_{\min} > 0$ for all $\mathbf{y} \in [-1/2, 1/2]^s$, we can define

$$b_j := \frac{\beta_j}{a_{\min}} \quad \text{for all } j \geq 1,$$

and estimate the parametric regularity of the first order mixed partial derivatives as

$$\left| \frac{\partial^{|\mathbf{u}|}}{\partial \mathbf{y}_{\mathbf{u}}} F(\mathbf{y}) \right| \leq \frac{1}{a_{\min}} |\mathbf{u}|! \prod_{j \in \mathbf{u}} b_j \quad \text{for all } \mathbf{u} \subseteq \{1 : s\}.$$

Step 2: Estimate the weighted Sobolev norm. It is easy to see that

$$\|F(\cdot - \frac{1}{2})\|_{s,\gamma}^2 \lesssim \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \prod_{j \in \mathfrak{u}} b_j^2.$$

Step 3: Plugging this into the CBC error bound

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \leq \left(\frac{1}{\varphi(n)} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/(2\lambda)} \|F(\cdot - \frac{1}{2})\|_{s,\gamma}$$

yields

$$\begin{aligned} \sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} &\lesssim \left(\frac{1}{\varphi(n)} \right)^{1/(2\lambda)} \left(\sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/(2\lambda)} \\ &\quad \times \left(\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \prod_{j \in \mathfrak{u}} b_j^2 \right)^{1/2}. \end{aligned}$$

(We have separated the dependence on the number of QMC nodes n since this is unaffected by the choice of weights. The weights only affect the constant in the error bound, which we try to minimize next.)

Step 4: Choosing the weights. Note that the square of the objective functional has the form

$$g(\gamma) := \left(\sum_{\mathfrak{u}} \alpha_{\mathfrak{u}} \gamma_{\mathfrak{u}}^{\lambda} \right)^{1/\lambda} \left(\sum_{\mathfrak{u}} \beta_{\mathfrak{u}} \gamma_{\mathfrak{u}}^{-1} \right),$$

which is minimized by

$$\gamma_{\mathfrak{u}} := c \left(\frac{\beta_{\mathfrak{u}}}{\alpha_{\mathfrak{u}}} \right)^{1/(1+\lambda)} \quad \text{for arbitrary } c > 0.$$

(In fact, with $c = 1$, the minimizer is equivalent to setting the summands equal: $\alpha_{\mathfrak{u}} \gamma_{\mathfrak{u}}^{\lambda} = \beta_{\mathfrak{u}} \gamma_{\mathfrak{u}}^{-1}$.)

Thus the minimizing weights for our problem are the *product-and-order (POD) dependent* weights:

$$\gamma_{\mathfrak{u}} := \left(|\mathfrak{u}|! \prod_{j \in \mathfrak{u}} \frac{b_j}{\sqrt{\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}}}} \right)^{2/(1+\lambda)}, \quad \mathfrak{u} \subseteq \{1 : s\}.$$

(The POD form is important since it doesn't contain "too many degrees of freedom": the cost of fast CBC used to find the generating vector satisfying the CBC error bound is $\mathcal{O}(s n \log n + s^2 n)$ with these weights.)

Step 5: Plugging the optimized POD weights into the QMC error bound results in

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \lesssim \left(\frac{1}{\varphi(n)} \right)^{1/(2\lambda)} C(s, \gamma, \lambda)^{(1+\lambda)/(2\lambda)},$$

where

$$C(s, \gamma, \lambda) := \sum_{\mathfrak{u} \subseteq \{1:s\}} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda} \right)^{|\mathfrak{u}|/(1+\lambda)} (|\mathfrak{u}|!)^{2\lambda/(1+\lambda)} \prod_{j \in \mathfrak{u}} b_j^{2\lambda/(1+\lambda)}.$$

In complete analogy to the 11th lecture, we have the following:

Lemma

By choosing

$$\lambda = \begin{cases} \frac{p}{2-p} & \text{when } p \in (2/3, 1) \\ \frac{1}{2-2\delta} \text{ for arbitrary } \delta \in (0, 1/2) & \text{when } p \in (0, 2/3], \end{cases}$$

*there exists a constant $C(\gamma, \lambda) < \infty$ independently of s
s.t. $C(s, \gamma, \lambda) \leq C(\gamma, \lambda) < \infty$.*

Using randomly shifted rank-1 lattice rules to estimate the integral

$$\int_{[-1/2,1/2]^s} F(\mathbf{y}) d\mathbf{y}, \quad F(\mathbf{y}) := \frac{1}{1 + \sum_{j=1}^s \beta_j y_j},$$

we can conclude the following:

For arbitrary $\delta \in (0, 1/2)$, we can choose the POD weights

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

as inputs to the CBC algorithm to obtain a generating vector. If the number of QMC nodes n is prime or a prime power, then the resulting randomly shifted rank-1 lattice rule satisfies the root-mean-square error bound

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \lesssim n^{\max\{-1/p+1/2, -1+\delta\}}, \quad (3)$$

where the implied coefficient is independent of the dimension s .

Note that this rate is always better than Monte Carlo, but cannot exceed linear convergence $\mathcal{O}(n^{-1})$ (i.e., double the Monte Carlo rate).

Uniform and affine model

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 $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient has the parameterization

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

where $a_0 \in L^\infty(D)$, there exist $a_{\min}, a_{\max} > 0$

s.t. $0 < a_{\min} \leq a(\mathbf{x}, \mathbf{y}) \leq a_{\max} < \infty$ for all $\mathbf{x} \in D$ and $\mathbf{y} \in U$, and the *stochastic fluctuations* $\psi_j: D \rightarrow \mathbb{R}$ are functions of the spatial variable such that

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

Proposition (Parametric regularity for the uniform and affine model)

For all $\mathbf{y} \in [-1/2, 1/2]^{\mathbb{N}}$ and $\boldsymbol{\nu} \in \mathcal{F}$, there holds

$$\|\partial^{\boldsymbol{\nu}} u(\cdot, \mathbf{y})\|_{H_0^1(D)} \leq \frac{C_P \|f\|_{L^2(D)}}{a_{\min}} \mathbf{b}^{\boldsymbol{\nu}} |\boldsymbol{\nu}|!,$$

where C_P is the Poincaré constant satisfying $\|v\|_{L^2(D)} \leq C_P \|v\|_{H_0^1(D)}$ for all $v \in H_0^1(D)$.

This parametric regularity bound is valid also for the dimensionally-truncated finite element solution $u_{s,h}$. If $G: H_0^1(D) \rightarrow \mathbb{R}$ is a bounded linear functional and we define $F(\mathbf{y}) := G(u_{s,h}(\cdot, \mathbf{y} - \frac{1}{2}))$ for $\mathbf{y} \in [0, 1]^s$, then

$$\|F\|_{s,\gamma}^2 \lesssim \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} (|\mathfrak{u}|!)^2 \prod_{j \in \mathfrak{u}} b_j^2,$$

and using the POD weights (3) as inputs to the CBC algorithm yields a randomly shifted rank-1 lattice rule satisfying the R.M.S. error

$$\sqrt{\mathbb{E}_{\Delta} |I_s F - Q_{n,s}^{\Delta} F|^2} \lesssim n^{\max\{-1/p+1/2, -1+\delta\}},$$

where the constant is independent of the dimension.

Of course, the truncation of the input random series and the finite element discretization incur additional errors.

- If $\|\psi_1\|_{L^\infty(D)} \geq \|\psi_2\|_{L^\infty(D)} \cdots$, then the error resulting from the dimension truncation has order $\mathcal{O}(s^{-2/p+1})$, where the constant is independent of s .
- If $D \subset \mathbb{R}^d$ is a bounded, convex polyhedron, a_0 and ψ_j are Lipschitz for all $j \geq 1$ with $\sum_{j=1}^{\infty} \|\psi_j\|_{W^{1,\infty}(D)} < \infty$, and $G : L^2(D) \rightarrow \mathbb{R}$ is a bounded linear functional, then—if the FE mesh has been obtained from an initial, regular triangulation of D by recursive, uniform bisection of simplices—the L^2 finite element error has order $\mathcal{O}(h^2)$, where $h > 0$ is the mesh size and the implied constant is independent of y , s , and h .

Lognormal model

Lognormal model: let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded Lipschitz domain, and let $f \in H^{-1}(D)$. Let $\psi_j \in L^\infty(D)$ and $b_j := \|\psi_j\|_{L^\infty}$ for $j \in \mathbb{N}$ such that $\sum_{j=1}^{\infty} b_j < \infty$, and set

$$U_b := \left\{ \mathbf{y} \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} b_j |y_j| < \infty \right\}.$$

Consider the problem of finding, for all $\mathbf{y} \in U$, $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient is assumed to have the parameterization

$$a(\mathbf{x}, \mathbf{y}) := a_0(\mathbf{x}) \exp \left(\sum_{j=1}^{\infty} y_j \psi_j(\mathbf{x}) \right), \quad \mathbf{x} \in D, \quad \mathbf{y} \in U_b,$$

where $a_0 \in L^\infty(D)$ is such that $a_0(\mathbf{x}) > 0$, $\mathbf{x} \in D$.

Standing assumptions for the lognormal model

- (B1) We have $a_0 \in L^\infty(D)$ and $\sum_{j=1}^{\infty} b_j < \infty$.
- (B2) For every $\mathbf{y} \in U_b$, the expressions $a_{\max}(\mathbf{y}) := \max_{\mathbf{x} \in \bar{D}} a(\mathbf{x}, \mathbf{y})$ and $a_{\min}(\mathbf{y}) := \min_{\mathbf{x} \in \bar{D}} a(\mathbf{x}, \mathbf{y})$ are well-defined and satisfy $0 < a_{\min}(\mathbf{y}) \leq a(\mathbf{x}, \mathbf{y}) \leq a_{\max}(\mathbf{y}) < \infty$.
- (B3) $\sum_{j=1}^{\infty} b_j^p < \infty$ for some $p \in (0, 1)$.

Remark: Note that in the lognormal case, $a(\mathbf{x}, \mathbf{y})$ can take values which are arbitrarily close to 0 or arbitrarily large. Thus, the best we can do is to find \mathbf{y} -dependent lower and upper bounds $a_{\min}(\mathbf{y})$ and $a_{\max}(\mathbf{y})$. This will lead to a \mathbf{y} -dependent *a priori* bound and, consequently, \mathbf{y} -dependent parametric regularity bounds. This will make the QMC analysis more involved, leading one to consider “special” weighted, unanchored Sobolev spaces.

In this setting, we have

$$I_s(F) := \int_{\mathbb{R}^s} F(\mathbf{y}) \prod_{j=1}^s \phi(y_j) d\mathbf{y} = \int_{(0,1)^s} F(\Phi^{-1}(\mathbf{w})) d\mathbf{w}.$$

where $\phi(y) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$ is the probability density function of $\mathcal{N}(0, 1)$ and $\Phi^{-1}(\mathbf{w}) = [\Phi^{-1}(w_1), \dots, \Phi^{-1}(w_s)]^T$ denotes the corresponding (componentwise) inverse cumulative distribution function. We use the randomly shifted QMC rules

$$Q_{n,s}^{\Delta_r}(F) = \frac{1}{n} \sum_{k=1}^n F(\Phi^{-1}(\{\mathbf{t}_k + \Delta_r\})),$$

$$\overline{Q}_{n,R}(F) := \frac{1}{R} \sum_{r=1}^R Q_{n,s}^{\Delta_r}(F),$$

where we have R independent random shifts $\Delta_1, \dots, \Delta_R$ drawn from $\mathcal{U}([0, 1]^s)$, $\mathbf{t}_k := \{\frac{k\mathbf{z}}{n}\}$, with generating vector $\mathbf{z} \in \mathbb{N}^s$.

The appropriate function space for unbounded integrands is a “special” weighted, unanchored Sobolev space equipped with the norm

$$\|F\|_{s,\gamma} = \left[\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \left(\int_{\mathbb{R}^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \mathbf{y}_{\mathfrak{u}}} F(\mathbf{y}) \left(\prod_{j \in \{1:s\} \setminus \mathfrak{u}} \phi(y_j) \right) d\mathbf{y}_{-\mathfrak{u}} \right)^2 \times \left(\prod_{j \in \mathfrak{u}} \varpi_j^2(y_j) \right) d\mathbf{y}_{\mathfrak{u}} \right]^{1/2}$$

where we have the weights

$$\varpi_j^2(y) := \exp(-2\alpha_j|y_j|), \quad \alpha_j > 0.$$

Theorem (Graham, Kuo, Nichols, Scheichl, Schwab, Sloan (2015))

Let F belong to the special weighted space over \mathbb{R}^s with weights γ , with ϕ being the standard normal density, and the weight functions ϖ_j defined as above. A randomly shifted lattice rule in s dimensions with n being a prime power can be constructed by a CBC algorithm such that

$$\sqrt{\mathbb{E}_{\Delta}|I_s F - Q_{n,s}^{\Delta} F|^2} \leq \left(\frac{2}{n} \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \prod_{j \in \mathfrak{u}} \varrho_j(\lambda) \right)^{1/(2\lambda)} \|F\|_{s,\gamma},$$

where $\lambda \in (1/2, 1]$ and

$$\varrho_j(\lambda) = 2 \left(\frac{\sqrt{2\pi} \exp(\alpha_j^2/\eta_*)}{\pi^{2-2\eta_*} (1-\eta_*) \eta_*} \right)^{\lambda} \zeta(\lambda + \tfrac{1}{2}) \quad \text{and} \quad \eta_* = \frac{2\lambda - 1}{4\lambda},$$

with $\zeta(x) := \sum_{k=1}^{\infty} k^{-x}$ denoting the Riemann zeta function for $x > 1$.

The steps for QMC analysis are the same as in the uniform case: (1) estimate $\|\cdot\|_{s,\gamma}$ for a given integrand (2) find weights γ which minimize the upper bound (3) plug the weights into the new error bound and estimate the constant (which ideally can be bounded independently of s). 370

Proposition (Parametric regularity bound for the lognormal model)

For all $\mathbf{y} \in U_b$ and $\nu \in \mathcal{F}$, there holds

$$\|\partial^\nu u(\cdot, \mathbf{y})\|_{H_0^1(D)} \leq \frac{C_P \|f\|_{L^2(D)}}{\min_{x \in \bar{D}} a_0(x)} \frac{|\nu|!}{(\log 2)^{|\nu|}} b^\nu \prod_{j \geq 1} \exp(b_j |y_j|).$$

This parametric regularity bound is valid also for the dimensionally-truncated finite element solution $u_{s,h}$. If $G: H_0^1(D) \rightarrow \mathbb{R}$ is a bounded linear functional and $F(\mathbf{y}) := G(u_{s,h}(\cdot, \mathbf{y}))$ for $\mathbf{y} \in \mathbb{R}^s$, then

$$\|F\|_{s,\gamma}^2 \leq \sum_{\mathbf{u} \subseteq \{1:s\}} \frac{(|\mathbf{u}|!)^2}{\gamma_{\mathbf{u}}} \left(\prod_{j=1}^s 2 \exp(2b_j^2) \Phi(2b_j) \right) \left(\prod_{j \in \mathbf{u}} \frac{b_j^2}{2(\log 2)^2 \exp(2b_j^2) \Phi(2b_j) (\alpha_j - b_j)} \right).$$

By choosing $\alpha_j = \frac{1}{2}(b_j + \sqrt{b_j^2 + 1 - \frac{1}{2\lambda}})$ and using the POD weights

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

as inputs to the CBC algorithm yields a randomly shifted rank-1 lattice rule satisfying the R.M.S. error

$$\sqrt{\mathbb{E}_\Delta |I_s F - Q_{n,s}^\Delta F|^2} \lesssim n^{\max\{-1/p+1/2, -1+\delta\}},$$

where the constant is independent of the dimension.

Similarly to the uniform and affine setting, the truncation of the input random series and the finite element discretization incur a *dimension truncation error* and a *finite element discretization error*, respectively. However, the analysis is more complicated in the lognormal case and has been omitted.

Computational implementation

Consider the task of approximating $\int_{[0,1]^s} f(\mathbf{y}) d\mathbf{y}$ using a randomly shifted rank-1 lattice rule with R random shifts.

Once a generating vector $\mathbf{z} \in \mathbb{N}^s$ has been obtained for a given number n of QMC nodes and dimension s (using, e.g., the CBC algorithm), then:

Remarks:

```
for r = 1, ..., R, do
    draw Δ(r) ~ U([0, 1]s);
    initialize Qr = 0;
    for i = 1, ..., n, do
        set ti = mod(iz/n + Δ(r), 1);
        set Qr = Qr + f(ti);
    end for
    set Qr = Qr/n;
end for
return Q̄ = Q1 + ... + QR/R;
(This is the QMC estimator
with R random shifts.)
```

- If integrating

$$\int_{\mathbb{R}^s} f(\mathbf{y}) \prod_{j=1}^s \frac{e^{-\frac{1}{2}y_j^2}}{\sqrt{2\pi}} d\mathbf{y}$$

then use $t_i = \Phi^{-1}(\text{mod}(iz/n + \Delta^{(r)}, 1))$, where Φ^{-1} is the (componentwise) inverse cumulative distribution function of $\mathcal{N}(0, 1)$.

- The R.M.S. error can be estimated by

R.M.S. error

$$\approx \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^R (\bar{Q} - Q_r)^2}.$$

The end!