Quasi-Monte Carlo methods for PDE uncertainty quantification

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

Winter School of SIAM/GAMM Student Chapter Berlin

February 21–22, 2024

Practical matters

- Lecture notes and MATLAB programs are available at https://www.iki.fi/vesakaar/winterschool24
- Schedule:

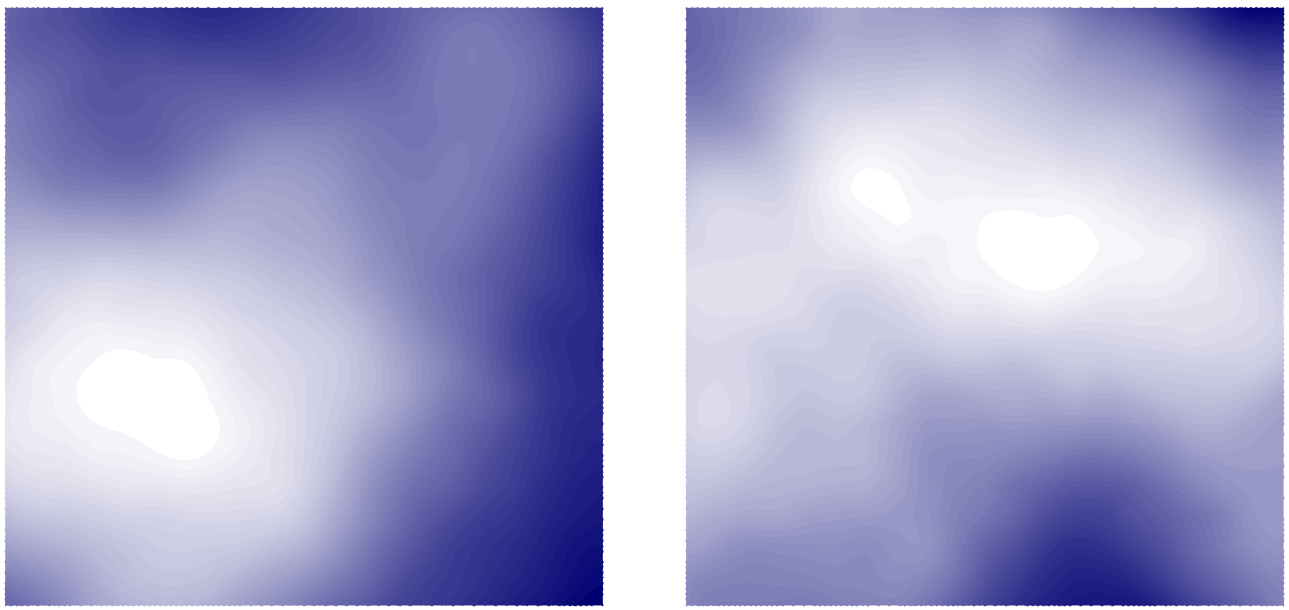
	Wednesday	**Thursday**
09:10-09:30	Coffee/Opening	Coffee
09:30-10:50	V. Kaarnioja	D. Walter
10:50-11:10	Coffee Break	Coffee Break
11:10-12:30	V. Kaarnioja	D. Walter
12:30-13:30	Lunch Break	Lunch Break
13:30-14:50	D. Walter	V. Kaarnioja
14:50-15:10	Coffee Break	Coffee Break
15:10-16:30	D. Walter	V. Kaarnioja

Uncertainty in groundwater flow

Risk analysis of radwaste disposal or CO_2 sequestration.

Darcy's law	$\boldsymbol{q}(\boldsymbol{x}) + \boldsymbol{a}(\boldsymbol{x})\nabla \boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})$
mass conservation law	$\nabla \cdot \boldsymbol{q}(\boldsymbol{x}) = 0$

in $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$
together with boundary conditions

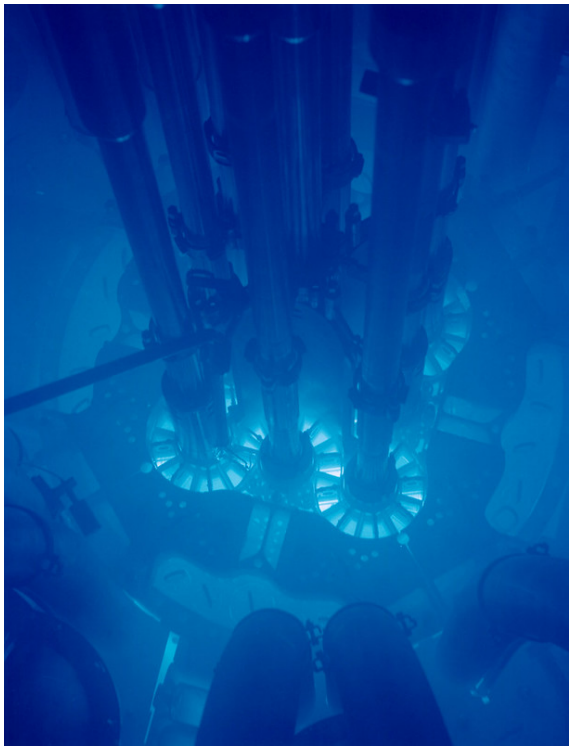


Uncertainty in $\boldsymbol{a}(\boldsymbol{x}, \omega)$ leads to uncertainty in $\boldsymbol{q}(\boldsymbol{x}, \omega)$ and $\boldsymbol{p}(\boldsymbol{x}, \omega)$

Criticality problem for nuclear reactors

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

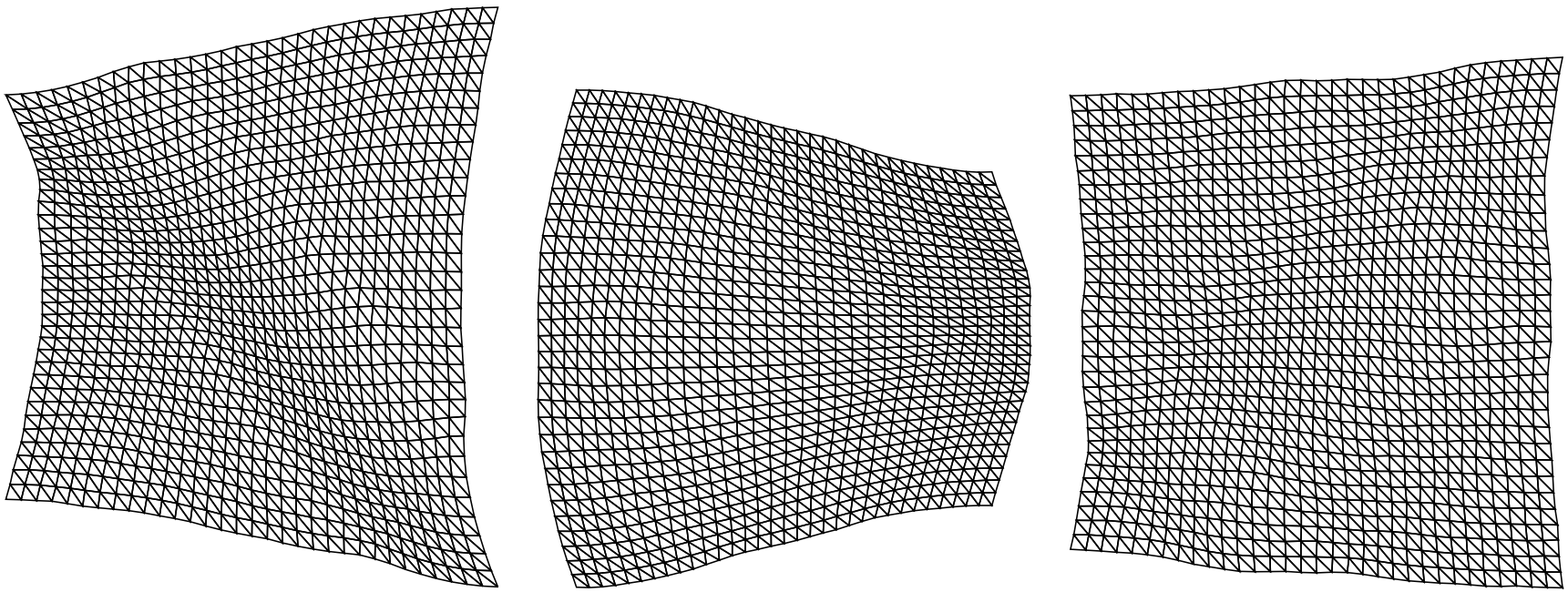
- The smallest eigenvalue $\lambda_1 \in \mathbb{R}$ measures *criticality* of a reactor.
- Eigenfunction $u_1(\boldsymbol{x})$ is the *neutron flux* at the point $\boldsymbol{x}$.



Source: Argonne National Laboratory on Flickr

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

Domain uncertainty quantification

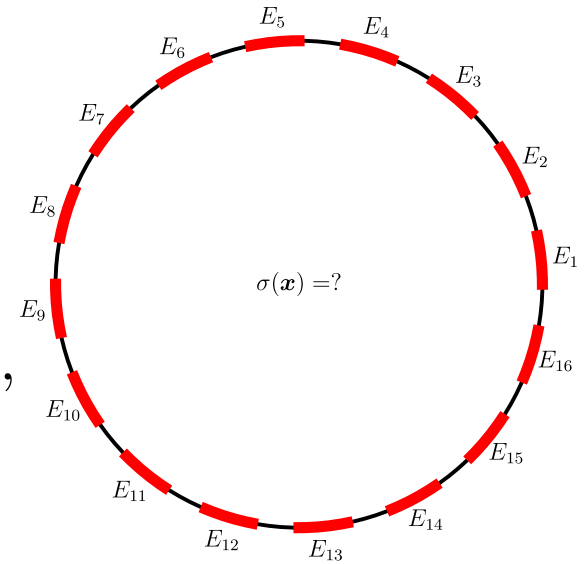


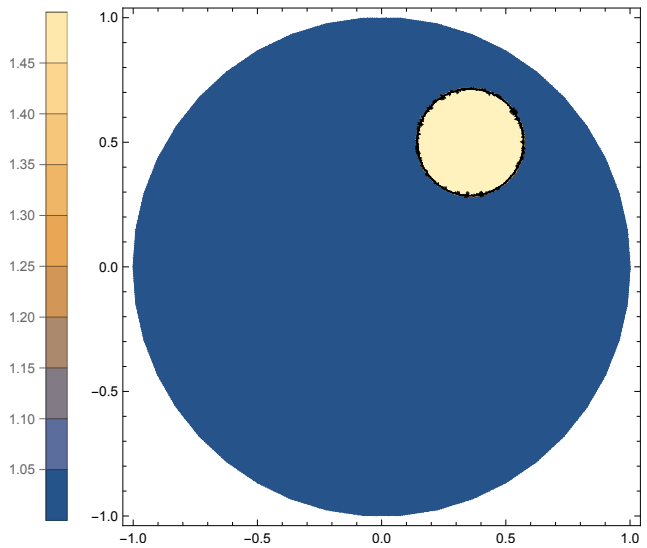
Three realizations of a random spatial domain

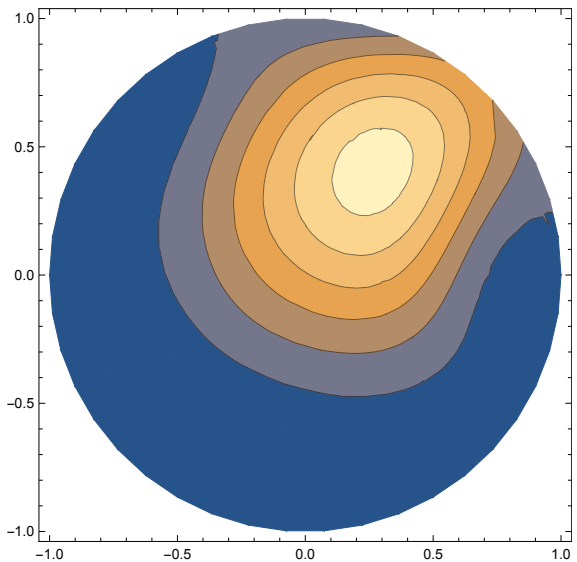
Electrical impedance tomography

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

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







Consider the elliptic PDE problem:

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

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.
$\Rightarrow$ 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

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

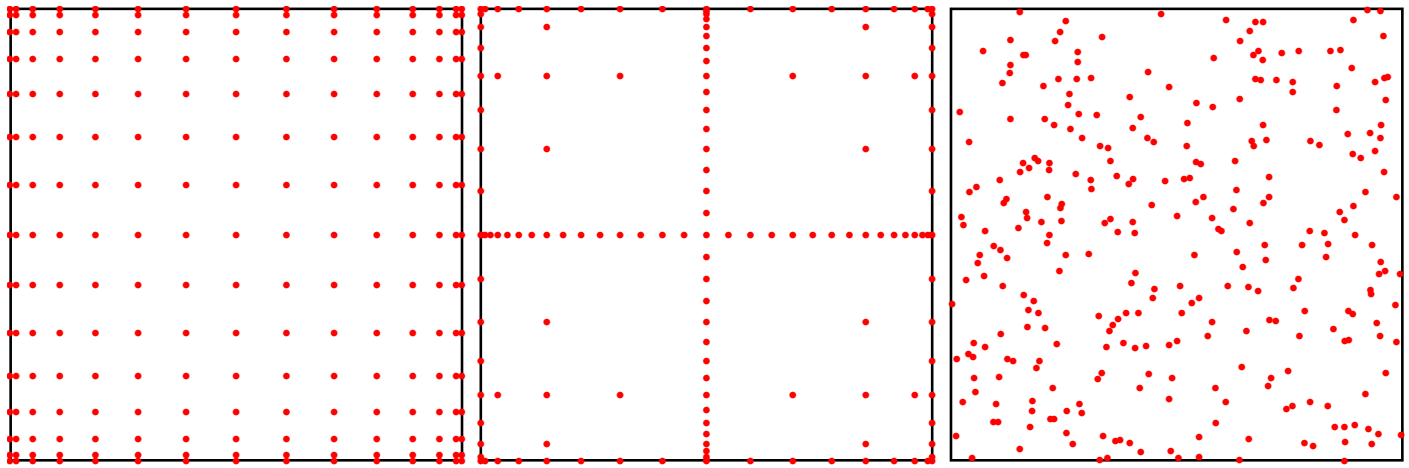


Figure: Tensor product grid, sparse grid, Monte Carlo nodes (not QMC rules)

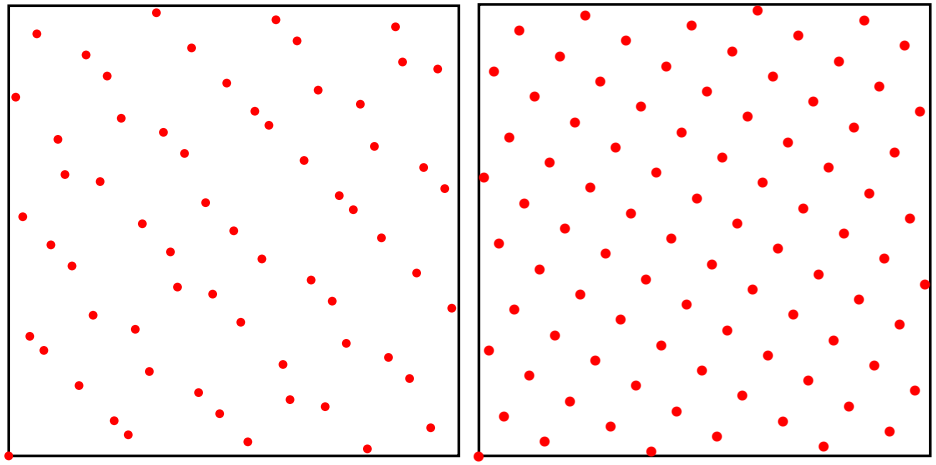


Figure: Sobol′ points, lattice rule (examples of QMC rules)

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

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

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

The nodes $(\boldsymbol{t}_i)_{i=1}^n$ 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.

Table of contents

General reference on QMC:


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

QMC for PDE uncertainty quantification:


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. https://doi.org/10.1007/s10208-016-9329-5


F. Y. Kuo and D. Nuyens. Application of quasi-Monte Carlo methods to PDEs with random coefficients – an overview and tutorial. In: A. B. Owen, P. W. Glynn (eds.), *Monte Carlo and Quasi-Monte Carlo Methods 2016*, 53–71, 2018. https://doi.org/10.1007/978-3-319-91436-7_3

QMC for Bayesian inverse problems:


R. Scheichl and A. M. Stuart and A. L. Teckentrup. Quasi-Monte Carlo and multilevel Monte Carlo methods for computing posterior expectations in elliptic inverse problems. *SIAM/ASA J. Uncertain. Quantif.* **5**(1):493–518, 2017. https://doi.org/10.1137/16M1061692


J. Dick, R. N. Gantner, Q. T. Le Gia, and Ch. Schwab. Higher order quasi-Monte Carlo integration for Bayesian PDE inversion. *Comput. Math. Appl.* **77**(1):144–172, 2019. https://doi.org/10.1016/j.camwa.2018.09.019


R. N. Gantner. *Computational Higher-Order Quasi-Monte Carlo for Random Partial Differential Equations*. PhD thesis, ETH Zürich, 2017. https://doi.org/10.3929/ethz-b-000182695


L. Herrmann, M. Keller, and Ch. Schwab. Quasi-Monte Carlo Bayesian estimation under Besov priors in elliptic inverse problems. *Math. Comp.* **90**:1831–1860, 2021. https://www.ams.org/journals/mcom/2021-90-330/S0025-5718-2021-03615-7

1. Preliminaries

Elliptic PDE

Many physical phenomena can be modeled using elliptic partial differential equations of the form

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

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 $-\nabla \cdot (a(\boldsymbol{x}) \nabla u(\boldsymbol{x})) = f(\boldsymbol{x})$ for a.e. $\boldsymbol{x} \in D$ with $u|_{\partial D} = 0$.

Let $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be a nonempty open set.

$L^2(D) := \{v \colon D \to \mathbb{R} \mid v \text{ is measurable, } \|v\|_{L^2(D)} := \big(\int_D |v(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x}\big)^{1/2} < \infty\}$,

$H^1(D) := \{v \in L^2(D) \mid \partial_j v \in L^2(D) \text{ for all } j \in \{1, \dots, d\}\}$,

with $\|v\|_{H^1(D)} := (\|v\|^2_{L^2(D)} + \|\nabla v\|^2_{L^2(D)})^{1/2}$,

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

where $\operatorname{supp}(v) := \overline{\{\boldsymbol{x} \in D \mid v(\boldsymbol{x}) \neq 0\}}$,

$H_0^1(D) := \operatorname{cl}_{H^1(D)}(C_0^\infty(D))$,

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

The spaces $L^2(D)$, $H^1(D)$, $H_0^1(D)$, and $H^{-1}(D)$ are Hilbert spaces.

The duality pairing $\langle f, v\rangle_{H^{-1}(D), H_0^1(D)}$ represents the bounded, linear functional $v \mapsto f(v)$ for $f \in H^{-1}(D)$ and $v \in H_0^1(D)$.

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

Lipschitz domain: A nonempty domain $D \subset \mathbb{R}^d$ is said to be a Lipschitz domain if, for every $\boldsymbol{x} \in \partial D$, there exists a rigid transformation (i.e., a rotation plus a translation) $f : \mathbb{R}^d \to \mathbb{R}^d$, a radius $r > 0$, and a Lipschitz function $\xi : \mathbb{R}^{d-1} \to \mathbb{R}$ such that

$$f(D) \cap B(f(\boldsymbol{x}), r) = \{(y_1, \ldots, y_d) \in \mathbb{R}^d \mid y_d < \xi(y_1, \ldots, y_{d-1})\} \cap B(f(\boldsymbol{x}), r),$$

where $B(\boldsymbol{x}, r)$ denotes the open ball of radius r centered at $\boldsymbol{x}$.

Trace theorem: Let D be a bounded Lipschitz domain. Then the trace operator

$$\gamma : C^\infty(\overline{D}) \to C^\infty(\partial D), \ \gamma u = u|_{\partial D},$$

has a unique extension to a bounded linear operator $\gamma : H^1(D) \to L^2(\partial D)$.

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 $H^1_0(D)$:

$$u \in H^1_0(D) \quad \Leftrightarrow \quad \gamma u = 0 : \partial D \to \mathbb{R}.$$

Q: *How to solve such PDE problems in practice?*

A: We consider the *weak formulation* of the PDE problem: given $f \in H^{-1}(D)$, find $u \in H_0^1(D)$ such that

$$\underbrace{\int_D a(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}_{=:B(u,v)} = \langle f, v \rangle_{H^{-1}(D), H_0^1(D)} \quad \text{for all } v \in H_0^1(D). \quad (1)$$

If there exist $a_{\min}, a_{\max} > 0$ s.t. $0 < a_{\min} \leq a(\boldsymbol{x}) \leq a_{\max} < \infty$ for all $\boldsymbol{x} \in D$, then the bilinear form $B \colon H_0^1(D) \times H_0^1(D) \to \mathbb{R}$ is bounded, i.e.,

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

for all $u, v \in H_0^1(D)$, and coercive, i.e.,

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

then 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

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

Let V_m be spanned by $\phi_1, \ldots, \phi_m$. We can write the solution as $u_m = \sum_{i=1}^m c_i \phi_i$. The above system reduces to the linear system of equations

$$\begin{bmatrix} \int_D a(\boldsymbol{x}) \nabla \phi_1(\boldsymbol{x}) \cdot \nabla \phi_1(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} & \cdots & \int_D a(\boldsymbol{x}) \nabla \phi_1(\boldsymbol{x}) \cdot \nabla \phi_m(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ \vdots & \ddots & \vdots \\ \int_D a(\boldsymbol{x}) \nabla \phi_m(\boldsymbol{x}) \cdot \nabla \phi_1(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} & \cdots & \int_D a(\boldsymbol{x}) \nabla \phi_m(\boldsymbol{x}) \cdot \nabla \phi_m(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \langle f, \phi_1 \rangle_{H^{-1}(D), H_0^1(D)} \\ \vdots \\ \langle f, \phi_m \rangle_{H^{-1}(D), H_0^1(D)} \end{bmatrix}.$$

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 $\phi_1, \ldots, \phi_m$ which are constructed to satisfy

$$\phi_i(\boldsymbol{n}_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

  where $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_m$ are the *interior* nodes of the triangulation.
- The finite element solution can be written as $u_m(\boldsymbol{x}) = \sum_{i=1}^m c_i \phi_i(\boldsymbol{x}) \in V_m$, where the expansion coefficients are solved from the Galerkin system. Note that $u_m(\boldsymbol{n}_j) = c_j$.
- If $v_m(\boldsymbol{x}) = \sum_{i=1}^m c_i \phi_i(\boldsymbol{x}) \in V_m$, then, e.g., $\|v_h\|_{L^2(D)} = \sqrt{\boldsymbol{c}^{\mathrm{T}} M \boldsymbol{c}}$, where $\boldsymbol{c} := [c_1, \ldots, c_m]^{\mathrm{T}}$ and $M = [M_{i,j}]_{i,j=1}^m$ is the mass matrix defined elementwise by $M_{i,j} := \int_D \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$, $i, j \in \{1, \ldots, m\}$.

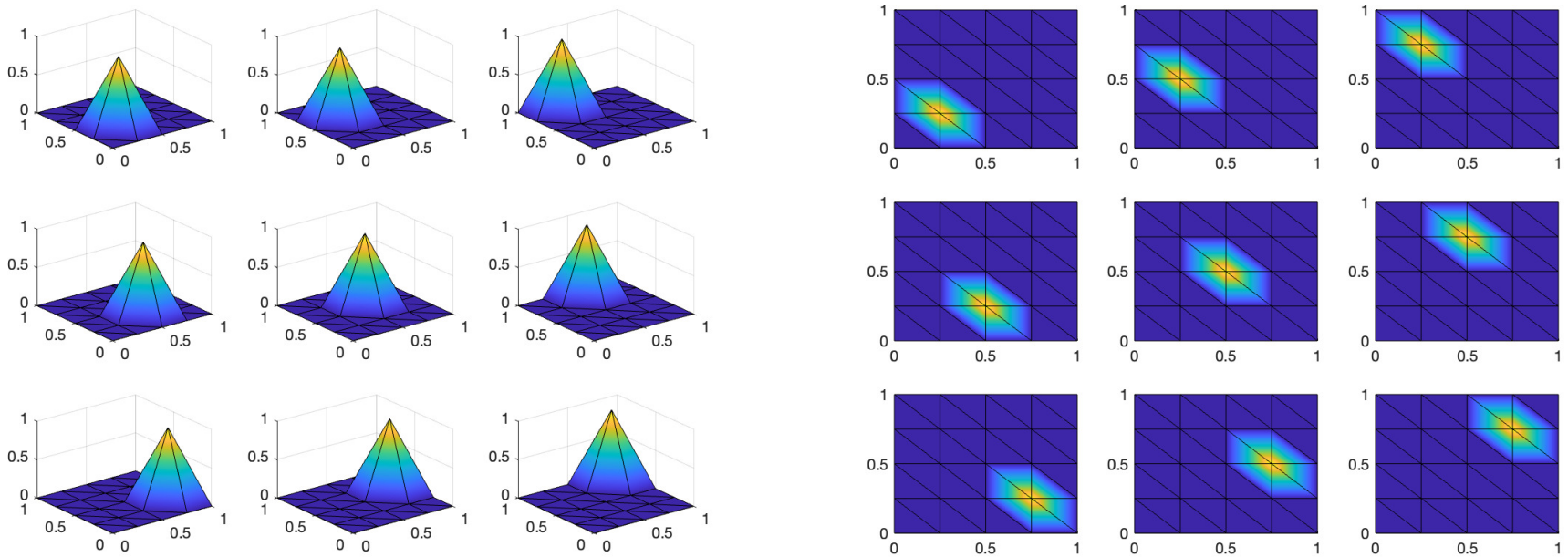


Figure: Left: An illustration of global, piecewise linear FE basis functions spanning V_m over a regular, uniform triangulation of $(0,1)^2$. Right: Bird's-eye view of the same global FE basis functions.

Random field

Definition

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

Definition

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

$$\int_\Omega |A(\boldsymbol{x}, \omega)|^2 \, \mu(\mathrm{d}\omega) < \infty \quad \text{for all } \boldsymbol{x} \in D.$$

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

Mercer's theorem

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

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

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

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

Mercer's theorem: the covariance operator $\mathcal{C}\colon L^2(D) \to L^2(D)$,

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

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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\lambda_k \to 0$ and the eigenfunctions form an orthonormal basis for $L^2(D)$.
Note that this means:

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

[†] In this context, positive definite means: for all choices of finitely many points $x_1, \ldots, x_k \in D$, $k \in \mathbb{N}$, 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 $a\colon D \times \Omega \to \mathbb{R}$ be a square-integrable random field with continuous, symmetric, positive definite covariance $K(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[(a(\boldsymbol{x}, \cdot) - \bar{a}(\boldsymbol{x}))(a(\boldsymbol{x}', \cdot) - \bar{a}(\boldsymbol{x}'))]$. Then the eigensystem $(\lambda_k, \psi_k) \in \mathbb{R}_+ \times L^2(D)$ of the covariance operator $\mathcal{C}\colon L^2(D) \to L^2(D)$, as described on the previous slide, can be used to write

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

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

where the convergence is in L^2 w.r.t. the stochastic parameter and uniform in $\boldsymbol{x}$. 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 $a(\boldsymbol{x}, \omega)$ can be written as

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

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

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

- The random variables ξ_k are centered and uncorrelated, but not necessarily independent.
- If the random field $a(\boldsymbol{x}, \omega)$ is Gaussian – by definition, this means that $(a(\boldsymbol{x}_1, \omega), \ldots, a(\boldsymbol{x}_k, \omega))$ is a multivariate Gaussian random variable for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in D$, $k \in \mathbb{N}$ – 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 the 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 $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be a Lipschitz domain and consider the PDE problem

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

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

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

where $a_0 \in L^\infty(D)$ is such that $a_0(\boldsymbol{x}) > 0$ 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 $a(\boldsymbol{x}, \boldsymbol{y}) \equiv a(\boldsymbol{x}, \boldsymbol{y}(\omega))$ and $u(\boldsymbol{x}, \boldsymbol{y}) \equiv u(\boldsymbol{x}, \boldsymbol{y}(\omega))$, where (formally) $\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}}$ is a *parametric vector* endowed with a product Gaussian measure.

Example (Uniform and affine input random field)

Let $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be a Lipschitz domain, $f : D \to \mathbb{R}$ is a fixed (deterministic) source term, and consider the PDE problem

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

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

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

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 $a(\boldsymbol{x}, \boldsymbol{y}) \equiv a(\boldsymbol{x}, \boldsymbol{y}(\omega))$ and $u(\boldsymbol{x}, \boldsymbol{y}) \equiv u(\boldsymbol{x}, \boldsymbol{y}(\omega))$, where $\boldsymbol{y} \in [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ 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 (formally) given by

$$\mathbb{E}[u(\boldsymbol{x},\cdot)] = \lim_{s\to\infty} \int_{\mathbb{R}^s} u(\boldsymbol{x},\boldsymbol{y}) \prod_{j=1}^{s} \frac{\mathrm{e}^{-\frac{1}{2}y_j^2}}{\sqrt{2\pi}} \,\mathrm{d}\boldsymbol{y}$$

while in the *affine and uniform model* the expected value of the PDE solution is (formally) given by

$$\mathbb{E}[u(\boldsymbol{x},\cdot)] = \lim_{s\to\infty} \int_{[-1/2,1/2]^s} u(\boldsymbol{x},\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y}.$$

- 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*.

2. Quasi-Monte Carlo (QMC) methods

Lattice rules

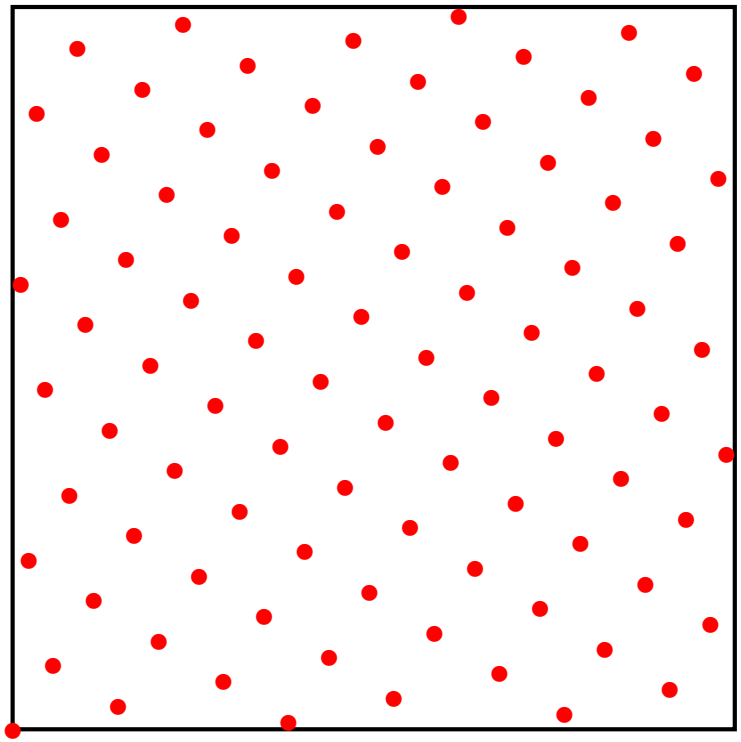
Rank-1 lattice rules

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

have the points

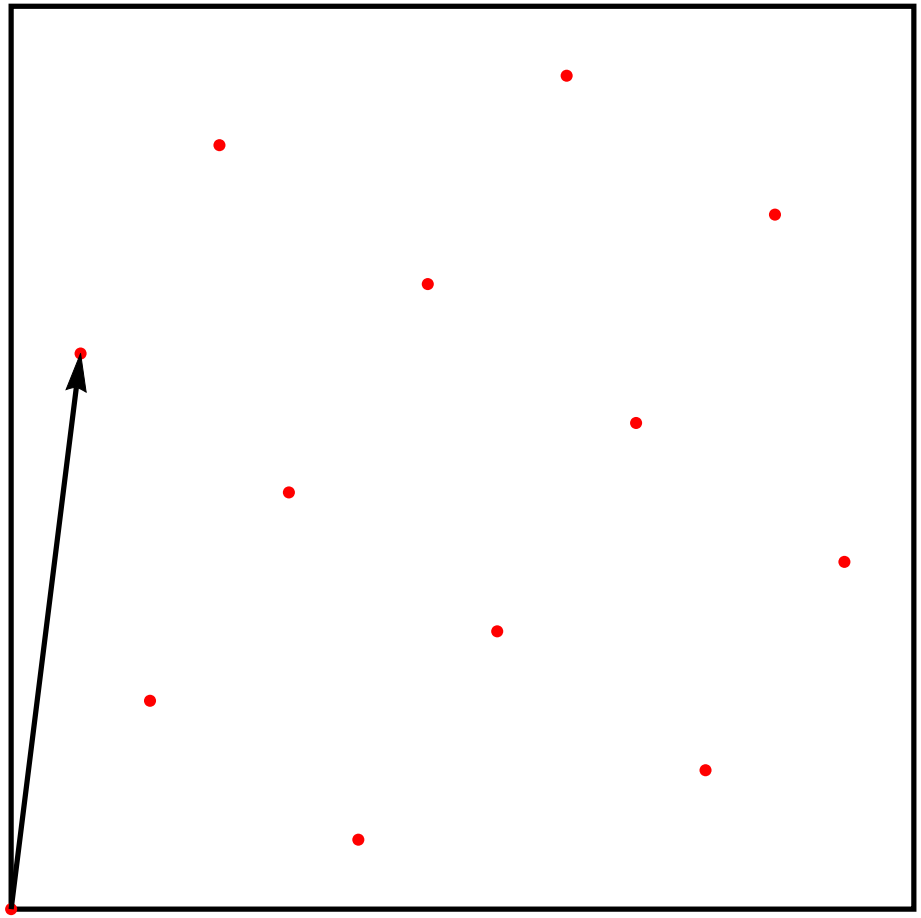
$$\boldsymbol{t}_i = \left\{\frac{i\boldsymbol{z}}{n}\right\} = \mathrm{mod}\left(\frac{i\boldsymbol{z}}{n}, 1\right), \quad i \in \{1, \ldots, n\},$$

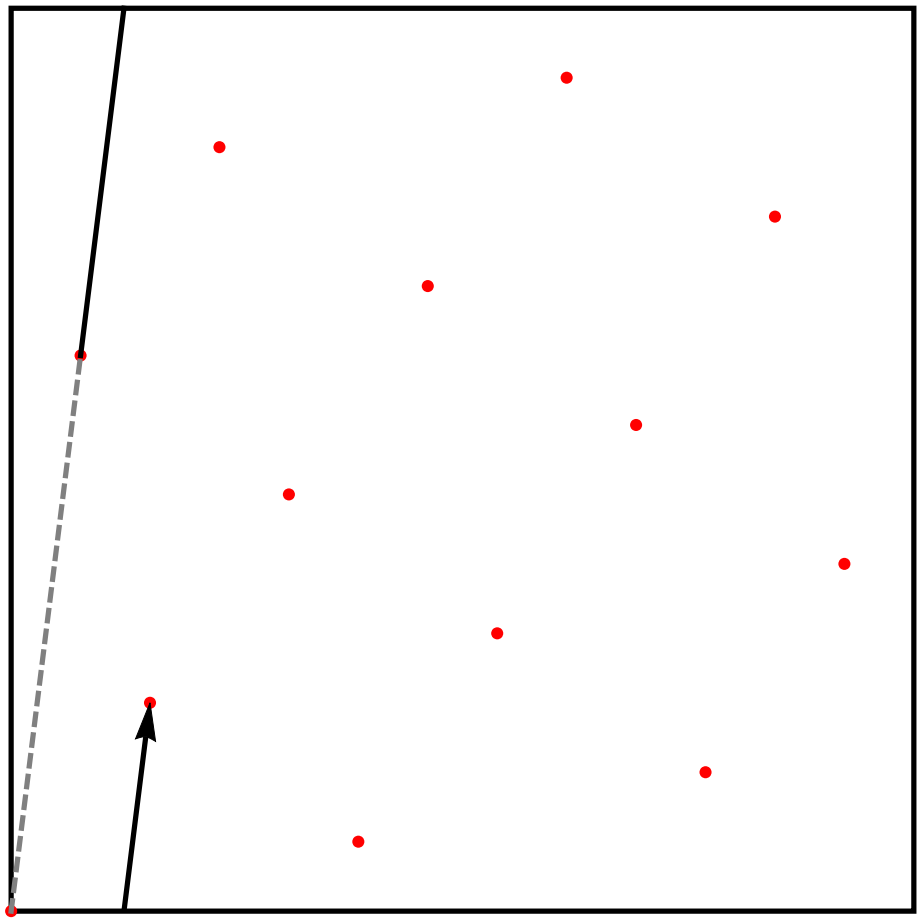
where the entire point set is determined by the *generating vector* $\boldsymbol{z} \in \mathbb{N}^s$, with all components *coprime* to n. The braces $\{\cdot\}$ denote the componentwise fractional part of a vector.

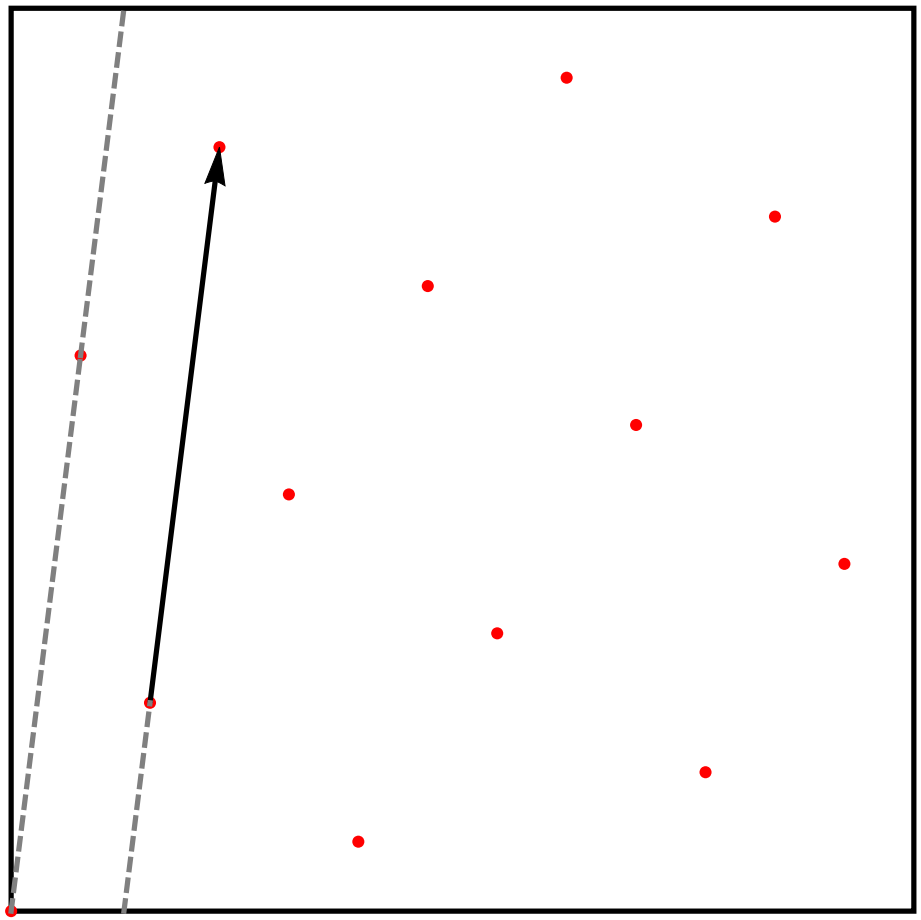


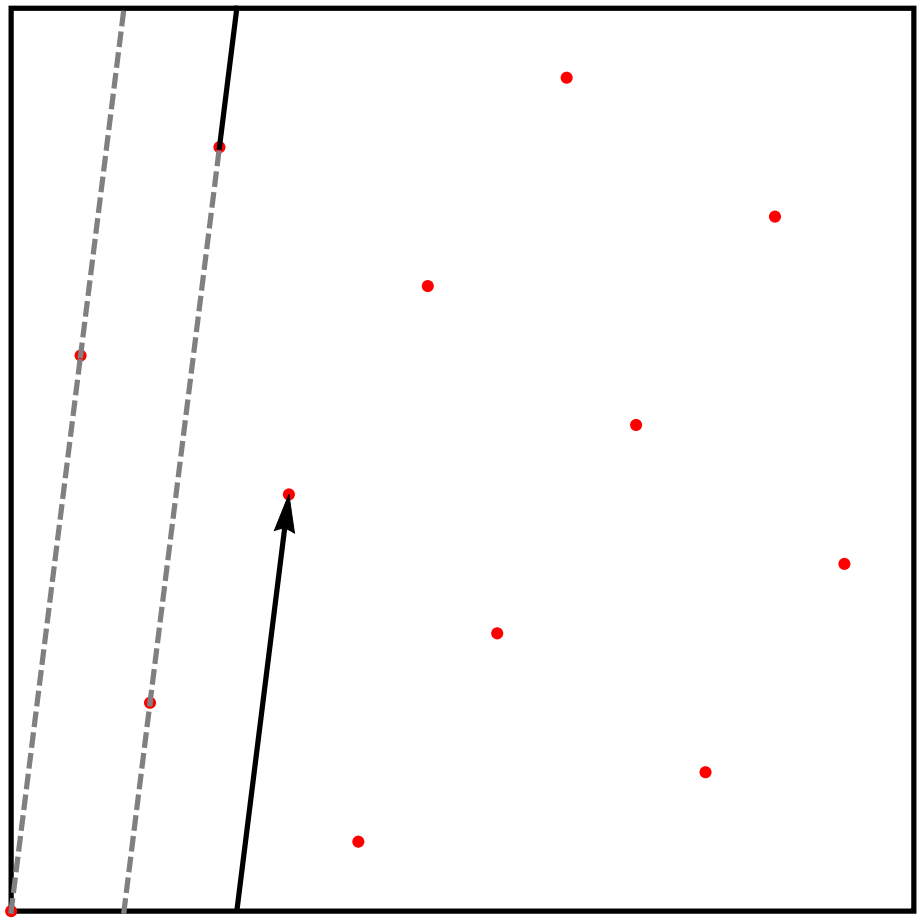
Lattice rule with $\boldsymbol{z} = (1, 55)$ and $n = 89$ nodes in $[0, 1]^2$

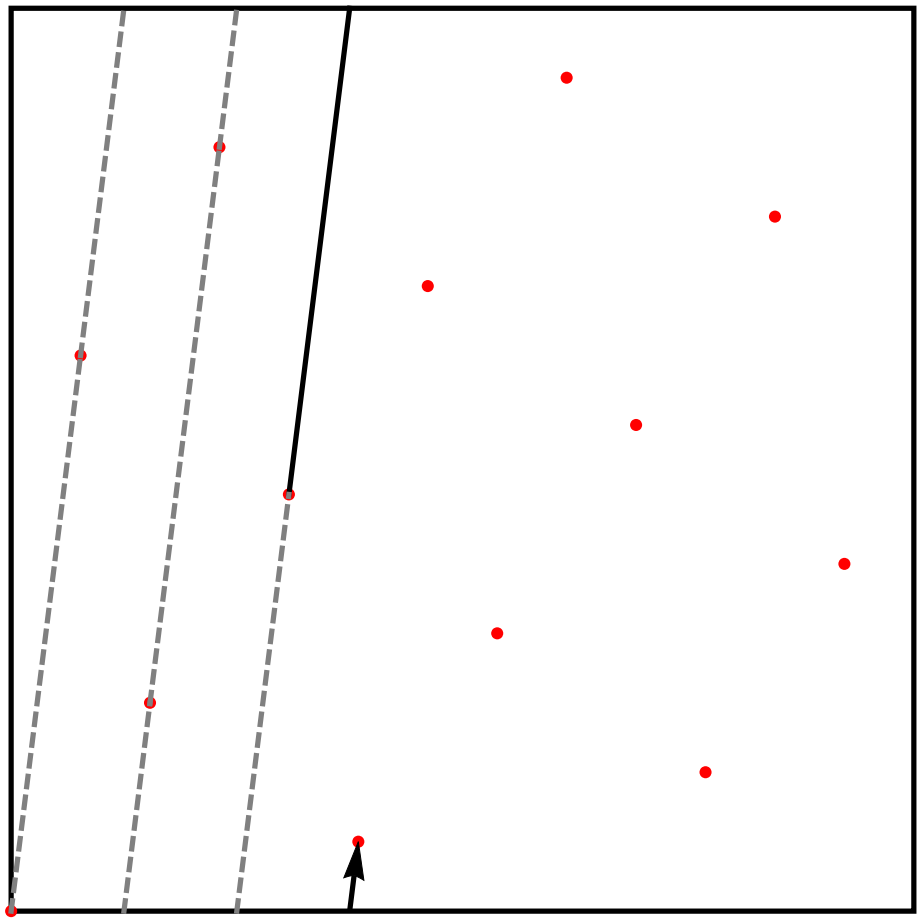
The quality of the lattice rule is determined by the choice of $\boldsymbol{z}$.

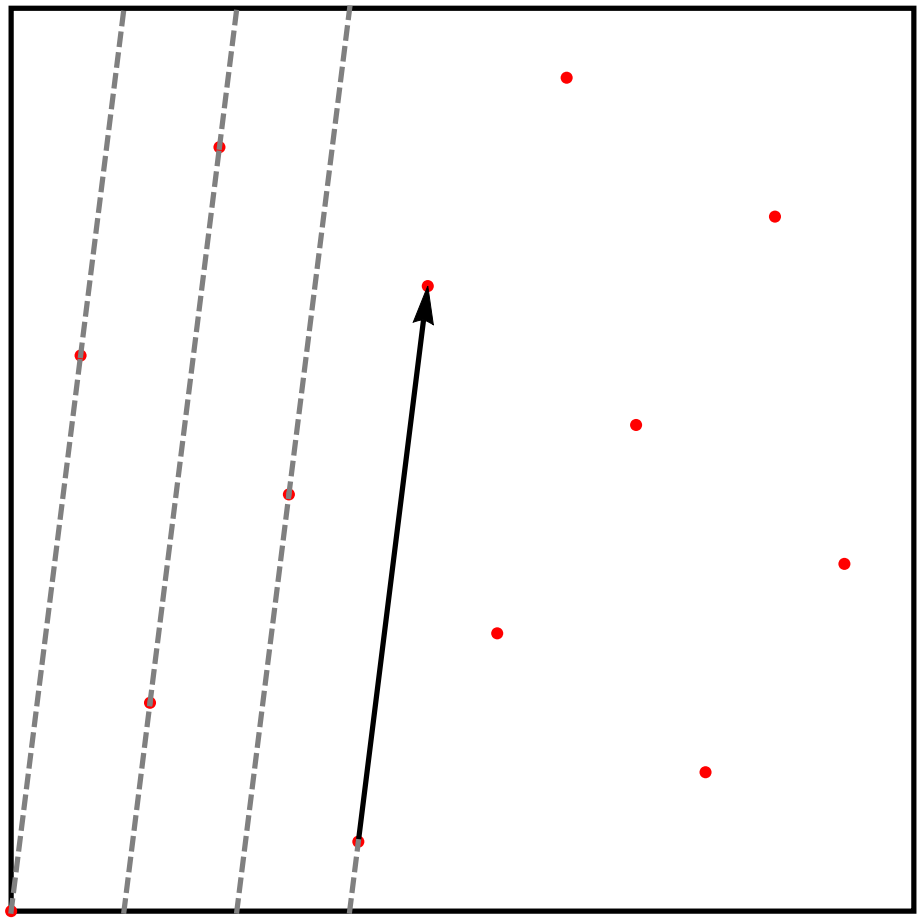


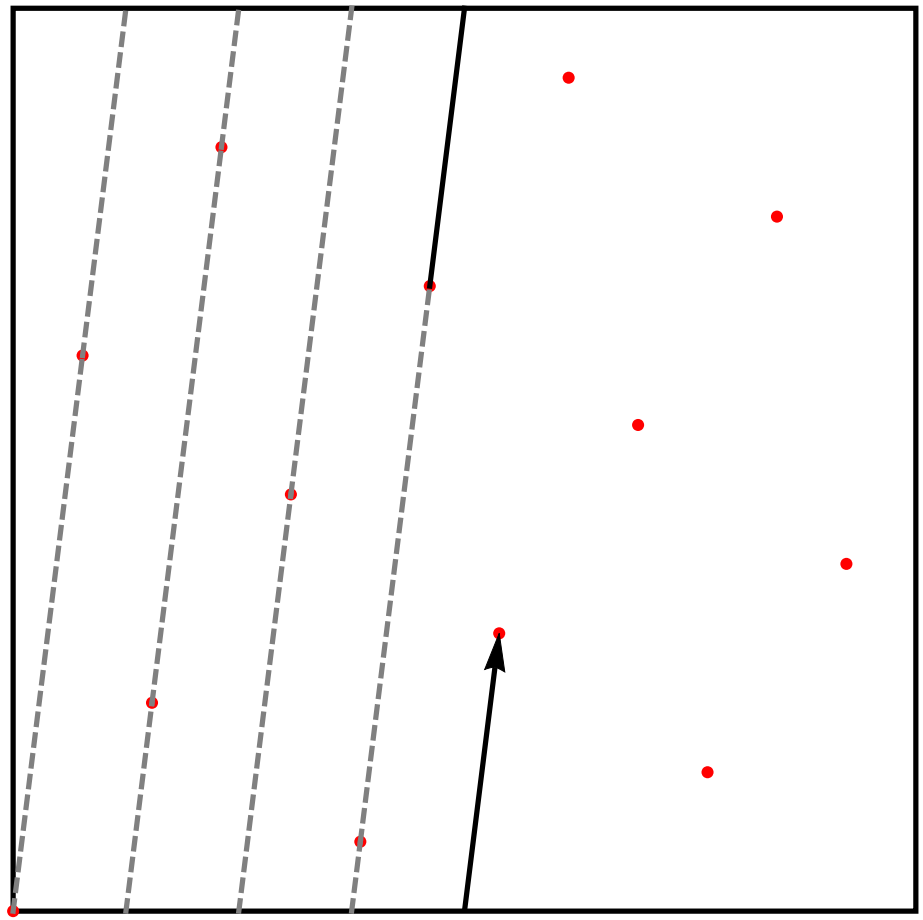


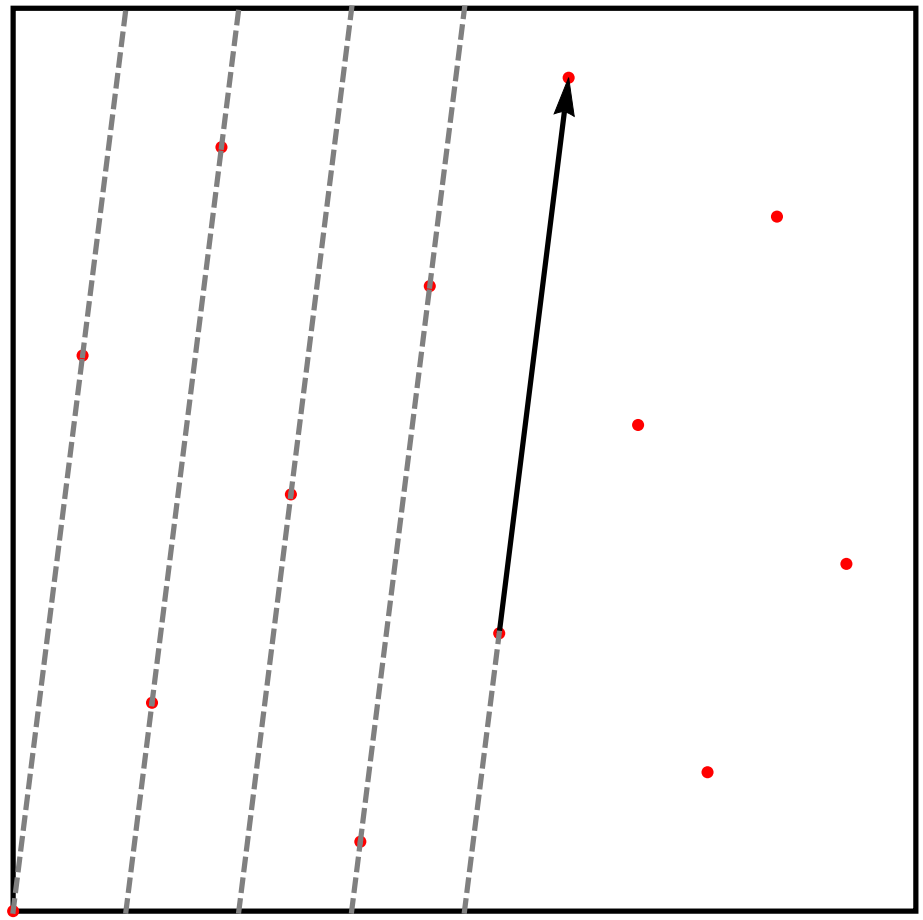


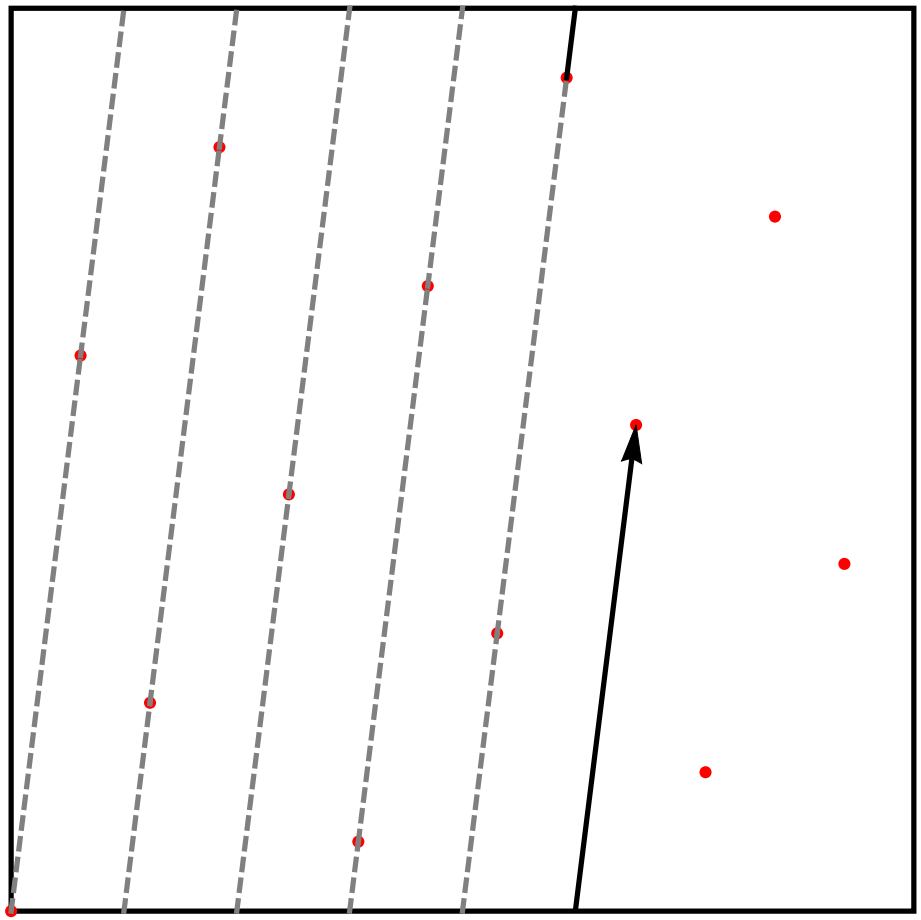


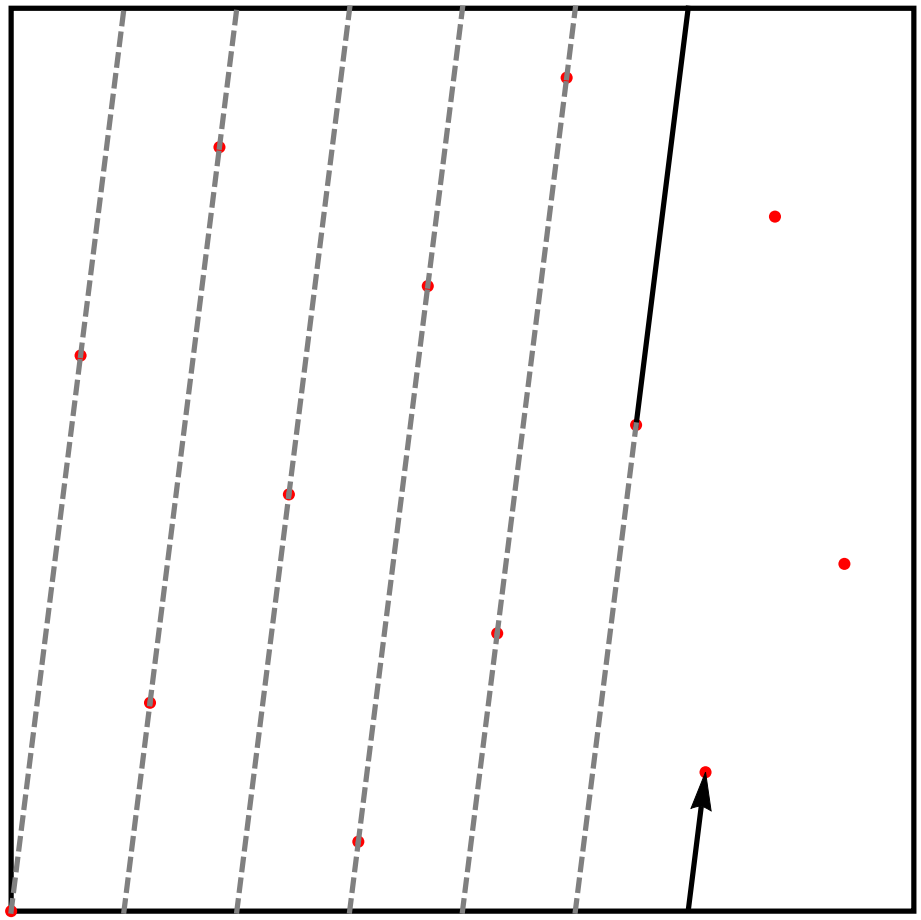


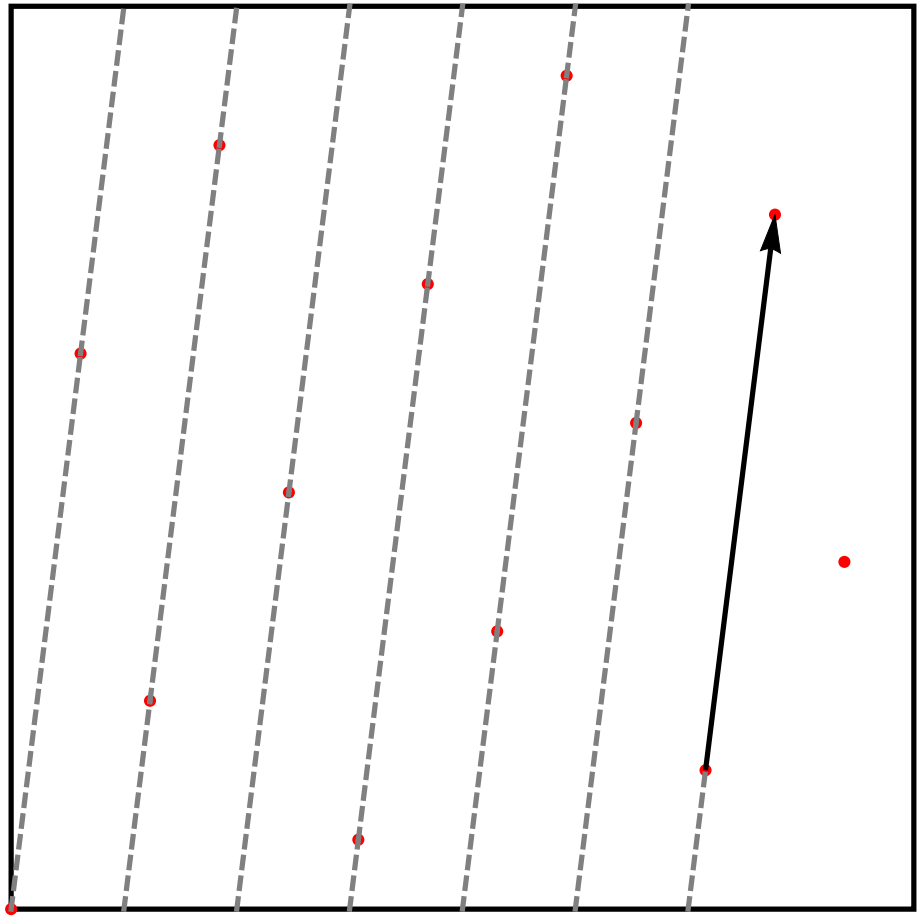


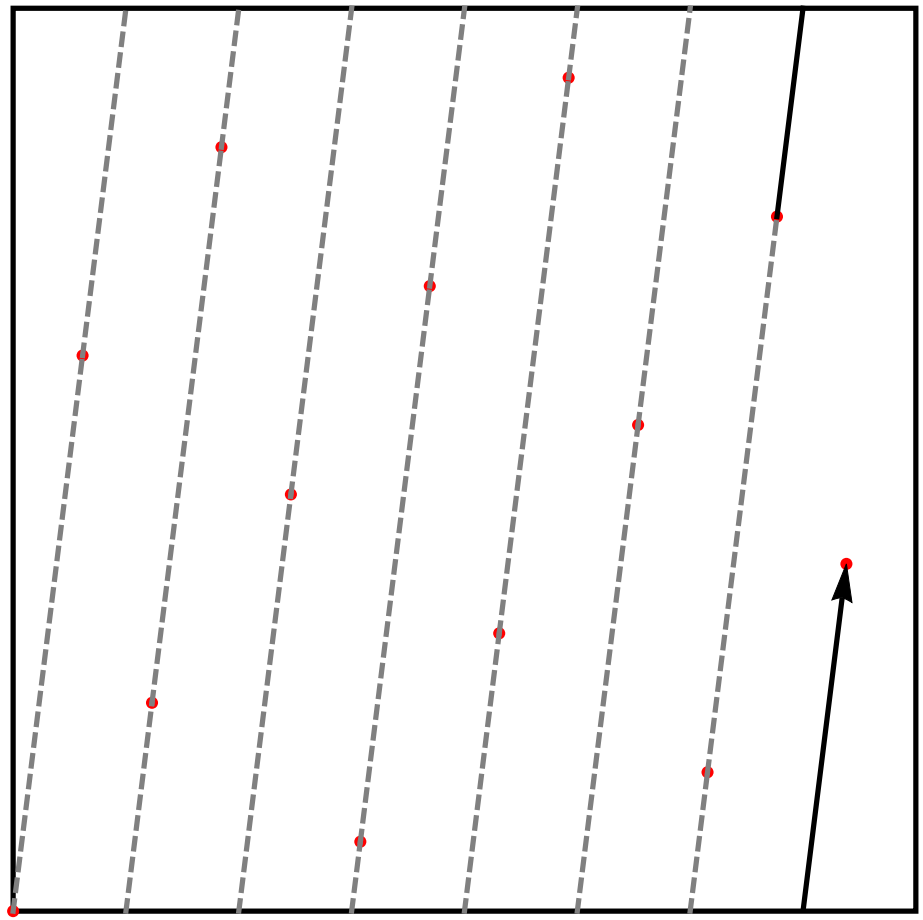












Worst-case error

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 $I_s : H \to \mathbb{R}$ be an integral operator

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

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

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

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

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

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

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 $\boldsymbol{y} \in D$, let

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

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

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

The function $K(\boldsymbol{x}, \boldsymbol{y})$ 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 $K : D \times D \to \mathbb{R}$ which satisfies

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

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 $A\colon H \to \mathbb{R}$, its representer $a \in H$ satisfying $A(f) = \langle f, a\rangle$ for all $f \in H$ is given by

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

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

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

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

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

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

$$\langle f, g\rangle = \left(\int_0^1 f(x)\,\mathrm{d}x\right)\left(\int_0^1 g(x)\,\mathrm{d}x\right) + \int_0^1 f'(x)g'(x)\,\mathrm{d}x.$$

Then this space has the *reproducing kernel*

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

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

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

there holds

$$\langle f, K(\cdot,y)\rangle = \Big(\int_0^1 f(x)\,\mathrm{d}x\Big)\underbrace{\Big(\int_0^1 K(x,y)\,\mathrm{d}x\Big)}_{=1} + \int_0^1 f'(x)\Big(x - \frac{1}{2} - \frac{1}{2}\mathrm{sign}(x-y)\Big)\,\mathrm{d}x$$

$$= \int_0^1 f(x)\,\mathrm{d}x + \int_0^1 f'(x)x\,\mathrm{d}x - \frac{1}{2}\int_0^1 f'(x)\,\mathrm{d}x + \frac{1}{2}\int_0^y f'(x)\,\mathrm{d}x - \frac{1}{2}\int_y^1 f'(x)\,\mathrm{d}x$$

$$= \cancel{\int_0^1 f(x)\,\mathrm{d}x} + \cancel{f(1)} - \cancel{\int_0^1 f(x)\,\mathrm{d}x} - \cancel{\frac{1}{2}f(1)} + \cancel{\frac{1}{2}f(0)} + \frac{1}{2}f(y) - \cancel{\frac{1}{2}f(0)} - \cancel{\frac{1}{2}f(1)} + \frac{1}{2}f(y)$$

$$= f(y)$$

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

Theorem

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

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

Then

$$e^2_{n,s}(P; H_s(K)) = \int_{[0,1]^s} \int_{[0,1]^s} K(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} - \frac{2}{n} \sum_{i=0}^{n-1} \int_{[0,1]^s} K(\boldsymbol{t}_i, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K(\boldsymbol{t}_i, \boldsymbol{t}_j). \tag{2}$$

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

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

Similarly, we find that

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

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

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

which will hold for all the kernels we shall consider.

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

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

where

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

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 $f = \xi / \|\xi\|_H \in H$ and we obtain

$$e_{n,s}^2(P; H) = \left\| \int_{[0,1]^s} K(\cdot, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \frac{1}{n} \sum_{i=0}^{n-1} K(\boldsymbol{x}, \boldsymbol{t}_i) \right\|^2$$

$$= \int_{[0,1]^s} \int_{[0,1]^s} K(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} - \frac{2}{n} \sum_{i=0}^{n-1} \int_{[0,1]^s} K(\boldsymbol{x}, \boldsymbol{t}_i) \, \mathrm{d}\boldsymbol{x} + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K(\boldsymbol{t}_i, \boldsymbol{t}_j),$$

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 $\boldsymbol{t}_k := \{\frac{k\boldsymbol{z}}{n}\}$ for some generating vector $z \in \mathbb{N}^s$ and fixed $n \in \mathbb{N}$. Given a vector $\boldsymbol{\Delta} \in [0,1]^s$, known as the *shift*, the $\boldsymbol{\Delta}$-shift of the QMC points $\boldsymbol{t}_0, \ldots, \boldsymbol{t}_{n-1}$ is defined as the point set

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

Shifting preserves the lattice structure. In practice, we will generate a number of independent random shifts $\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_{R-1}$ from $\mathcal{U}([0,1]^s)$ and take the average of $\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_{R-1}$-shifted QMC rules as our approximation of I_s.

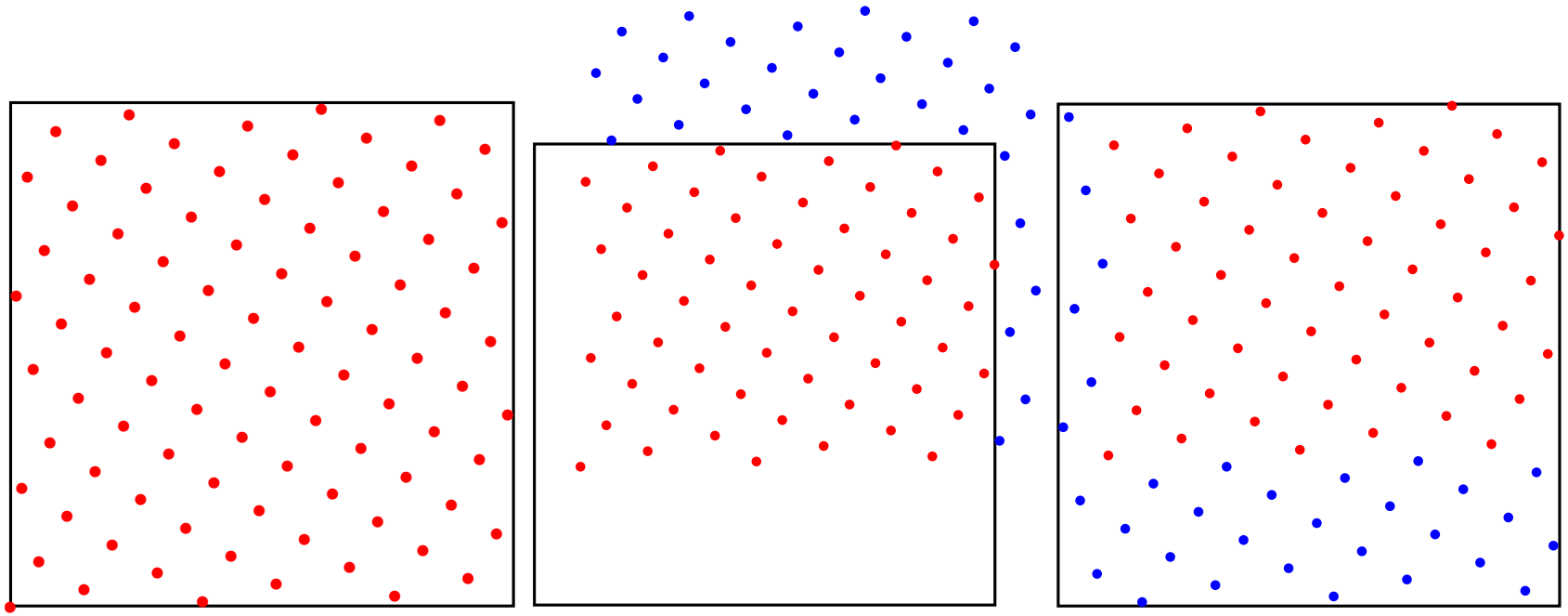
Advantages:

- 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

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

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 $\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_{R-1}$ from $\mathcal{U}([0,1]^s)$.
- For a given QMC rule with points $(\boldsymbol{t}_i)_{i=0}^{n-1} \subset [0,1]^s$, form the approximations $Q_{n,s}^{(0)} f, \ldots, Q_{n,s}^{(R-1)} f$, where

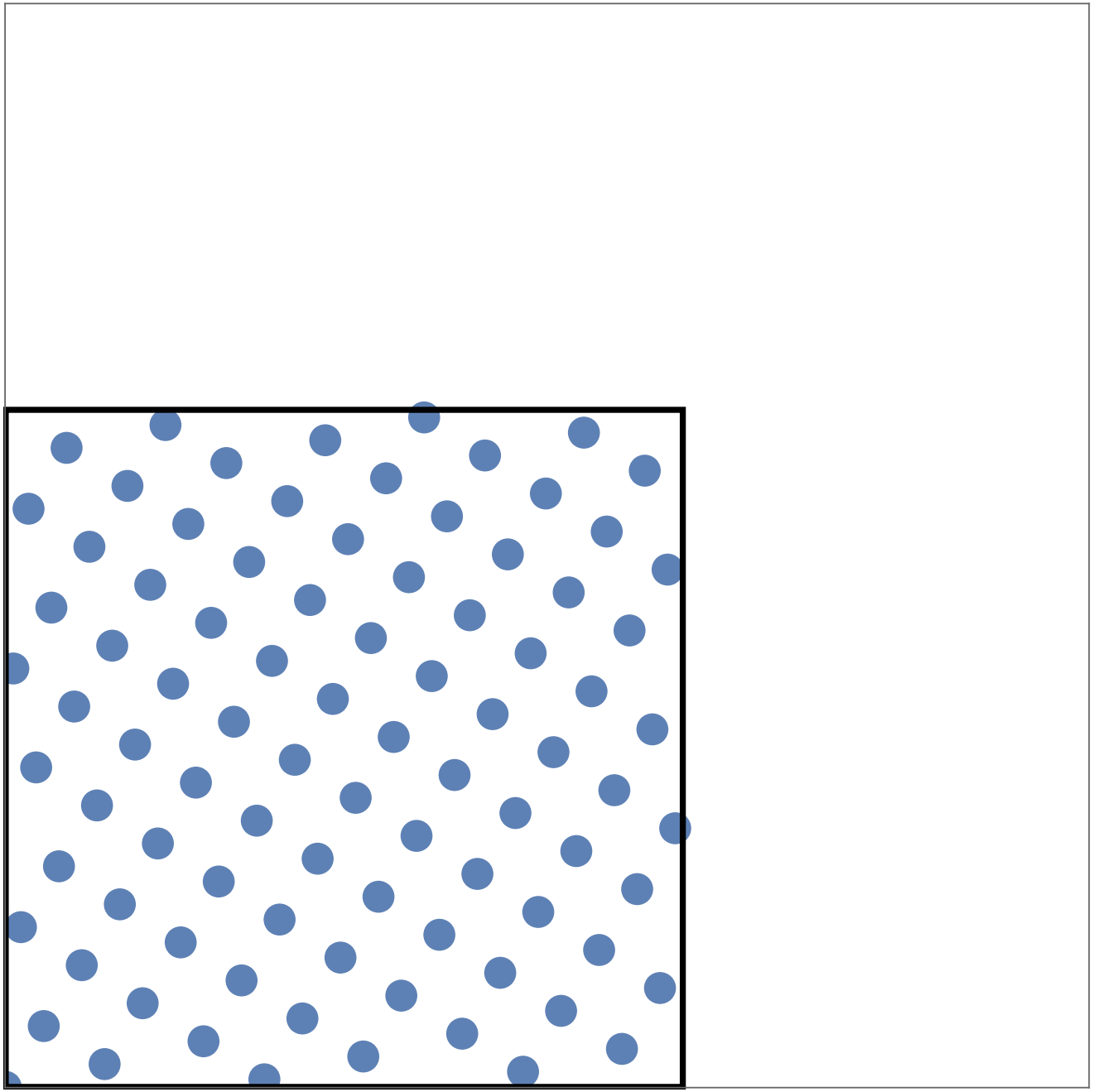
$$Q_{n,s}^{\boldsymbol{\Delta}_r} f = \frac{1}{n} \sum_{i=0}^{n-1} f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_r\}), \quad r = 0, \ldots, R-1,$$

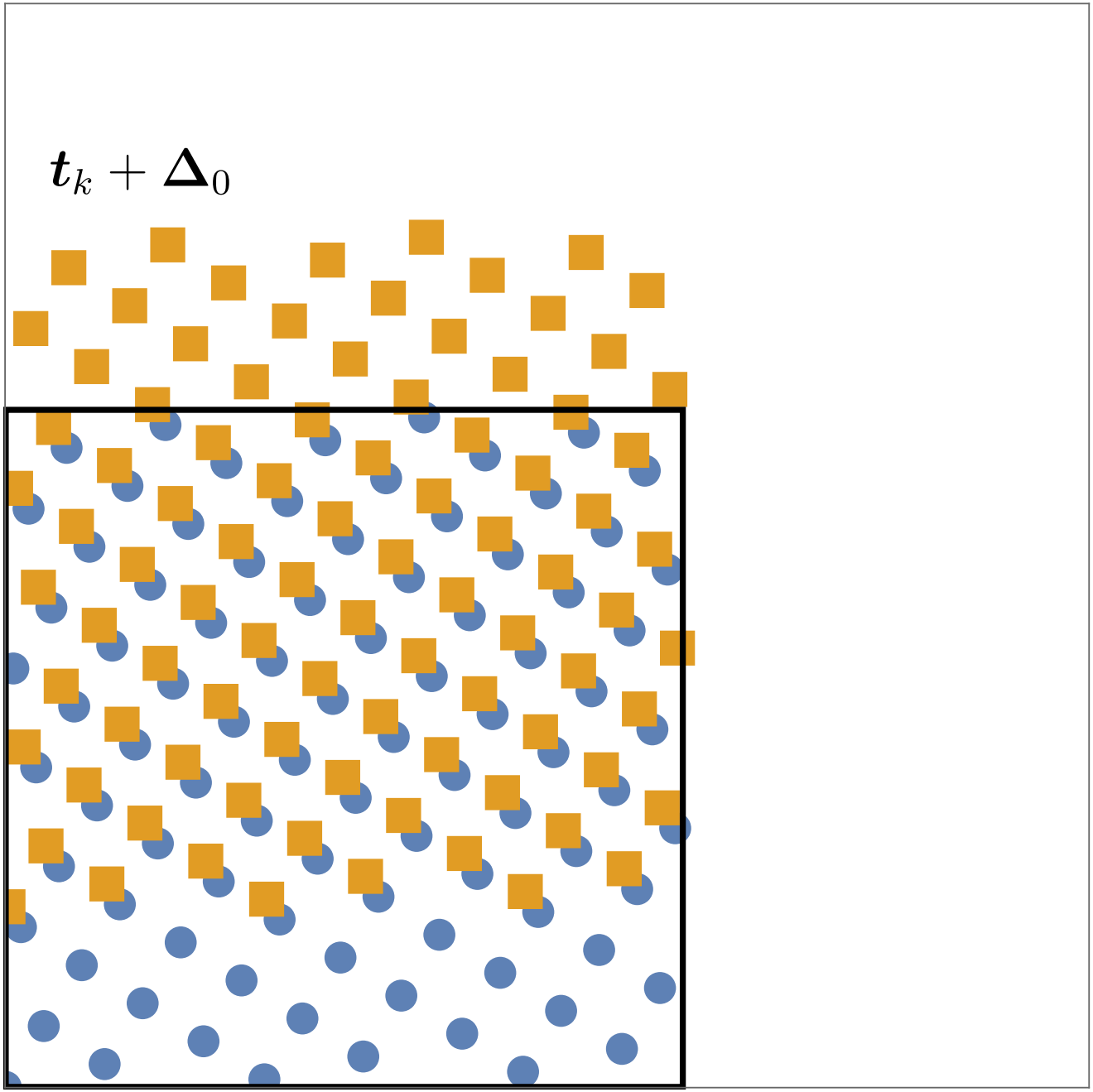
  is the approximation of the integral using a $\boldsymbol{\Delta}_r$-shift of the original QMC rule.
- We take the *average*

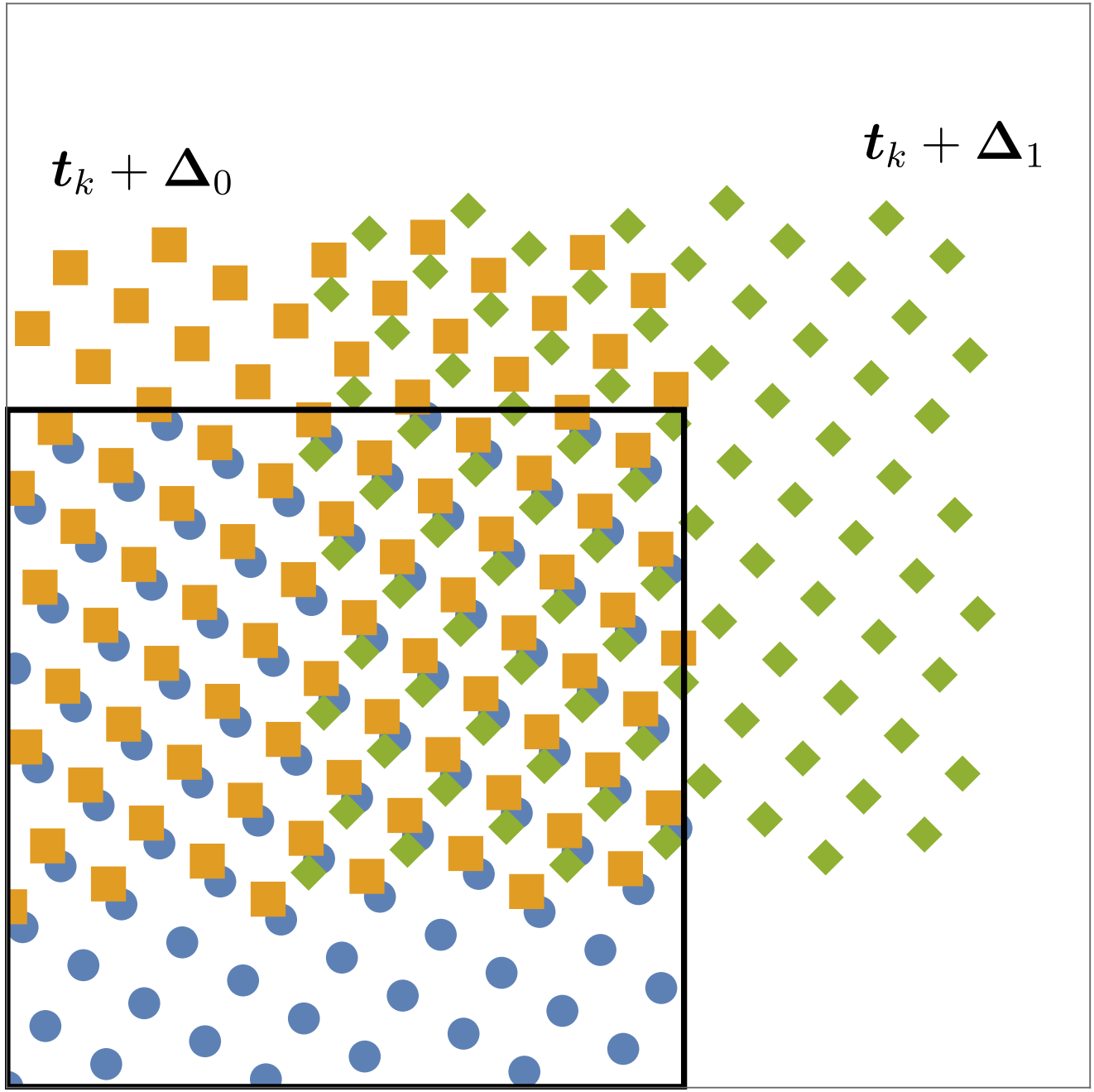
$$\overline{Q}_{n,s,R} f = \frac{1}{R} \sum_{r=0}^{R-1} Q_{n,s}^{\boldsymbol{\Delta}_r} f$$

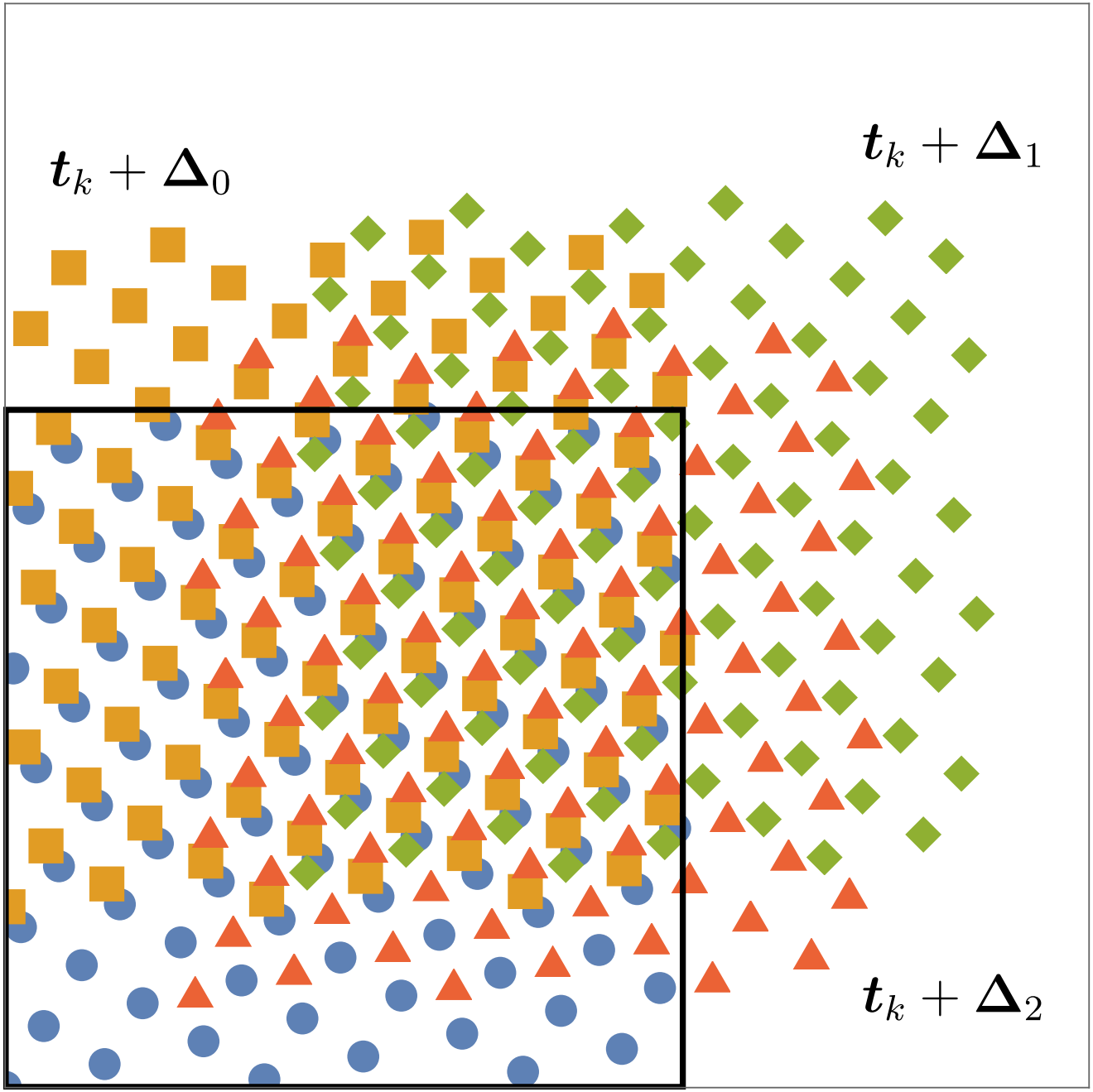
  as our *final* approximation of the integral.
- An *unbiased* estimate for the mean-square error of $\overline{Q}_{n,s,R} f$ is given by

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

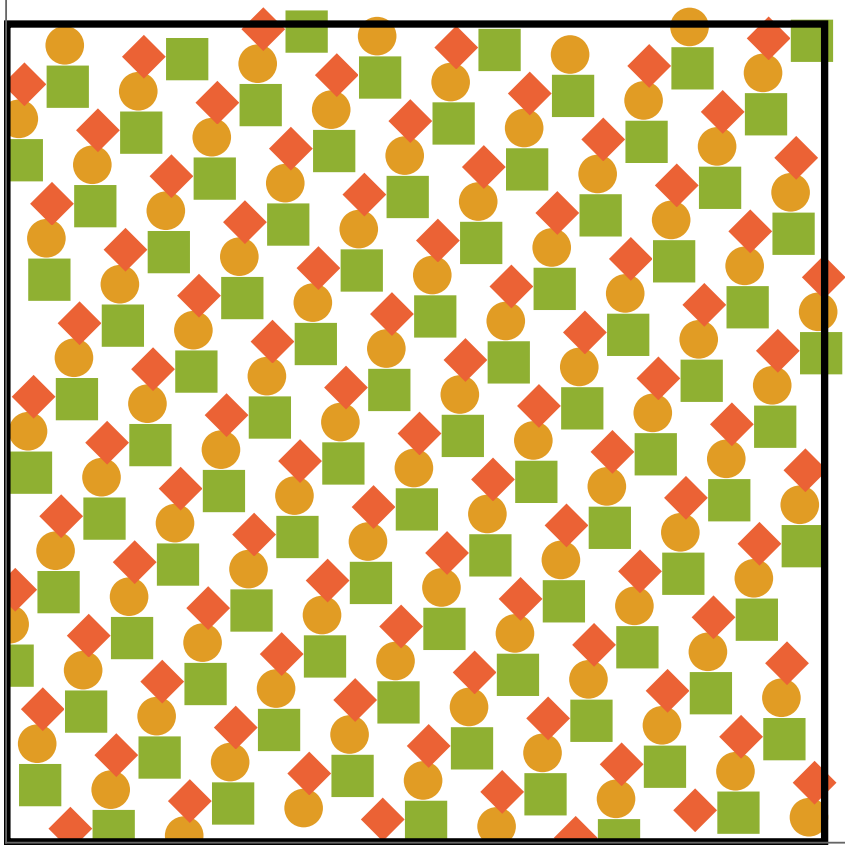



$t_k + \Delta_0$


$t_k + \Delta_0$
$t_k + \Delta_1$


$t_k + \Delta_0$
$t_k + \Delta_1$
$t_k + \Delta_2$

$$\{\{\boldsymbol{t}_k + \boldsymbol{\Delta}_0\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}_1\}, \{\boldsymbol{t}_k + \boldsymbol{\Delta}_2\}\}$$



$Q_{n,s}^{\boldsymbol{\Delta}_0} f = \frac{1}{n}\sum_{i=0}^{n-1} f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_0\})$, $Q_{n,s}^{\boldsymbol{\Delta}_1} f = \frac{1}{n}\sum_{i=0}^{n-1} f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_1\})$, $Q_{n,s}^{\boldsymbol{\Delta}_2} f = \frac{1}{n}\sum_{i=0}^{n-1} f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_2\})$

QMC approximation with 3 random shifts: $\overline{Q}_{n,s,3} f = \frac{Q_{n,s}^{\boldsymbol{\Delta}_0} f + Q_{n,s}^{\boldsymbol{\Delta}_1} f + Q_{n,s}^{\boldsymbol{\Delta}_2} f}{3}$.

Shift-averaged worst-case error

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

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

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

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

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

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

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

$$e_{n,s}^{\mathrm{sh}}(P; H) := \sqrt{\int_{[0,1]^s} e_{n,s}^2(P + \boldsymbol{\Delta}; H)\, \mathrm{d}\boldsymbol{\Delta}}$$

is called the *shift-averaged worst-case error*.

Theorem (Formula for the shift-averaged worst-case error)

$$[e^{\mathrm{sh}}_{n,s}(P; H_s(K))]^2 = -\int_{[0,1]^s}\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} + \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} K^{\mathrm{sh}}(\boldsymbol{t}_i,\boldsymbol{t}_j),$$

where

$$K^{\mathrm{sh}}(\boldsymbol{x},\boldsymbol{y}) := \int_{[0,1]^s} K(\{\boldsymbol{x}+\boldsymbol{\Delta}\},\{\boldsymbol{y}+\boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta}, \quad \boldsymbol{x},\boldsymbol{y}\in[0,1]^s.$$

Proof. The definition of shift-averaged WCE and (2) imply

$$\begin{aligned}
[e^{\mathrm{sh}}_{n,s}(P; H_s(K))]^2 &= \int_{[0,1]^s} e^2_{n,s}(P+\boldsymbol{\Delta}; H)\,\mathrm{d}\boldsymbol{\Delta} \\
&= \int_{[0,1]^s}\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - \frac{2}{n}\sum_{i=0}^{n-1}\int_{[0,1]^s}\int_{[0,1]^s} K(\{\boldsymbol{t}_i+\boldsymbol{\Delta}\},\boldsymbol{y})\,\mathrm{d}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{y} \\
&\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\int_{[0,1]^s} K(\{\boldsymbol{t}_i+\boldsymbol{\Delta}\},\{\boldsymbol{t}_j+\boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta}.
\end{aligned}$$

The result follows by a change of variables $\boldsymbol{x} = \{\boldsymbol{t}_i + \boldsymbol{\Delta}\}$ in the second term. □

Remarks

$$K^{\mathrm{sh}}(\boldsymbol{x},\boldsymbol{y}) := \int_{[0,1]^s} K(\{\boldsymbol{x}+\boldsymbol{\Delta}\},\{\boldsymbol{y}+\boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta}, \quad \boldsymbol{x},\boldsymbol{y}\in[0,1]^s.$$

- The function K^{sh} is actually a reproducing kernel, with the *shift-invariant property*

$$K^{\mathrm{sh}}(\boldsymbol{x},\boldsymbol{y}) = K^{\mathrm{sh}}(\{\boldsymbol{x}+\boldsymbol{\Delta}\},\{\boldsymbol{y}+\boldsymbol{\Delta}\}) \quad \text{for all } \boldsymbol{x},\boldsymbol{y},\boldsymbol{\Delta}\in[0,1].$$

  Equivalently,

$$K^{\mathrm{sh}}(\boldsymbol{x},\boldsymbol{y}) = K^{\mathrm{sh}}(\{\boldsymbol{x}-\boldsymbol{y}\},\boldsymbol{0}) \quad \text{for all } \boldsymbol{x},\boldsymbol{y}\in[0,1].$$

- 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 $(\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ 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

$$K_{s,\gamma}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}} \prod_{j\in\mathfrak{u}} \eta(x_j,y_j), \quad \eta(x,y) := \tfrac{1}{2}B_2(|x-y|)+(x-\tfrac{1}{2})(y-\tfrac{1}{2}),$$

where $B_2(x) := x^2 - x + \frac{1}{6}$ is the Bernoulli polynomial of degree 2 and we use the notation $\{1:s\} := \{1,\dots,s\}$.

The norm $\|f\|_{s,\gamma} = \sqrt{\langle f,f\rangle_{s,\gamma}}$ is induced by the inner product

$$\langle f,g\rangle_{s,\gamma} = \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[0,1]^{|\mathfrak{u}|}} \left(\int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}_{-\mathfrak{u}} \right) \times \left(\int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}_{-\mathfrak{u}} \right) \mathrm{d}\boldsymbol{x}_{\mathfrak{u}},$$

where $\mathrm{d}\boldsymbol{x}_{\mathfrak{u}} := \prod_{j\in\mathfrak{u}} \mathrm{d}x_j$ and $\mathrm{d}\boldsymbol{x}_{-\mathfrak{u}} := \prod_{j\in\{1:s\}\setminus\mathfrak{u}} \mathrm{d}x_j$.

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 $\boldsymbol{x}_{\mathfrak{u}} = \{x_j \mid j \in \mathfrak{u}\}$. We refer to these as the "active" variables, and denote the remaining "inactive" variables by $\boldsymbol{x}_{-\mathfrak{u}}$.
- The cardinality $|\mathfrak{u}|$ of the set $\mathfrak{u}$ is referred to as the "order" of the subset of variables $\boldsymbol{x}_{\mathfrak{u}}$. There is a *weight* parameter $\gamma_{\mathfrak{u}}$ associated with every subset of variables $\boldsymbol{x}_{\mathfrak{u}}$. 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 $\frac{\partial^{|\mathfrak{u}|} f}{\partial \boldsymbol{x}_{\mathfrak{u}}}$ must also be small.
- Note that $\|\cdot\|_{s,\boldsymbol{\gamma}}$ and $\|\cdot\|_{s,c\boldsymbol{\gamma}}$ are equivalent norms for any $c > 0$.[†] Therefore we do not lose any generality by assuming that the weights have been normalized s.t. $\gamma_{\varnothing} = 1$. WLOG, we will *always* use the convention that $\gamma_{\varnothing} := 1$.

[†]Here, $c\boldsymbol{\gamma} = (c\gamma_{\mathfrak{u}})_{\mathfrak{u} \subseteq \{1:s\}}$.

Special forms of weights

- *Product weights*: we have a sequence of numbers satisfying $\gamma_1 \geq \gamma_2 \geq \cdots$ and we take

$$\gamma_{\mathfrak{u}} = \prod_{j \in \mathfrak{u}} \gamma_j.$$

  In this case, the reproducing kernel is given by the product

$$K_{s,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s} \Big(1 + \gamma_j \eta(x_j, y_j)\Big).$$

- *Finite order weights*: there exists $q \in \mathbb{N}$ s.t. $\gamma_{\mathfrak{u}} = 0$ for all $|\mathfrak{u}| > q$.
- *Order dependent weights*: we have a sequence of numbers $\Gamma_1, \Gamma_2, \ldots$, and take

$$\gamma_{\mathfrak{u}} = \Gamma_{|\mathfrak{u}|}.$$

- *Product-and-order dependent (POD) weights*: we have two sequences $\gamma_1, \gamma_2, \ldots$ and $\Gamma_1, \Gamma_2, \ldots$, and take

$$\gamma_{\mathfrak{u}} = \Gamma_{|\mathfrak{u}|} \prod_{j \in \mathfrak{u}} \gamma_j.$$

Why weighted spaces are interesting

Theorem (Sloan and Woźniakowski 1998)

Consider $H_{s,\boldsymbol{\gamma}}$ equipped with product weights $\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j$. Then there exist point sets $P_n \subset [0,1]^s$ for $n = 1, 2, \ldots$ such that the worst-case error $e_{n,s}(P_n; H_{s,\boldsymbol{\gamma}})$ 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 $s \to \infty$.
- 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

$$K_{s,\gamma}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}} \prod_{j\in\mathfrak{u}} \eta(x_j,y_j), \quad \eta(x,y) := \tfrac{1}{2}B_2(|x-y|)+(x-\tfrac{1}{2})(y-\tfrac{1}{2}),$$

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

Lemma

$$\int_{[0,1]^s} K_{s,\gamma}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = 1,$$

$$\int_{[0,1]^s}\int_{[0,1]^s} K_{s,\gamma}(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} = 1,$$

$$\int_{[0,1]^s} K_{s,\gamma}(\boldsymbol{x},\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}}(\tfrac{1}{6})^{|\mathfrak{u}|}.$$

Recall that $H_{s,\gamma}$ is defined via the reproducing kernel

$$K_{s,\gamma}(\boldsymbol{x},\boldsymbol{y}) = \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}} \prod_{j\in\mathfrak{u}} \eta(x_j,y_j), \quad \eta(x,y) := \tfrac{1}{2}B_2(|x-y|)+(x-\tfrac{1}{2})(y-\tfrac{1}{2}),$$

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

$$\begin{aligned} K^{\mathrm{sh}}_{s,\gamma}(\boldsymbol{x},\boldsymbol{y}) &:= \int_{[0,1]^s} K_{s,\gamma}(\{\boldsymbol{x}+\boldsymbol{\Delta}\},\{\boldsymbol{y}+\boldsymbol{\Delta}\})\,\mathrm{d}\boldsymbol{\Delta} \\ &= \sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}} \prod_{j\in\mathfrak{u}} B_2(|x_j-y_j|). \end{aligned}$$

Proof. This is an immediate consequence of

$$\int_0^1 \eta(\{x+\Delta\},\{y+\Delta\})\,\mathrm{d}\Delta = B_2(|x-y|). \quad \square$$

Let

$$P = \left\{ \left\{ \frac{k\boldsymbol{z}}{n} \right\} \mid k = 0, \ldots, n-1 \right\}$$

be a rank-1 lattice point set corresponding to generating vector $\boldsymbol{z} \in \mathbb{N}^s$ and $n \in \mathbb{N}$.

When dealing with the shift-invariant kernel corresponding to the unanchored, weighted Sobolev space $H_{s,\boldsymbol{\gamma}}$, we use the shorthand notation

$$e^{\mathrm{sh}}_{n,s}(\boldsymbol{z}) := e^{\mathrm{sh}}_{n,s}(P; H_{s,\boldsymbol{\gamma}}).$$

Lemma

The shift-averaged worst-case error for a rank-1 lattice rule in the weighted unanchored Sobolev space satisfies

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

Proof. Let $\boldsymbol{t}_j = \left\{\frac{j\boldsymbol{z}}{n}\right\}$. We have the kernel

$$K_{s,\gamma}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} \eta(x_j, y_j), \quad \eta(x, y) := \tfrac{1}{2} B_2(|x-y|) + (x-\tfrac{1}{2})(y-\tfrac{1}{2}),$$

which satisfies $\int_{[0,1]^s} \int_{[0,1]^s} K_{s,\gamma}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} = 1$. We showed that the shift-invariant kernel related to K is given by

$$K^{\mathrm{sh}}_{s,\gamma}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{k \in \mathfrak{u}} B_2(|x_k - y_k|).$$

Moreover, we showed that the shift-averaged WCE is given by

$$[e^{\mathrm{sh}}_{n,s}(\boldsymbol{z})]^2 = -\int_{[0,1]^s} \int_{[0,1]^s} K_{s,\gamma}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} + \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} K^{\mathrm{sh}}_{s,\gamma}(\boldsymbol{t}_i, \boldsymbol{t}_j).$$

Making the obvious substitutions, we arrive at

$$[e^{\mathrm{sh}}_{n,s}(\boldsymbol{z})]^2 = -1 + \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{\mathfrak{u}\subseteq\{1:s\}}\gamma_{\mathfrak{u}}\prod_{k\in\mathfrak{u}}B_2\left(\left\{\frac{(i-j)z_k}{n}\right\}\right) \quad (\gamma_\varnothing := 1)$$

$$= \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s\}}\gamma_{\mathfrak{u}}\prod_{k\in\mathfrak{u}}B_2\left(\left\{\frac{\mathrm{mod}(i-j,n)z_k}{n}\right\}\right)$$

$$= \frac{1}{n^2}\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s\}}\gamma_{\mathfrak{u}}\underbrace{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\prod_{k\in\mathfrak{u}}B_2\left(\left\{\frac{\mathrm{mod}(i-j,n)z_k}{n}\right\}\right)}_{=n\sum_{\ell=0}^{n-1}\prod_{k\in\mathfrak{u}}B_2\left(\left\{\frac{\ell z_k}{n}\right\}\right)}.$$

Final step: as i and j range from 0 to $n-1$, the values of $\mathrm{mod}(i-j,n)$ are just $0,\ldots,n-1$ in a different order (see next slide for illustration), with each value occurring n times. Thus

$$[e^{\mathrm{sh}}_{n,s}(\boldsymbol{z})]^2 = \frac{1}{n}\sum_{\ell=0}^{n-1}\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s\}}\gamma_{\mathfrak{u}}\prod_{k\in\mathfrak{u}}B_2\left(\left\{\frac{\ell z_k}{n}\right\}\right),$$

as desired. □

An illustration of the counting argument used on the previous slide

i/j	0	1	2	3	4	$\cdots$	$n-1$
0	$f(0)$	$f(1)$	$f(2)$	$f(3)$	$f(4)$	$\cdots$	$f(n-1)$
1	$f(n-1)$	$f(0)$	$f(1)$	$f(2)$	$f(3)$	$\cdots$	$f(n-2)$
2	$f(n-2)$	$f(n-1)$	$f(0)$	$f(1)$	$f(2)$	$\cdots$	$f(n-3)$
3	$f(n-3)$	$f(n-2)$	$f(n-1)$	$f(0)$	$f(1)$	$\cdots$	$f(n-4)$
4	$f(n-4)$	$f(n-3)$	$f(n-2)$	$f(n-1)$	$f(0)$	$\cdots$	$f(n-5)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$n-1$	$f(1)$	$f(2)$	$f(3)$	$f(4)$	$f(5)$	$\cdots$	$f(0)$

Table of the values $f(\mathrm{mod}(i-j,n))$, when $i,j \in \{0,1,\ldots,n-1\}$.

By a simple counting argument we can write

$$\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} f(\mathrm{mod}(i-j,n)) = n\sum_{\ell=0}^{n-1} f(\ell)$$

for *any* function $f : \{0,1,\ldots,n-1\} \to \mathbb{R}$.

Component-by-component construction

The components of the generating vector $\boldsymbol{z}$ can be restricted to the set

$$\mathbb{U}_n := \{z \in \mathbb{Z} \mid 1 \leq z \leq n-1 \text{ and } \gcd(z,n) = 1\},$$

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

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} |I_s f - Q_{n,s}^{\boldsymbol{\Delta}} f|^2} \leq e_{n,s}^{\mathrm{sh}}(\boldsymbol{z}) \|f\|_{s,\gamma}.$$

Finding $\boldsymbol{z}^* = \arg\min_{\boldsymbol{z} \in \mathbb{U}_n} e_{n,s}^{\mathrm{sh}}(\boldsymbol{z})$ is not computationally feasible: the search space contains altogether up to $(n-1)^s$ possible choices for $\boldsymbol{z}$. 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 $(\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$.

1. Set $z_1 = 1$.
2. For $k = 2, 3, \ldots, s$, choose $z_k \in \mathbb{U}_n$ to minimize $[e^{\mathrm{sh}}_{n,k}(z_1, \ldots, z_k)]^2$.

Remarks:

- Note that we have the (in principle computable) expression

$$[e^{\mathrm{sh}}_{n,k}(\boldsymbol{z})]^2 = \frac{1}{n} \sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:k\}} \gamma_{\mathfrak{u}} \sum_{\ell=0}^{n-1} \prod_{j\in\mathfrak{u}} B_2\left(\left\{\frac{\ell z_j}{n}\right\}\right). \tag{6}$$

- We will show that when the weights $(\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ are so-called *product-and-order dependent* (*POD*) weights, i.e., they can be written in the form

$$\gamma_{\mathfrak{u}} := \Gamma_{|\mathfrak{u}|} \prod_{j\in\mathfrak{u}} \gamma_j, \quad \mathfrak{u} \subseteq \{1:s\},$$

  where $\gamma_{\varnothing} := 1$, $(\Gamma_k)_{k=1}^{\infty}$ 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 $e^{\mathrm{sh}}_{n,s}(\boldsymbol{z})$. Regardless, we **can** produce an error estimate for the *QMC rule based on a generating vector constructed by the CBC algorithm!*

Theorem (CBC error bound)

The generating vector $\boldsymbol{z} \in \mathbb{U}_n^s$ constructed by the CBC algorithm, minimizing the squared shift-averaged worst-case error $[e_{n,s}^{\mathrm{sh}}(\boldsymbol{z})]^2$ for the weighted unanchored Sobolev space in each step, satisfies

$$[e_{n,s}^{\mathrm{sh}}(\boldsymbol{z})]^2 \le \left(\frac{1}{\varphi(n)} \sum_{\varnothing \ne \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}}\right)^{|\mathfrak{u}|}\right)^{1/\lambda} \quad \text{for all } \lambda \in (1/2, 1],$$

where $\zeta(x) := \sum_{k=1}^{\infty} k^{-x}$ denotes the Riemann zeta function for $x > 1$.

Significance: Suppose that $f \in H_{s,\boldsymbol{\gamma}}$ for all $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u} \subseteq \{1:s\}}$. Then for any given sequence of weights $\boldsymbol{\gamma}$, we can use the CBC algorithm to obtain a generating vector satisfying the error bound

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} |I_s f - Q_{n,s}^{\boldsymbol{\Delta}} f|^2} \le \left(\frac{1}{\varphi(n)} \sum_{\varnothing \ne \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/(2\lambda)} \|f\|_{s,\boldsymbol{\gamma}} \quad (7)$$

for all $\lambda \in (1/2, 1]$. We can use the following strategy:

- For a given integrand f, estimate the norm $\|f\|_{s,\boldsymbol{\gamma}}$.
- Find weights $\boldsymbol{\gamma}$ which *minimize* the error bound (7).
- Using the optimized weights $\boldsymbol{\gamma}$ as input, use the CBC algorithm to find a generating vector which *satisfies* the error bound (7).

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) $n \ge 3$, $\frac{1}{\varphi(n)} \le \frac{\mathrm{e}^{\gamma} \log\log n + \frac{3}{\log\log n}}{n}$, where $\gamma = 0.57721566\ldots$ (Euler–Mascheroni constant).
- The optimal convergence rate close to $\mathcal{O}(n^{-1})$ is obtained with $\lambda \to 1/2$, but note that $\lambda = 1/2$ is not permitted since $\zeta(2\lambda) \to \infty$ as $\lambda \to 1/2$.

3. Constructing lattice rules

Naïve CBC construction

We write the error criterion as

$$[e^{\mathrm{sh}}_{n,d}(z_1,\dots,z_d)]^2 = \frac{1}{n}\sum_{k=0}^{n-1}\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:d\}}\gamma_{\mathfrak{u}}\prod_{j\in\mathfrak{u}}B_2\left(\left\{\frac{kz_j}{n}\right\}\right)$$

$$= \frac{1}{n}\sum_{k=0}^{n-1}\sum_{\ell=1}^{d}\underbrace{\sum_{\substack{|\mathfrak{u}|=\ell\\ \mathfrak{u}\subseteq\{1:d\}}}\gamma_{\mathfrak{u}}\prod_{j\in\mathfrak{u}}B_2\left(\left\{\frac{kz_j}{n}\right\}\right)}_{=:p_{d,\ell}(k)}.$$

By plugging in the POD weights $\gamma_{\mathfrak{u}} := \Gamma_{|\mathfrak{u}|}\prod_{j\in\mathfrak{u}}\gamma_j$, note that we have the following recursion (we split the sum over $\mathfrak{u}$ in two parts depending on whether $d\in\mathfrak{u}$):

$$p_{d,\ell}(k) = \sum_{\substack{|\mathfrak{u}|=\ell\\ \mathfrak{u}\subseteq\{1:d\}}}\Gamma_\ell\left(\prod_{j\in\mathfrak{u}}\gamma_j B_2\left(\left\{\frac{kz_j}{n}\right\}\right)\right)$$

$$= \sum_{\substack{|\mathfrak{u}|=\ell\\ \mathfrak{u}\subseteq\{1:d-1\}}}\Gamma_\ell\left(\prod_{j\in\mathfrak{u}}\gamma_j B_2\left(\left\{\frac{kz_j}{n}\right\}\right)\right)$$

$$+ \sum_{\substack{|\mathfrak{u}|=\ell-1\\ \mathfrak{u}\subseteq\{1:d-1\}}}\Gamma_\ell\gamma_d B_2\left(\left\{\frac{kz_d}{n}\right\}\right)\left(\prod_{j\in\mathfrak{u}}\gamma_j B_2\left(\left\{\frac{kz_j}{n}\right\}\right)\right)$$

$$= p_{d-1,\ell}(k) + \frac{\Gamma_\ell}{\Gamma_{\ell-1}}\gamma_d B_2\left(\left\{\frac{kz_d}{n}\right\}\right)p_{d-1,\ell-1}(k).$$

Plugging the recurrence

$$p_{d,\ell}(k) = p_{d-1,\ell}(k) + \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d B_2\left(\left\{\frac{k z_d}{n}\right\}\right) p_{d-1,\ell-1}(k)$$

into the expression for the squared shift-averaged WCE yields

$$\begin{aligned}
[e^{\mathrm{sh}}_{n,d}(z_1,\ldots,z_d)]^2 &= \frac{1}{n}\sum_{k=0}^{n-1}\sum_{\ell=1}^{d} p_{d,\ell}(k) \\
&= \frac{1}{n}\sum_{k=0}^{n-1}\sum_{\ell=1}^{d} p_{d-1,\ell}(k) + \frac{1}{n}\sum_{k=0}^{n-1}\sum_{\ell=1}^{d} \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d B_2\left(\left\{\frac{k z_d}{n}\right\}\right) p_{d-1,\ell-1}(k) \\
&= [e^{\mathrm{sh}}_{n,d-1}(z_1,\ldots,z_{d-1})]^2 + \frac{1}{n}\sum_{k=0}^{n-1} B_2\left(\left\{\frac{k z_d}{n}\right\}\right) \sum_{\ell=1}^{d} \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d p_{d-1,\ell-1}(k).
\end{aligned}$$

Recall that in the d^{th} step of the CBC algorithm, the components $z_1,\ldots,z_{d-1}$ are fixed and it is therefore sufficient to find $z_d \in \mathbb{U}_n$ which minimizes the expression $\sum_{k=0}^{n-1} B_2\left(\left\{\frac{k z_d}{n}\right\}\right) \sum_{\ell=1}^{d} \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d p_{d-1,\ell-1}(k)$.

Let us introduce the matrix $\Omega_n := \left[B_2\left(\left\{\frac{kz}{n}\right\}\right)\right]_{\substack{z\in\mathbb{U}_n \\ k\in\{0,\ldots,n-1\}}}$ and define a set of n-vectors recursively via

$$\boldsymbol{p}_{d,\ell} = \boldsymbol{p}_{d-1,\ell} + \gamma_d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \Omega_n(z_d, :) .* \boldsymbol{p}_{d-1,\ell-1}$$

starting from the initial values

$$\begin{aligned} \boldsymbol{p}_{d,0} &= \mathbf{1}_n \quad \text{for all } d \geq 1, \\ \boldsymbol{p}_{d,\ell} &= \mathbf{0}_n \quad \text{for all } d \geq 1 \text{ and } \ell > d, \end{aligned}$$

with $.*$ denoting the componentwise product between two vectors.

Then the value of $\sum_{k=0}^{n-1} B_2\left(\left\{\frac{kz_d}{n}\right\}\right) \sum_{\ell=1}^{d} \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d p_{d-1,\ell-1}(k)$ in the d^{th} step of the CBC algorithm can be obtained for all $z_d \in \mathbb{U}_n$ via

$$\Omega_n \boldsymbol{x}, \quad \text{where } \boldsymbol{x} = \sum_{\ell=1}^{d} \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d \boldsymbol{p}_{d-1,\ell-1}.$$

CBC algorithm – naïve version

1. Define the matrix $\Omega_n := \left[B_2\left(\left\{\frac{kz}{n}\right\}\right)\right]_{\substack{z\in\mathbb{U}_n \\ k\in\{0,\ldots,n-1\}}}$ and initialize the n-vectors

$$\boldsymbol{p}_{d,0} = \mathbf{1}_n \quad \text{for all } d \geq 1,$$
$$\boldsymbol{p}_{d,\ell} = \mathbf{0}_n \quad \text{for all } d \geq 1 \text{ and } \ell > d.$$

for $d = 1, \ldots, s$, **do**

2. Pick the value $z_d \in \{1, \ldots, n-1\}$ corresponding to the smallest entry in the matrix-vector product

$$\Omega_n \boldsymbol{x}, \quad \text{where } \boldsymbol{x} = \sum_{\ell=1}^{d} \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_d \boldsymbol{p}_{d-1,\ell-1}. \tag{8}$$

3. Update $\boldsymbol{p}_{d,\ell} = \boldsymbol{p}_{d-1,\ell} + \gamma_d \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \Omega_n(z_d, :) .* \boldsymbol{p}_{d-1,\ell-1}$.

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 $\Omega_n \boldsymbol{x}$ in (8), 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 $\Omega_n \boldsymbol{x}$ 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 $\Omega_n \boldsymbol{x}$ 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 $\Omega'_n := \Omega_n(1 : \mathtt{m}, 2 : \mathtt{m} + 1)$, 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 $\varphi(n) := |\{k \in \mathbb{N} \mid 1 \leq k \leq n-1,\ \gcd(k,n) = 1\}|$. 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 $d \in \left\{ \frac{\varphi(n)}{p_1}, \ldots, \frac{\varphi(n)}{p_\ell} \right\}$. To see this, let d be any proper divisor of $\varphi(n)$. Then there exists j such that $d | \frac{\varphi(n)}{p_j}$, 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 $g^{\frac{\varphi(n)}{p_j}} \not\equiv 1 \pmod{n}$ for all $j \in \{1, \ldots, \ell\}$.

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 $g = 1, 2, \ldots, n-1$ and, for each number, check whether it is a primitive root by doing the following:

   2. Calculate $\mathrm{mod}(g^{\frac{\varphi(n)}{p_j}}, n)$ for all $j \in \{1, \ldots, \ell\}$.
   3. If all the calculated values are different from 1, then g is a primitive root.

Remark: In MATLAB, the quantities in step 2 can be computed, e.g., via `powermod(g,eulerPhi(n)/pj,n)`

Discrete and fast Fourier transform

The *discrete Fourier transform* of (complex) vector $\boldsymbol{x} := (x_j)_{j=1}^n$ is defined as the vector $\boldsymbol{y} := (y_j)_{j=1}^n$ with

$$y_j = \sum_{k=1}^{n} x_k \mathrm{e}^{-2\pi \mathrm{i}(j-1)(k-1)/n}, \quad j \in \{1, \ldots, n\},$$

and the *inverse discrete Fourier transform* is given by

$$x_j = \frac{1}{n} \sum_{k=1}^{n} y_k \mathrm{e}^{2\pi \mathrm{i}(j-1)(k-1)/n}, \quad j \in \{1, \ldots, n\}.$$

The *fast Fourier transform (FFT)* can be used to carry out these operations in $\mathcal{O}(n \log n)$ time. In MATLAB, one has $\boldsymbol{y} = \texttt{fft}(\boldsymbol{x})$ and $\boldsymbol{x} = \texttt{ifft}(\boldsymbol{y})$.

Circular convolution

Let $\boldsymbol{x} := (x_i)_{i=1}^n$ and $\boldsymbol{y} := (y_i)_{i=1}^n$ be (complex) vectors. Then the sequence $\boldsymbol{z} := (z_i)_{i=1}^n$ defined by

$$z_i = \sum_{k=1}^{n} x_k y_{\mathrm{mod}(i-k,n)+1}, \quad i \in \{1, \ldots, n\},$$

is called the *circular convolution* of $\boldsymbol{x}$ and $\boldsymbol{y}$ and we denote it by $\boldsymbol{z} := \boldsymbol{x} \star \boldsymbol{y}$.

Similarly to the continuous convolution, we have the following identity using discrete/fast Fourier transform:

$$\texttt{fft}(\boldsymbol{x} \star \boldsymbol{y}) = \texttt{fft}(\boldsymbol{x}).*\texttt{fft}(\boldsymbol{y}),$$

where $\boldsymbol{x}.*\boldsymbol{y} := (x_i y_i)_{i=1}^n$ 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:

$$A\boldsymbol{x} = \boldsymbol{a} \star \boldsymbol{x}, \tag{9}$$

  where $\boldsymbol{a} := [a_0, a_1, \ldots, a_{n-1}]^{\mathrm{T}}$ is the first column of matrix A.
- The identity (9) implies that a circulant matrix-vector product can be implemented in $\mathcal{O}(n \log n)$ time as $A\boldsymbol{x} = \texttt{ifft}(\texttt{fft}(\boldsymbol{a}).*\texttt{fft}(\boldsymbol{x}))$.

Putting it all together

The matrix-vector product $\Omega_n \boldsymbol{x}$ 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 $\rightarrow$ the matrix-vector product can be implemented in $\mathcal{O}(n \log n)$ operations using FFT.

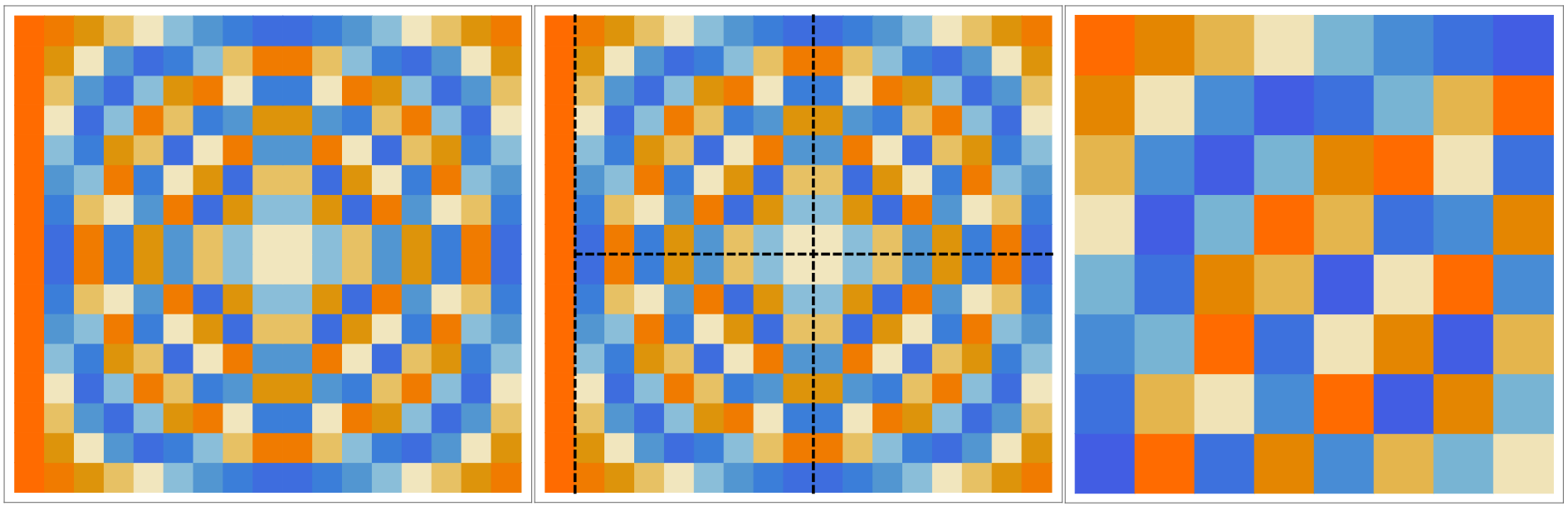


Figure: Example with Ω_{17}. Note that the first column is a constant and can be left out (the components of $\Omega_n \boldsymbol{x}$ are shifted by a constant $\rightarrow$ the smallest component stays invariant). Noting the obvious symmetries in the remaining four blocks, we can focus on the top left block.

When n is prime, it is possible to use the so-called Rader transformation to permute the top-left $m \times m$ matrix Ω' into circulant form:

$$\Omega_n^g(i,j) = \Omega'_n(g^i, (g^{-1})^j), \quad i,j \in \{1,\ldots,m\},$$

where g is the primitive root modulo n. Here, g^{-1} denotes the modular multiplicative inverse $gg^{-1} \equiv 1 \pmod{n}$.

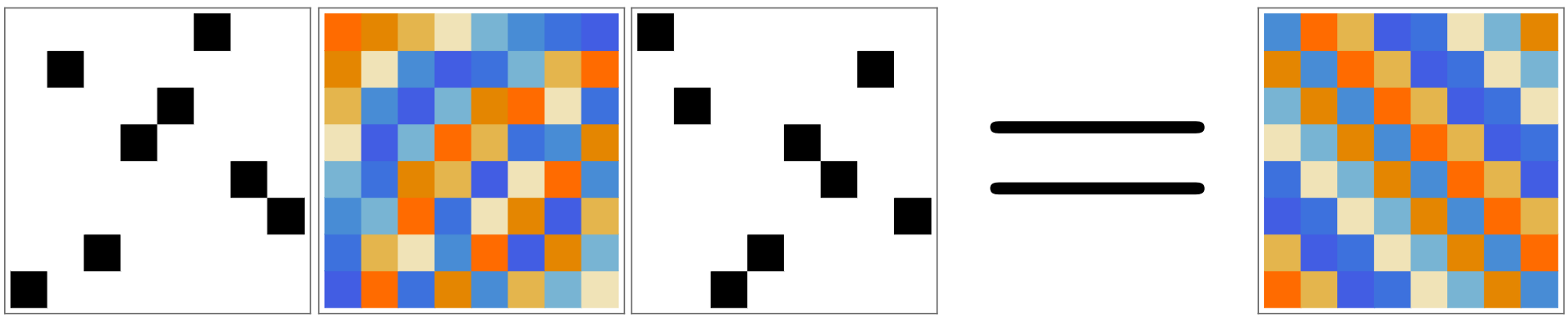


Figure: The original block matrix is multiplied from both sides by Rader permutation matrices (the black elements indicate the value 1 and white elements indicate the value 0) to obtain a circulant matrix.

Example with $n = 1009$

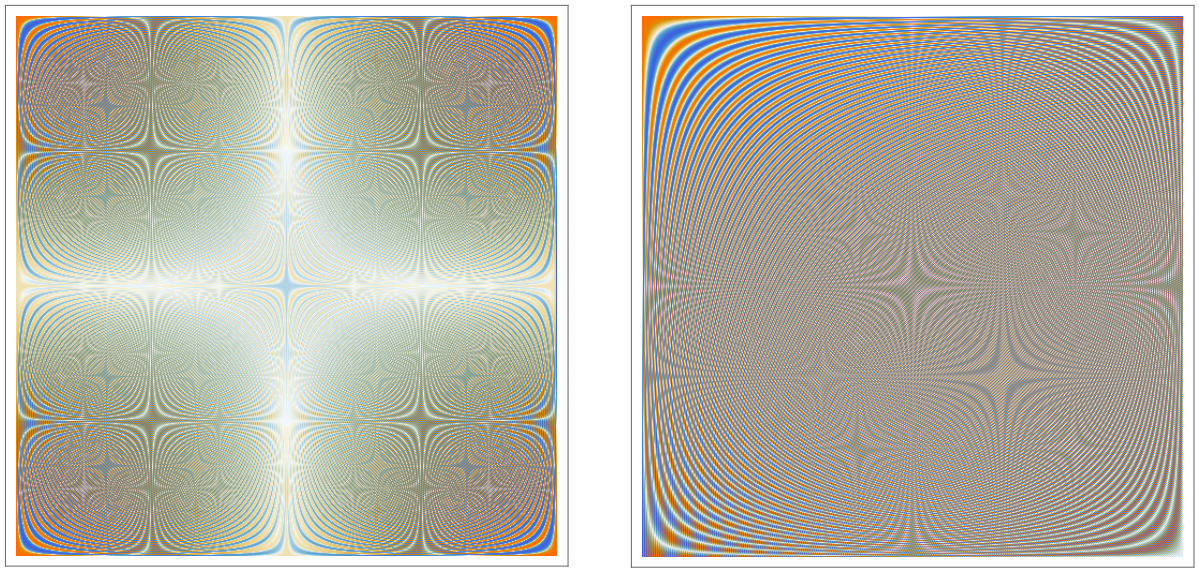


Figure: LHS: Original Ω_{1009}. RHS: top left block of Ω_{1009} (sans first column).

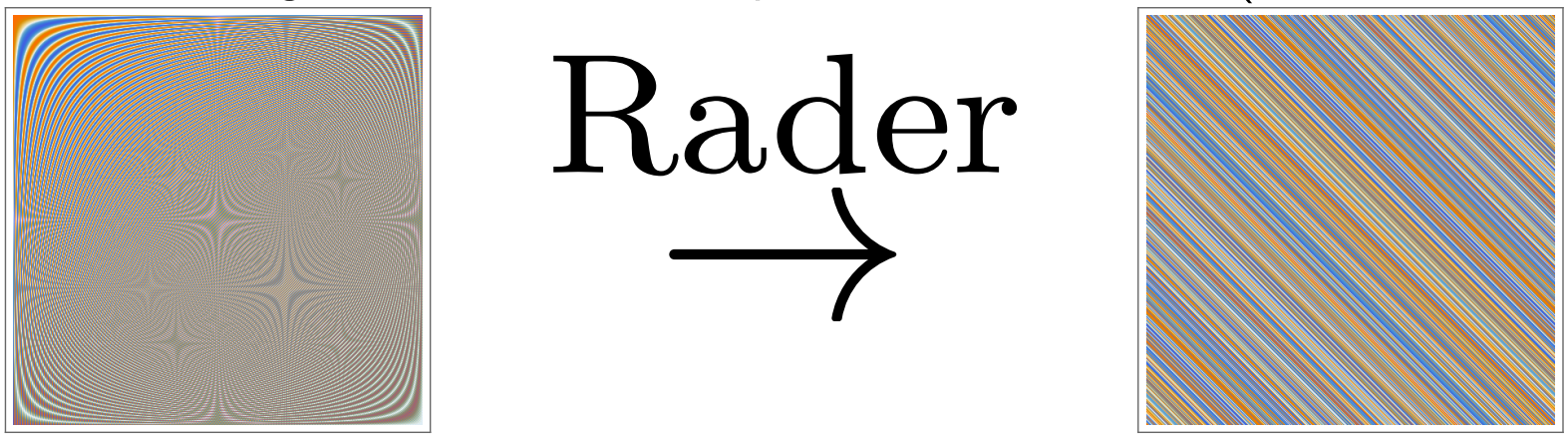


Figure: Rader transformation turns the top left block matrix circulant.

- 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/, ...

4. QMC methods for forward and inverse uncertainty quantification of elliptic PDEs with random coefficients

Uniform and affine model

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

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

where the diffusion coefficient has the parameterization

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

where $a_0 \in L^\infty(D)$, there exist $a_{\min}, a_{\max} > 0$ s.t. $0 < a_{\min} \leq a(\boldsymbol{x},\boldsymbol{y}) \leq a_{\max} < \infty$ for all $\boldsymbol{x} \in D$ and $\boldsymbol{y} \in U$, and the *stochastic fluctuations* $\psi_j : D \to \mathbb{R}$ 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 $\boldsymbol{y} \in [-1/2, 1/2]^{\mathbb{N}}$ to a finite number of terms. Moreover, the PDE needs to be discretized spatially using, e.g., the finite element method.

Let $u_s(\cdot, \boldsymbol{y}) := u_s(y_1, \ldots, y_s, 0, 0, \ldots)$ denote the dimensionally-truncated PDE solution for $\boldsymbol{y} \in [-1/2, 1/2]^s$, and let $u_{s,h}(\cdot, \boldsymbol{y}) \in V_h$ denote the dimensionally-truncated FE solution in the FE subspace spanned by piecewise linear FE basis functions. Furthermore, let $(\boldsymbol{t}_i)_{i=1}^n = (\{\frac{i\boldsymbol{z}}{n}\} - \frac{\mathbf{1}}{2})_{i=1}^n$ 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 $u(\cdot, \boldsymbol{y}) \in H_0^1(D)$ is the PDE solution for $\boldsymbol{y} \in [-1/2, 1/2]^{\mathbb{N}}$ and $G\colon H_0^1(D) \to \mathbb{R}$ is a linear functional (quantity of interest). We decompose the total error as

$$
\begin{aligned}
&\int_{[-1/2,1/2]^{\mathbb{N}}} G(u(\cdot, \boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} G(u_{s,h}(\cdot, \boldsymbol{t}_i)) \\
&= \int_{[-1/2,1/2]^{\mathbb{N}}} (G(u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y}_{\le s})))\,\mathrm{d}\boldsymbol{y} \\
&+ \int_{[-1/2,1/2]^{s}} G(u_s(\cdot, \boldsymbol{y}) - u_{s,h}(\cdot, \boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} \\
&+ \int_{[-1/2,1/2]^{s}} G(u_{s,h}(\cdot, \boldsymbol{y}))\,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} G(u_{s,h}(\cdot, \boldsymbol{t}_i)).
\end{aligned}
$$

Using the triangle inequality, we are left with the total error decomposition

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

We focus on the cubature error.

- If $\|\psi_1\|_{L^\infty(D)} \geq \|\psi_2\|_{L^\infty(D)} \geq \cdots$, then the dimension truncation error decays like $\mathcal{O}(s^{-2/p+1})$ [Gantner (2018)].
- If D is a convex polygon (2d)/polyhedron (3d) and we have additional regularity, e.g., $f, G \in L^2(D)$, a is Lipschitz, and the family $\{V_h\}_h$ of first-order finite element spaces, indexed by the mesh size $h > 0$, is a sequence of regular, simplicial meshes in D obtained from an initial, regular triangulation of D by recursive, uniform bisection of simplices, then the L^2 finite element error satisfies $\mathcal{O}(h^2)$ as $h \to 0$ independently of s [Kuo, Schwab, Sloan (2012)].

Multi-index notation

We introduce the set of *finitely-supported multi-indices*

$$\mathscr{F} := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\mathrm{supp}(\boldsymbol{\nu})| < \infty\},$$

where the *support* of a multi-index $\boldsymbol{\nu}$ is defined as the set

$$\mathrm{supp}(\boldsymbol{\nu}) := \{i \in \mathbb{N} : \nu_i \neq 0\}.$$

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

$$\partial^{\boldsymbol{\nu}} := \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} := \prod_{j \in \mathrm{supp}(\boldsymbol{\nu})} \frac{\partial^{\nu_j}}{\partial y_j^{\nu_j}}, \quad \boldsymbol{x}^{\boldsymbol{\nu}} := \prod_{j \in \mathrm{supp}(\boldsymbol{\nu})} x_j^{\nu_j}, \quad \binom{\boldsymbol{\nu}}{\boldsymbol{m}} := \prod_{j \in \mathrm{supp}(\boldsymbol{\nu})} \binom{\nu_j}{m_j}.$$

Recursive bound

Consider the weak formulation

$$\int_D a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \langle f, v\rangle_{H^{-1}(D),H_0^1(D)}. \tag{10}$$

Noting that

$$\partial^{\boldsymbol{\nu}} a(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} a(\boldsymbol{x},\boldsymbol{y}) & \text{if } \boldsymbol{\nu}=\boldsymbol{0},\\ \psi_j(\boldsymbol{x}) & \text{if } \boldsymbol{\nu}=\boldsymbol{e}_j,\\ 0 & \text{otherwise,}\end{cases}$$

we let $\boldsymbol{\nu}\in\mathscr{F}\setminus\{\boldsymbol{0}\}$ and differentiate (10) on both sides with $\partial^{\boldsymbol{\nu}}$ and use the Leibniz product rule[†] to obtain

$$\partial^{\boldsymbol{\nu}}\int_D a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$$

$$\Leftrightarrow \quad \sum_{\boldsymbol{m}\le\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\int_D \partial^{\boldsymbol{m}} a(\boldsymbol{x})\nabla\partial^{\boldsymbol{\nu}-\boldsymbol{m}} u(\boldsymbol{x},\boldsymbol{y})\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$$

$$\Leftrightarrow \quad \int_D a(\boldsymbol{x},\boldsymbol{y})\nabla\partial^{\boldsymbol{\nu}} u(\boldsymbol{x},\boldsymbol{y})\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = -\sum_{j\in\mathrm{supp}(\boldsymbol{\nu})}\nu_j\int_D \psi_j(\boldsymbol{x})\nabla\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} u(\boldsymbol{x},\boldsymbol{y})\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

[†] $\partial^{\boldsymbol{\nu}}(fg) = \sum_{\boldsymbol{m}\le\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\partial^{\boldsymbol{m}} f\,\partial^{\boldsymbol{\nu}-\boldsymbol{m}} g$

Testing this against $v = \partial^{\boldsymbol{\nu}} u(\boldsymbol{x}, \boldsymbol{y})$ yields

$$\begin{aligned}
&a_{\min} \|\partial^{\boldsymbol{\nu}} u(\cdot, \boldsymbol{y})\|^2_{H^1_0(D)} \\
&\leq \int_D a(\boldsymbol{x}, \boldsymbol{y}) \|\nabla \partial^{\boldsymbol{\nu}} u(\boldsymbol{x}, \boldsymbol{y})\|^2 \, \mathrm{d}\boldsymbol{x} \\
&\leq \sum_{j \in \mathrm{supp}(\boldsymbol{\nu})} \nu_j \|\psi_j\|_{L^\infty(D)} \|\partial^{\boldsymbol{\nu} - \boldsymbol{e}_j} u(\cdot, \boldsymbol{y})\|_{H^1_0(D)} \|\partial^{\boldsymbol{\nu}} u(\cdot, \boldsymbol{y})\|_{H^1_0(D)}
\end{aligned}$$

Thus we obtain the recursive relation

$$\|\partial^{\boldsymbol{\nu}} u(\cdot, \boldsymbol{y})\|_{H^1_0(D)} \leq \sum_{j \in \mathrm{supp}(\boldsymbol{\nu})} \nu_j \underbrace{\frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}}_{=: b_j} \|\partial^{\boldsymbol{\nu} - \boldsymbol{e}_j} u(\cdot, \boldsymbol{y})\|_{H^1_0(D)}.$$

For later convenience, we introduce here the sequence $\boldsymbol{b} := (b_j)_{j \geq 1}$ 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 $\boldsymbol{b} \in \ell^p$ for some $p \in (0, 1)$.

Parametric regularity

Proposition

For all $\boldsymbol{y} \in [-1/2, 1/2]^{\mathbb{N}}$ *and* $\boldsymbol{\nu} \in \mathscr{F}$, *there holds*

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

Proof. By induction w.r.t. the order of the multi-index $\boldsymbol{\nu} \in \mathscr{F}$. If $\boldsymbol{\nu} = \mathbf{0}$, then this is the ordinary Lax–Milgram *a priori* bound

$$a_{\min} \underbrace{\int_D |\nabla u(\boldsymbol{x}, \boldsymbol{y})|^2 \,\mathrm{d}\boldsymbol{x}}_{= \|u(\cdot, \boldsymbol{y})\|^2_{H_0^1(D)}} \leq \int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla u(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla u(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x} = \langle f, u(\cdot, \boldsymbol{y}) \rangle_{H^{-1}(D), H_0^1(D)}$$

$$\leq \|f\|_{H^{-1}(D)} \|u(\cdot, \boldsymbol{y})\|_{H_0^1(D)}$$

whence

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

Next, let $\boldsymbol{\nu} \in \mathscr{F} \setminus \{\mathbf{0}\}$ and suppose that the claim has been proved for all multi-indices with order less than $|\boldsymbol{\nu}|$. Then using the recursive relation we derived previously, we obtain

$$\begin{aligned}
\|\partial^{\boldsymbol{\nu}} u(\cdot, \boldsymbol{y})\|_{H_0^1(D)} &\leq \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} \nu_j b_j \|\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} u(\cdot, \boldsymbol{y})\|_{H_0^1(D)} \\
&\leq \frac{\|f\|_{H^{-1}(D)}}{a_{\min}} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} \nu_j b_j |\boldsymbol{\nu} - \boldsymbol{e}_j|! \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{e}_j} \\
&= \frac{\|f\|_{H^{-1}(D)}}{a_{\min}} \boldsymbol{b}^{\boldsymbol{\nu}} (|\boldsymbol{\nu}| - 1)! \sum_{j \geq 1} \nu_j \\
&= \frac{\|f\|_{H^{-1}(D)}}{a_{\min}} \boldsymbol{b}^{\boldsymbol{\nu}} |\boldsymbol{\nu}|!,
\end{aligned}$$

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 $G\colon H_0^1(D) \to \mathbb{R}$ be a linear and bounded functional, $u_{s,h}$ the dimensionally-truncated FE solution, and define $F(\boldsymbol{y}) := G(u_{s,h}(\cdot, \boldsymbol{y} - \frac{\mathbf{1}}{\mathbf{2}}))$ for $\boldsymbol{y} \in [0,1]^s$. Let $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u} \subseteq \{1:s\}}$ be a sequence of positive weights. Then we know that the generating vector obtained by the CBC algorithm satisfies the error bound

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} |I_s F - Q_{n,s}^{\boldsymbol{\Delta}} F|^2} \le \left(\frac{1}{\varphi(n)} \sum_{\varnothing \ne \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{1/(2\lambda)} \|F\|_{s,\boldsymbol{\gamma}}$$

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

$$\begin{aligned} \|F\|_{s,\boldsymbol{\gamma}}^2 &= \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[0,1]^{|\mathfrak{u}|}} \left(\int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{x}_{\mathfrak{u}}} F(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}_{-\mathfrak{u}} \right)^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \\ &\le \left(\frac{\|G\|_{H^{-1}(D)} \|f\|_{H^{-1}(D)}}{a_{\min}} \right)^2 \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} (|\mathfrak{u}|!)^2 \prod_{j \in \mathfrak{u}} b_j^2. \end{aligned}$$

Plugging this norm bound back into the QMC error bound yields...

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}}|I_s F - Q^{\boldsymbol{\Delta}}_{n,s} F|^2} \lesssim \left(\frac{1}{\varphi(n)}\right)^{1/(2\lambda)} \left(\sum_{\varnothing \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}}\right)^{|\mathfrak{u}|}\right)^{1/(2\lambda)}$$
$$\times \left(\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} (|\mathfrak{u}|!)^2 \prod_{j \in \mathfrak{u}} b_j^2\right)^{1/2}.$$

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

$$g(\boldsymbol{\gamma}) := \left(\sum_i \alpha_i \gamma_i^{\lambda}\right)^{1/\lambda} \left(\sum_i \beta_i \gamma_i^{-1}\right)$$

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 $\gamma_i = c\big(\frac{\beta_i}{\alpha_i}\big)^{1/(1+\lambda)}$ into $\big(\sum_i \alpha_i \gamma_i^{\lambda}\big)^{1/(2\lambda)} \big(\sum_i \beta_i \gamma_i^{-1}\big)^{1/2}$ yields the expression $\big(\sum_i \alpha_i^{1/(1+\lambda)} \beta_i^{\lambda/(1+\lambda)}\big)^{(1+\lambda)/(2\lambda)}$. Thus, plugging the optimal POD weights into the QMC error bound results in

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} |I_s F - Q^{\boldsymbol{\Delta}}_{n,s} F|^2} \lesssim \left(\frac{1}{\varphi(n)}\right)^{1/(2\lambda)} C(s, \boldsymbol{\gamma}, \lambda)^{(1+\lambda)/(2\lambda)},$$

where

$$C(s, \boldsymbol{\gamma}, \lambda) := \sum_{\mathfrak{u} \subseteq \{1:s\}} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}}\right)^{|\mathfrak{u}|/(1+\lambda)} (|\mathfrak{u}|!)^{2\lambda/(1+\lambda)} \prod_{j \in \mathfrak{u}} b_j^{2\lambda/(1+\lambda)}.$$

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 $C(\boldsymbol{\gamma}, \lambda) < \infty$ independently of s s.t. $C(s, \boldsymbol{\gamma}, \lambda) \leq C(\boldsymbol{\gamma}, \lambda) < \infty$.

Proof. First observe that

$$C(s,\gamma,\lambda) = \sum_{\mathfrak{u}\subseteq\{1:s\}} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}\right)^{|\mathfrak{u}|/(1+\lambda)} (|\mathfrak{u}|!)^{2\lambda/(1+\lambda)} \prod_{j\in\mathfrak{u}} b_j^{2\lambda/(1+\lambda)}$$

$$= \sum_{\ell=0}^{s} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}\right)^{\ell/(1+\lambda)} (\ell!)^{2\lambda/(1+\lambda)} \sum_{\substack{|\mathfrak{u}|=\ell \\ \mathfrak{u}\subseteq\{1:s\}}} \prod_{j\in\mathfrak{u}} b_j^{2\lambda/(1+\lambda)}$$

$$\le \sum_{\ell=0}^{\infty} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}\right)^{\ell/(1+\lambda)} (\ell!)^{2\lambda/(1+\lambda)-1} \left(\sum_{j\ge1} b_j^{2\lambda/(1+\lambda)}\right)^{\ell}$$

where we used the inequality $\sum_{|\mathfrak{u}|=\ell,\mathfrak{u}\subseteq\mathbb{Z}_+} \prod_{j\in\mathfrak{u}} c_j \le \frac{1}{\ell!}\left(\sum_{j\ge1} c_j\right)^\ell$.

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

$$C(s,\gamma,\lambda) \le \sum_{\ell=0}^{\infty} \underbrace{\left(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}\right)^{\ell/(1+\lambda)} (\ell!)^{p-1} \left(\sum_{j\ge1} b_j^p\right)^{\ell}}_{=:a_\ell}$$

It is easy to see that $\frac{a_{\ell+1}}{a_\ell} \xrightarrow{\ell\to\infty} 0$. 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, $\|\boldsymbol{b}\|_{\ell^{2\lambda/(1+\lambda)}} \leq \|\boldsymbol{b}\|_{\ell^p}$, and we obtain from the estimate on the previous slide that

$$C(s, \boldsymbol{\gamma}, \lambda) \leq \sum_{\ell=0}^{\infty} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{\ell/(1+\lambda)} (\ell!)^{2\lambda/(1+\lambda)-1} \left(\sum_{j \geq 1} b_j^{2\lambda/(1+\lambda)} \right)^{\ell}$$

$$\leq \sum_{\ell=0}^{\infty} \underbrace{\left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{\ell/(1+\lambda)} (\ell!)^{2/(3-2\delta)-1} \left(\sum_{j \geq 1} b_j^p \right)^{2\ell/((3-2\delta)p)}}_{=: a_\ell}$$

Again, $\frac{a_{\ell+1}}{a_\ell} \xrightarrow{\ell \to \infty} 0$, 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 $G : H_0^1(D) \to \mathbb{R}$ be a bounded linear functional. Consider the problem of approximating

$$\mathbb{E}[G(u_{s,h})] = \int_{[-1/2,1/2]^s} G(u_{s,h}(\cdot, \boldsymbol{y}))\,\mathrm{d}\boldsymbol{y},$$

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 $\boldsymbol{z} \in \mathbb{N}^s$. The generating vector is designed to be used to compute the estimate

$$\overline{Q}_{n,s,R} G(u_{s,h}) := \frac{1}{R} \sum_{r=0}^{R-1} Q_{n,s}^{\boldsymbol{\Delta}_r} G(u_{s,h}),$$

where $Q_{n,s}^{\boldsymbol{\Delta}_r} F := \frac{1}{n} \sum_{i=0}^{n-1} f(\{\boldsymbol{t}_i + \boldsymbol{\Delta}_r\} - \frac{\mathbf{1}}{2})$, $\boldsymbol{t}_k := \{\frac{k\boldsymbol{z}}{n}\}$, and $\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_{R-1}$ 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

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

- If we instead wish to estimate $\mathbb{E}[u_{s,h}(\boldsymbol{x}, \cdot)]$ (i.e., leave out the *quantity of interest* $G \colon H_0^1(D) \to \mathbb{R}$), the same weights can be used as input to the CBC algorithm.

Numerical example

Let us consider the PDE problem

$$-\nabla \cdot (a(\boldsymbol{x}, \boldsymbol{y}) \nabla u(\boldsymbol{x}, \boldsymbol{y})) = x_1, \quad u(\cdot, \boldsymbol{y})|_{\partial D} = 0,$$

in the physical domain $D = (0,1)^2$ with the diffusion coefficient

$$a(\boldsymbol{x}, \boldsymbol{y}) = 1 + \sum_{j=1}^{s} y_j \psi_j(\boldsymbol{x}), \quad \boldsymbol{x} \in D, \ y_j \in [-\tfrac{1}{2}, \tfrac{1}{2}],$$

where $\psi_j(\boldsymbol{x}) = j^{-2} \sin(j\pi x_1) \sin(j\pi x_2)$. We compute $\mathbb{E}[G(u)]$ using QMC with $R = 8$ random shifts, where $G(v) = \int_D v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$.

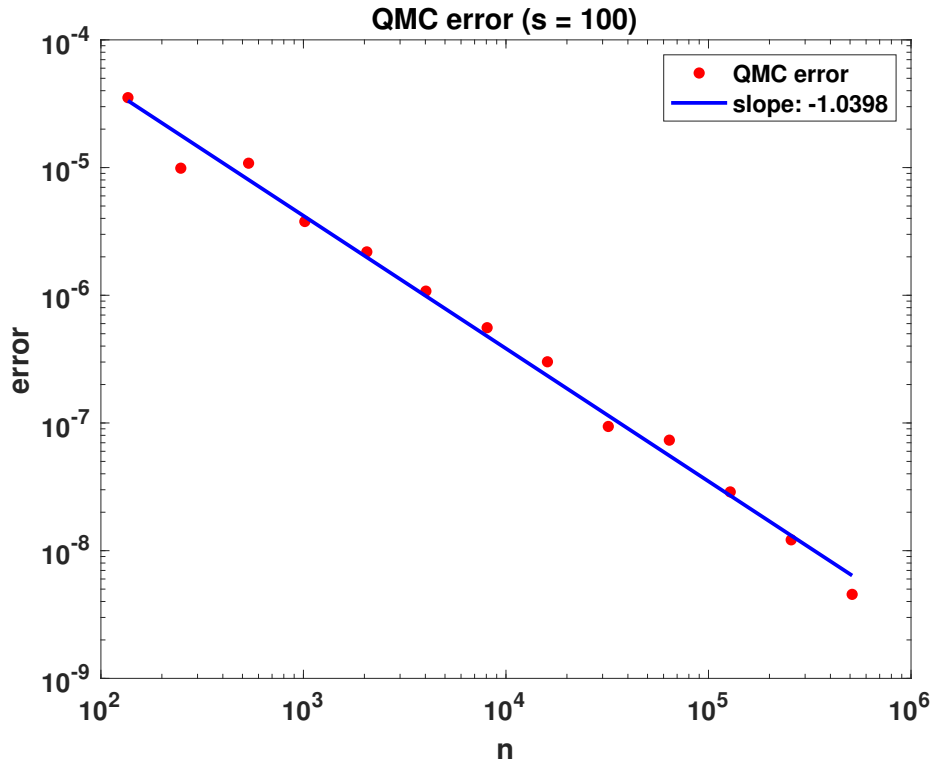


Figure: QMC with $s = 100$ constructed using the weights $\gamma_{\mathfrak{u}} = \left(|\mathfrak{u}|! \prod_{j \in \mathfrak{u}} \frac{b_j}{\sqrt{2\zeta(2\lambda)/(2\pi^2)^\lambda}}\right)^{\frac{2}{1+\lambda}}$, $\lambda = \frac{1}{2-2\delta}$, $\delta = 0.05$, for all $\mathfrak{u} \subseteq \{1, \ldots, s\}$.

Lognormal model (briefly)

Lognormal model: let $D \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, 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

$$U_{\boldsymbol{b}} := \left\{ \boldsymbol{y} \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} b_j |y_j| < \infty \right\}.$$

Consider the problem of finding, for all $\boldsymbol{y} \in U$, $u(\cdot, \boldsymbol{y}) \in H_0^1(D)$ such that

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

where the diffusion coefficient is assumed to have the parameterization

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

where $a_0 \in L^\infty(D)$ is such that $a_0(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in D$.

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 $\boldsymbol{y} \in U_{\boldsymbol{b}}$, the expressions $a_{max}(\boldsymbol{y}) := \max_{\boldsymbol{x} \in \overline{D}} a(\boldsymbol{x}, \boldsymbol{y})$ and $a_{\min}(\boldsymbol{y}) := \min_{\boldsymbol{x} \in \overline{D}} a(\boldsymbol{x}, \boldsymbol{y})$ are well-defined and satisfy $0 < a_{\min}(\boldsymbol{y}) \leq a(\boldsymbol{x}, \boldsymbol{y}) \leq a_{\max}(\boldsymbol{y}) < \infty$.

(B3) $\sum_{j=1}^\infty b_j^p < \infty$ for some $p \in (0, 1)$.

Remark: Note that in the lognormal case, $a(\boldsymbol{x}, \boldsymbol{y})$ can take values which are arbitrarily close to 0 or arbitrarily large. Thus, the best we can do is to find $\boldsymbol{y}$-dependent lower and upper bounds $a_{\min}(\boldsymbol{y})$ and $a_{\max}(\boldsymbol{y})$. This will lead to a $\boldsymbol{y}$-dependent *a priori* bound and, consequently, $\boldsymbol{y}$-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

$$I_s(F) := \int_{\mathbb{R}^s} F(\boldsymbol{y}) \prod_{j=1}^{s} \phi(y_j)\,\mathrm{d}\boldsymbol{y} = \int_{(0,1)^s} F(\Phi^{-1}(\boldsymbol{w}))\,\mathrm{d}\boldsymbol{w}.$$

where $\phi(y) := \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}y^2}$ is the probability density function of $\mathcal{N}(0,1)$ and $\Phi^{-1}(\boldsymbol{w}) = [\Phi^{-1}(w_1), \ldots, \Phi^{-1}(w_s)]^{\mathrm{T}}$ denotes the corresponding (componentwise) inverse cumulative distribution function. We use the randomly shifted QMC rules

$$Q_{n,s}^{\boldsymbol{\Delta}_r}(F) = \frac{1}{n}\sum_{k=1}^{n} F(\Phi^{-1}(\{\boldsymbol{t}_k + \boldsymbol{\Delta}_r\})),$$

$$\overline{Q}_{n,R}(F) := \frac{1}{R}\sum_{r=1}^{R} Q_{n,s}^{\boldsymbol{\Delta}_r}(F),$$

where we have R independent random shifts $\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_R$ drawn from $\mathcal{U}([0,1]^s)$, $\boldsymbol{t}_k := \{\frac{k\boldsymbol{z}}{n}\}$, with generating vector $\boldsymbol{z} \in \mathbb{N}^s$.

The appropriate function space for unbounded integrands is a "special" weighted, unanchored Sobolev space equipped with the norm

$$\|F\|_{s,\boldsymbol{\gamma}} = \Bigg[\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \Bigg(\int_{\mathbb{R}^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_{\mathfrak{u}}} F(\boldsymbol{y}) \Bigg(\prod_{j \in \{1:s\} \setminus \mathfrak{u}} \phi(y_j) \Bigg) \, \mathrm{d}\boldsymbol{y}_{-\mathfrak{u}} \Bigg)^2 \times \Bigg(\prod_{j \in \mathfrak{u}} \varpi_j^2(y_j) \Bigg) \, \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \Bigg]^{1/2}$$

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 $\boldsymbol{\gamma}$, 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

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}}|I_s F - Q_{n,s}^{\boldsymbol{\Delta}} F|^2} \le \left(\frac{2}{n} \sum_{\varnothing \ne \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \prod_{j \in \mathfrak{u}} \varrho_j(\lambda)\right)^{1/(2\lambda)} \|F\|_{s,\boldsymbol{\gamma}},$$

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 $\|\cdot\|_{s,\boldsymbol{\gamma}}$ for a given integrand (2) find weights $\boldsymbol{\gamma}$ 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).

Proposition (Parametric regularity bound for the lognormal model — Graham, Kuo, Nichols, Scheichl, Schwab, Sloan (2015))

For all $\boldsymbol{y} \in U_{\boldsymbol{b}}$ and $\boldsymbol{\nu} \in \mathscr{F}$, there holds

$$\|\partial^{\boldsymbol{\nu}} u(\cdot, \boldsymbol{y})\|_{H_0^1(D)} \leq \frac{\|f\|_{H^{-1}(D)}}{\min_{\boldsymbol{x} \in \overline{D}} a_0(\boldsymbol{x})} \frac{|\boldsymbol{\nu}|!}{(\log 2)^{|\boldsymbol{\nu}|}} \boldsymbol{b}^{\boldsymbol{\nu}} \prod_{j \geq 1} \exp(b_j |y_j|).$$

This parametric regularity bound is valid also for the dimensionally-truncated finite element solution $u_{s,h}$. If $G : H_0^1(D) \to \mathbb{R}$ is a bounded linear functional and $F(\boldsymbol{y}) := G(u_{s,h}(\cdot, \boldsymbol{y}))$ for $\boldsymbol{y} \in \mathbb{R}^s$, then

$$\|F\|_{s,\boldsymbol{\gamma}}^2 \leq \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{(|\mathfrak{u}|!)^2}{\gamma_{\mathfrak{u}}} \left(\prod_{j=1}^{s} 2 \exp(2b_j^2) \Phi(2b_j) \right) \left(\prod_{j \in \mathfrak{u}} \frac{b_j^2}{2(\log 2)^2 \exp(2b_j^2) \Phi(2b_j)(\alpha_j - b_j)} \right).$$

By choosing $\alpha_j = \frac{1}{2}(b_j + \sqrt{b_j^2 + 1 - \frac{1}{2\lambda}})$ and using the POD weights

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

as inputs to the CBC algorithm yields a randomly shifted rank-1 lattice rule satisfying the R.M.S. error

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} |I_s F - Q_{n,s}^{\boldsymbol{\Delta}} F|^2} \lesssim n^{\max\{-1/p+1/2, -1+\delta\}},$$

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.

Numerical example

Let us consider the PDE problem

$$-\nabla \cdot (a(\boldsymbol{x}, \boldsymbol{y}) \nabla u(\boldsymbol{x}, \boldsymbol{y})) = x_1, \quad u(\cdot, \boldsymbol{y})|_{\partial D} = 0,$$

in the physical domain $D = (0,1)^2$ with the diffusion coefficient

$$a(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\sum_{j=1}^{s} y_j \psi_j(\boldsymbol{x})\right), \quad \boldsymbol{x} \in D, \ y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

where $\psi_j(\boldsymbol{x}) = j^{-2} \sin(j\pi x_1) \sin(j\pi x_2)$. We compute $\mathbb{E}[G(u)]$ using QMC with $R = 8$ random shifts, where $G(v) = \int_D v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$.

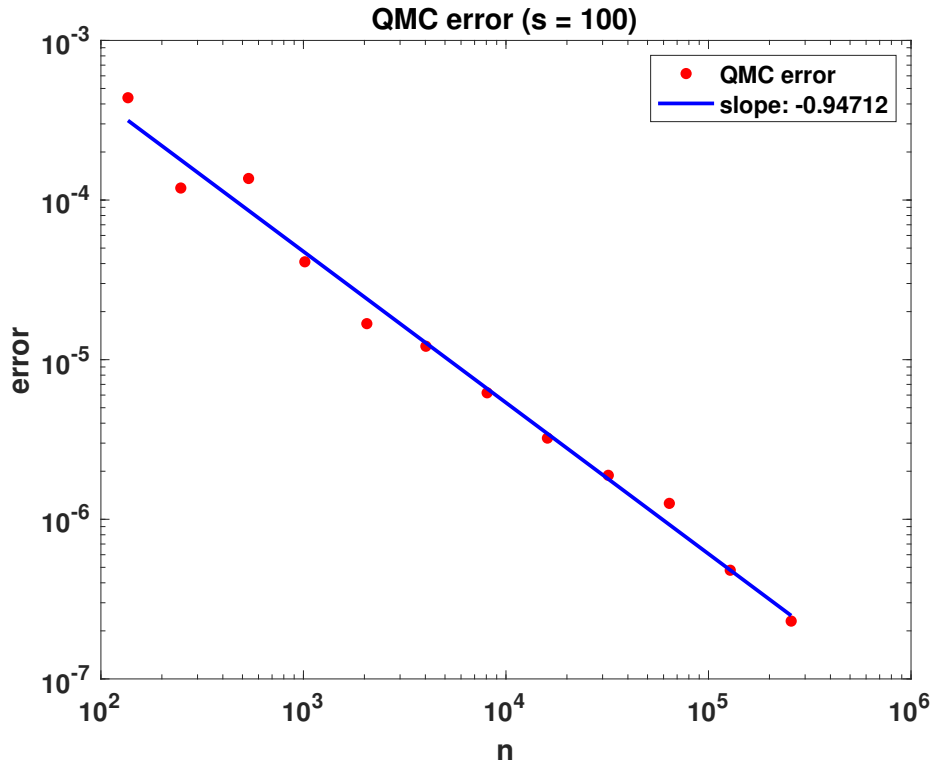


Figure: QMC with $s = 100$ constructed using the weights $\gamma_{\mathfrak{u}} = \big(|\mathfrak{u}|! \prod_{j\in\mathfrak{u}} \frac{b_j}{2(\log 2)\exp(b_j^2/2)\Phi(b_j)\sqrt{(\alpha_j - b_j)\varrho_j(\lambda)}}\big)^{\frac{2}{1+\lambda}}$, $\lambda = \frac{1}{2-2\delta}$, $\delta = 0.05$, for all $\mathfrak{u} \subseteq \{1, \ldots, s\}$.

Computational implementation

Consider the task of approximating $\int_{[0,1]^s} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$ using a randomly shifted rank-1 lattice rule with R random shifts.

Once a generating vector $\boldsymbol{z} \in \mathbb{N}^s$ has been obtained for a given number n of QMC nodes and dimension s (using, e.g., the CBC algorithm), then:

for $r = 1, \ldots, R$, **do**
  draw $\boldsymbol{\Delta}^{(r)} \sim \mathcal{U}([0,1]^s)$;
  initialize $Q_r = 0$;
  for $i = 1, \ldots, n$, **do**
    set $\boldsymbol{t}_i = \mathrm{mod}\left(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta}^{(r)}, 1\right)$;
    set $Q_r = Q_r + f(\boldsymbol{t}_i)$;
  end for
  set $Q_r = Q_r/n$;
end for

return $\overline{Q} = \frac{Q_1 + \cdots + Q_R}{R}$;
(This is the QMC estimator with R random shifts.)

Remarks:

- If integrating
$$\int_{\mathbb{R}^s} f(\boldsymbol{y}) \prod_{j=1}^{s} \frac{\mathrm{e}^{-\frac{1}{2}y_j^2}}{\sqrt{2\pi}}\,\mathrm{d}\boldsymbol{y}$$
then use $\boldsymbol{t}_i = \Phi^{-1}(\mathrm{mod}(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta}^{(r)}, 1))$, where Φ^{-1} is the (componentwise) inverse cumulative distribution function of $\mathcal{N}(0,1)$.
- The R.M.S. error can be estimated by
$$\text{R.M.S. error} \approx \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^{R} (\overline{Q} - Q_r)^2}.$$

Some perspectives on applying QMC for Bayesian inverse problems

Let $U \subseteq \mathbb{R}^s$ be a nonempty set of parameters and let $G\colon U \to \mathbb{R}^k$ be a forward mapping depending on some (unknown) parameter $\boldsymbol{y} \in U$.

Measurement model:

$$\boldsymbol{\delta} = G(\boldsymbol{y}) + \boldsymbol{\eta},$$

where $\boldsymbol{\delta} \in \mathbb{R}^k$ is the measurement data and $\boldsymbol{\eta} \in \mathbb{R}^k$ is Gaussian noise such that $\boldsymbol{\eta} \sim \mathcal{N}(0, \Gamma)$, where $\Gamma \in \mathbb{R}^{k\times k}$ is a symmetric, positive definite covariance matrix.

If we endow $\boldsymbol{y}$ with a *prior* density π_{pr} and $\boldsymbol{y}$ and $\boldsymbol{\eta}$ are independent, then Bayes' formula yields the *posterior distribution* with density

$$\pi(\boldsymbol{y}|\boldsymbol{\delta}) \propto \pi(\boldsymbol{\delta}|\boldsymbol{y})\pi_{\mathrm{pr}}(\boldsymbol{y}),$$

where we have the *likelihood* $\pi(\boldsymbol{\delta}|\boldsymbol{y}) \propto \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))}$.

The *conditional mean (CM)* estimator of the unknown parameter is

$$\boldsymbol{y}_{\mathrm{CM}} = \int_U \boldsymbol{y}\,\pi(\boldsymbol{y}|\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{y} = \frac{\int_U \boldsymbol{y}\,\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))}\pi_{\mathrm{pr}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}}{\int_U \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))}\pi_{\mathrm{pr}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}}.$$

For simplicity, let us make the following standing assumptions:

We have $U = U_s = [-\frac{1}{2}, \frac{1}{2}]^s$ and there is a constant $C \geq 1$ and a sequence $\boldsymbol{b} := (b_j)_{j \geq 1} \in \ell^p$ of nonnegative real numbers for some $p \in (0, 1)$ such that

(A1) the forward model satisfies

$$\|\partial^{\boldsymbol{\nu}} G(\boldsymbol{y})\|_{\mathbb{R}^k} \leq C|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}} \quad \text{for all } \boldsymbol{\nu} \in \mathscr{F} \text{ and } \boldsymbol{y} \in U, \text{ and}$$

(A2) $\pi_{\rm pr}(\boldsymbol{y}) = 1$ for $\boldsymbol{y} \in U$ and 0 otherwise.

(A3) The smallest eigenvalue of Γ is bounded from below by $0 < \mu_{\min} \leq 1$.

Example

Let $D \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, be a nonempty, bounded Lipschitz domain and let $f \in H^{-1}(D)$. For each $\boldsymbol{y} \in U$, there exists a weak solution $u(\cdot,\boldsymbol{y}) \in H_0^1(D)$ to

$$\begin{cases} -\nabla \cdot (a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})) = f(\boldsymbol{x}), & \boldsymbol{x} \in D,\ \boldsymbol{y} \in U, \\ u(\boldsymbol{x},\boldsymbol{y}) = 0, & \boldsymbol{x} \in \partial D,\ \boldsymbol{y} \in U, \end{cases}$$

where we assume that $\boldsymbol{y} = (y_j)_{j=1}^s$ are i.i.d. uniformly distributed in $[-1/2,1/2]$, and

$$a(\boldsymbol{x},\boldsymbol{y}) = a_0(\boldsymbol{x}) + \sum_{j=1}^{s} y_j \psi_j(\boldsymbol{x}), \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in [-1/2,1/2]^s,$$

with $a_0 \in L^\infty(D)$ and $\psi_j \in L^\infty(D)$, $j \geq 1$, such that $0 < a_{\min} \leq a(\boldsymbol{x},\boldsymbol{y}) \leq a_{\max} < \infty$ for all $\boldsymbol{x} \in D$, $\boldsymbol{y} \in [-1/2,1/2]^s$.

Let $\mathcal{O}_j : H_0^1(D) \to \mathbb{R}$ be linear, bounded observation operators for $j = 1,\dots,k$. Then $G(\boldsymbol{y}) = [\mathcal{O}_j(u(\cdot,\boldsymbol{y}))]_{j=1}^k$ with $\|\partial^{\boldsymbol{\nu}} G(\cdot,\boldsymbol{y})\|_{\mathbb{R}^k} \leq C|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}}$, where $C := \left(\sum_{j=1}^k \|\mathcal{O}_j\|_{H^{-1}(D)}^2\right)^{1/2}$ and $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$.

We are interested in the CM estimate

$$\boldsymbol{y}_{\mathrm{CM}} = \frac{1}{Z}\boldsymbol{Z'},$$

where

$$\boldsymbol{Z'} = \int_U \boldsymbol{y}\, \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))}\, \mathrm{d}\boldsymbol{y},$$

$$Z = \int_U \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))}\, \mathrm{d}\boldsymbol{y}.$$

It can be shown that

$$|\partial^{\boldsymbol{\nu}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))}| \leq 3.82^k \cdot C^{|\boldsymbol{\nu}|} 2^{|\boldsymbol{\nu}|-1} \mu_{\min}^{-|\boldsymbol{\nu}|/2} |\boldsymbol{\nu}|! \boldsymbol{b}^{\boldsymbol{\nu}} \quad \text{for } \boldsymbol{\nu} \neq \mathbf{0}$$

and

$$\begin{aligned}
&\left| \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_{\mathfrak{u}}} y_\ell \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\delta}-G(\boldsymbol{y}))^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{\delta}-G(\boldsymbol{y}))} \right| \\
&\leq 3.82^k \cdot C^{|\mathfrak{u}|} 2^{|\mathfrak{u}|-1} \mu_{\min}^{-|\mathfrak{u}|/2} |\mathfrak{u}|! \left(1 + \frac{1}{b_s}\right) \prod_{j \in \mathfrak{u}} b_j
\end{aligned}$$

for $\varnothing \neq \mathfrak{u} \subseteq \{1:s\}$.

For the QMC approximation of both the numerator and the denominator, this suggests choosing the weights

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

where $\delta \in (0, 1/2]$ is arbitrary and

$$\beta_j = 2C\mu_{\min}^{-1/2} b_j, \quad j = 1, \ldots, s,$$

as inputs to the CBC algorithm. The QMC rule obtained in this way has

- a dimension-independent QMC convergence rate $\mathcal{O}(n^{\max\{-1/p+1/2, -1+\delta\}})$ for the denominator when the number of QMC nodes n is a prime power.
- a dimension-dependent QMC convergence rate $\mathcal{O}\big(\big(1 + \frac{1}{b_\ell}\big) n^{\max\{-1/p+1/2, -1+\delta\}}\big)$ for *the* ℓ^{th} *component of the vector* $\boldsymbol{Z}'$ when the number of QMC nodes n is a prime power. For example, if $b_s \propto s^{-2}$, then $(1 + \frac{1}{b_s}) \propto s^2$.

(Note that the constant in the error bounds also depends on k, the dimension of the measurement data!)

Numerical example

Let us consider the PDE problem

$$-\nabla \cdot (a(\boldsymbol{x}) \nabla u(\boldsymbol{x})) = x_1, \quad u|_{\partial D} = 0,$$

in the physical domain $D = (0,1)^2$, where the diffusion coefficient is assumed to be *unknown*.

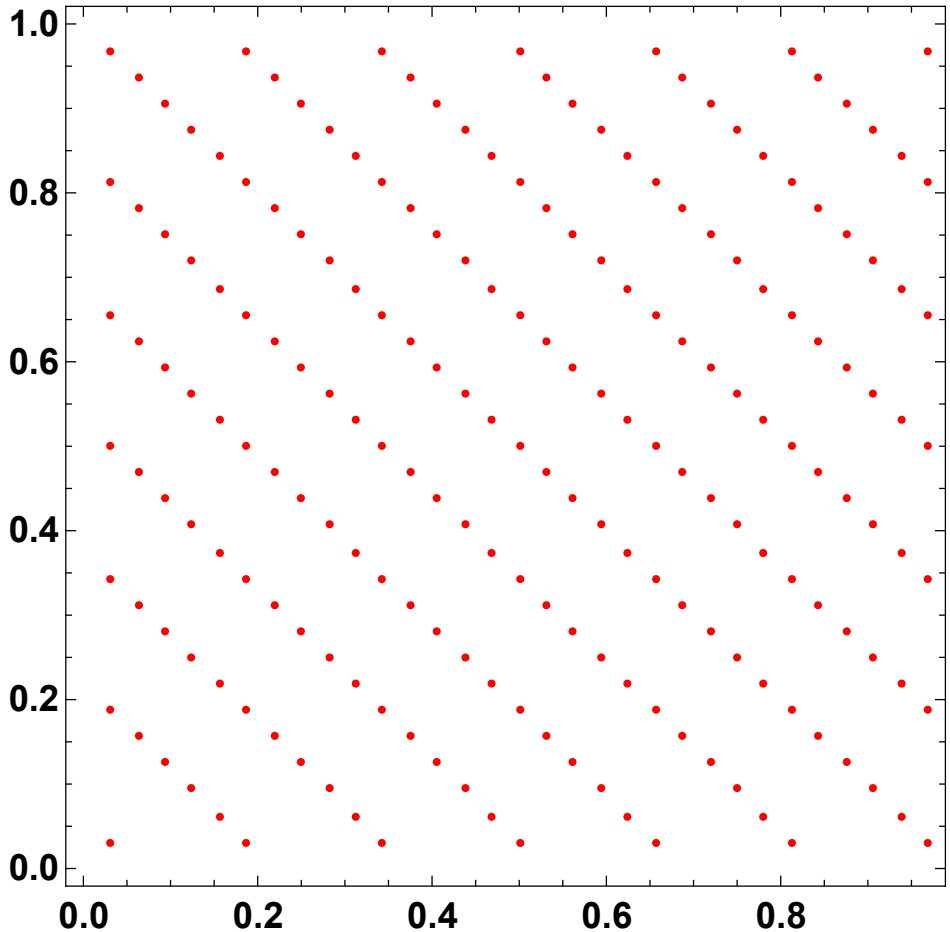
Given some noisy observations $\boldsymbol{y} = [u_1, \ldots, u_k]^{\mathrm{T}}$ of the PDE solution $G(u) = [\mathcal{O}_1(u), \ldots, \mathcal{O}_k(u)]^{\mathrm{T}} = [u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_k)]^{\mathrm{T}}$ over a point set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in D$, we wish to recover the diffusion coefficient which caused the observations.

We model the uncertain diffusion coefficient using the affine parameterization

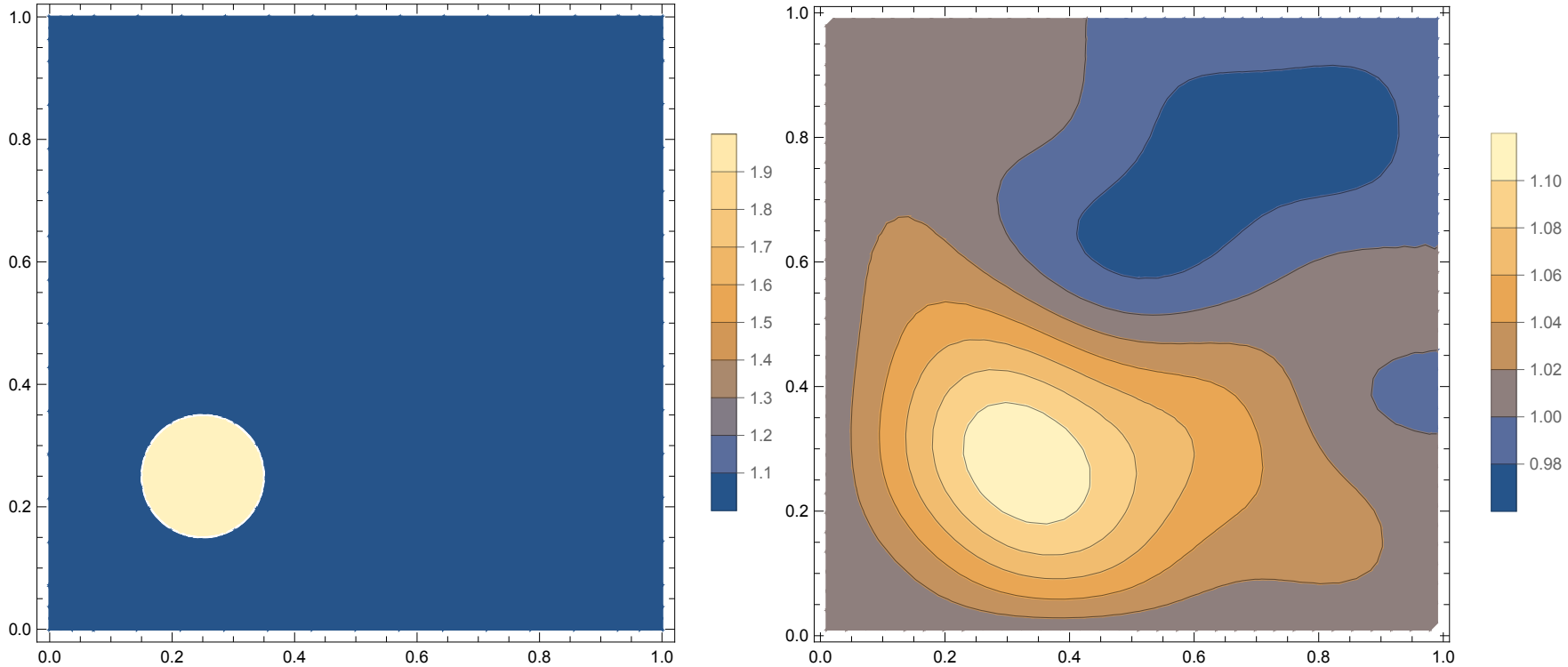
$$a(\boldsymbol{x}, \boldsymbol{y}) = 1 + \sum_{j=1}^{30} y_j \psi_j(\boldsymbol{x}), \quad y_j \in (-\tfrac{1}{2}, \tfrac{1}{2}),$$

where $\psi_j(\boldsymbol{x}) = \frac{1}{(k_j^2+\ell_j^2)^{1.1}} \sin(\pi k_j x_1) \sin(\pi \ell_j x_2)$ and the sequence $(k_j, \ell_j)_{j\geq 1}$ is an ordering of the elements of $\mathbb{N} \times \mathbb{N}$ so that the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ is non-increasing. This implies that $\|\psi_j\|_{L^\infty(D)} \sim j^{-1.1}$.

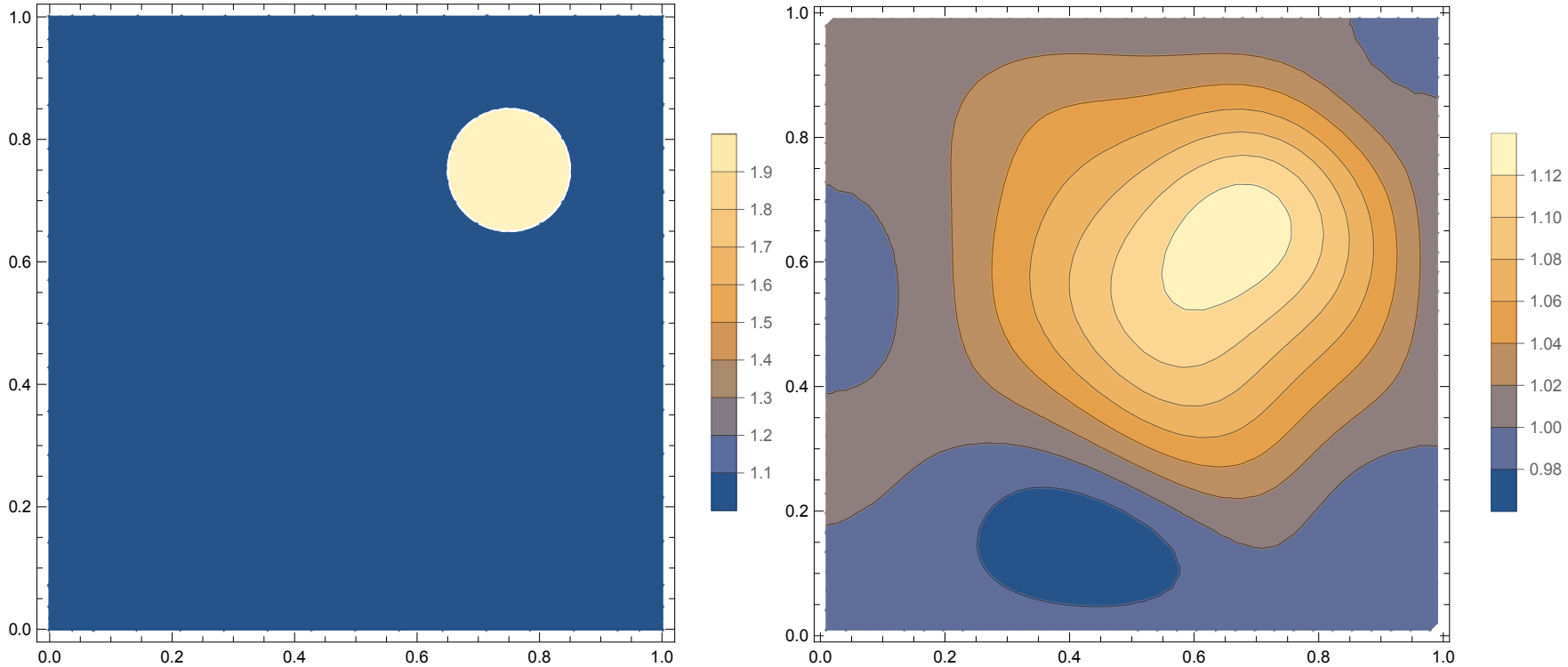
As the reconstruction, we compute the CM estimate $\boldsymbol{y}_{\mathrm{CM}} \in [-1/2, 1/2]^{30}$ of the parametric diffusion coefficient which fits the observations and plot $a(\boldsymbol{x}, \boldsymbol{y}_{\mathrm{CM}})$. The observations were simulated using a FE mesh with $h = 2^{-7}$ and contaminated with 1% relative white noise. The CM estimate was approximated using QMC with a single random shift and $n = 2^{14}$ nodes, and the PDE was discretized using a coarser FE mesh $h = 2^{-5}$ in order to avoid the so-called "inverse crime".



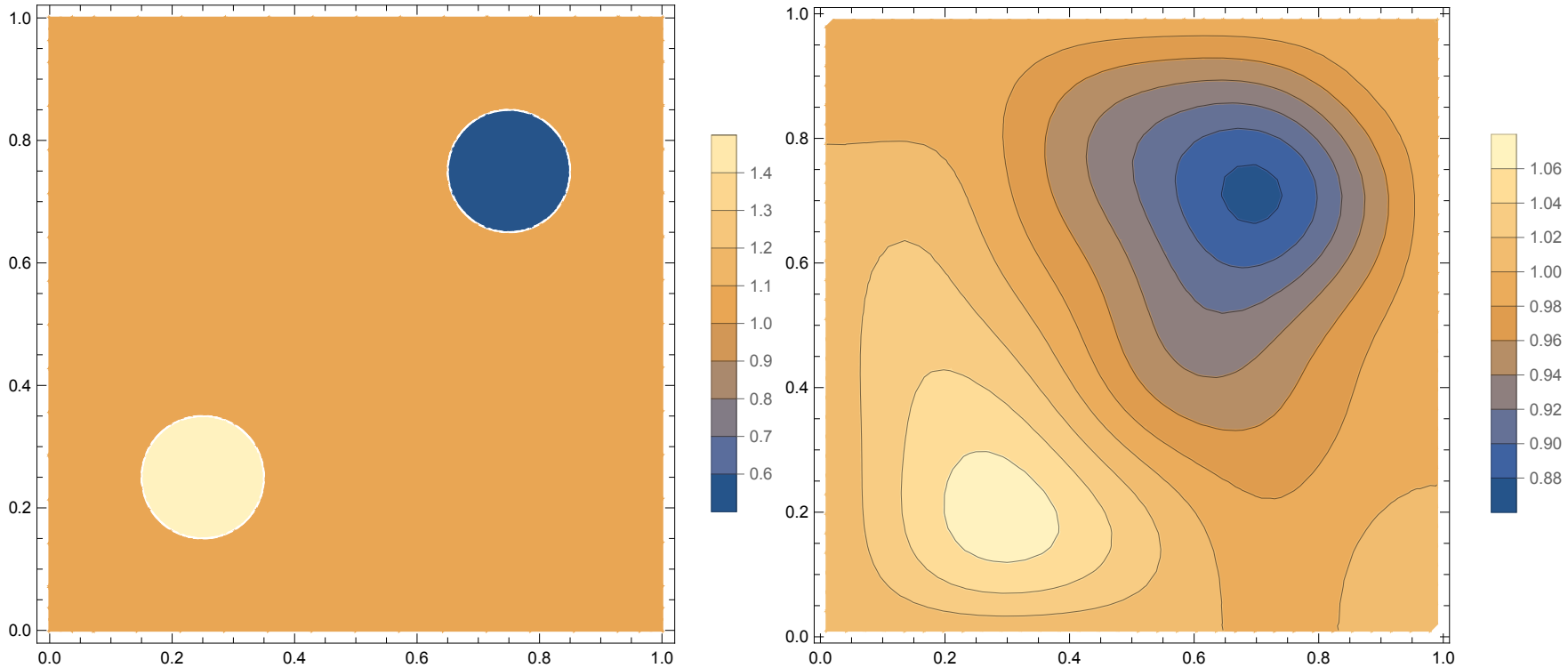
The point set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in D$ in the spatial domain where the noisy measurements $u_1, \ldots, u_k$ were collected with $k = 193$.



Left: the ground truth diffusion with $a_{\text{background}} \equiv 1$ and $a_{\text{inclusion}} \equiv 2$.
Right: the reconstruction $a(\boldsymbol{x}, \boldsymbol{y}_{\text{CM}})$.



Left: the ground truth diffusion with $a_{\text{background}} \equiv 1$ and $a_{\text{inclusion}} \equiv 2$.
Right: the reconstruction $a(\boldsymbol{x}, \boldsymbol{y}_{\text{CM}})$.



Left: the ground truth diffusion with $a_{\text{background}} \equiv 1$, $a_{\text{inclusion,bottom left}} \equiv 1.5$, and $a_{\text{inclusion,top right}} \equiv 0.5$.
Right: the reconstruction $a(\boldsymbol{x}, \boldsymbol{y}_{\text{CM}})$.