# Quasi-Monte Carlo methods for optimal control problems subject to parabolic PDE constraints under uncertainty

August 19, 2022

Vesa Kaarnioja (FU Berlin)
Joint work with Philipp Guth (Uni. Mannheim), Frances Kuo (UNSW Sydney), Claudia Schillings (FU Berlin), and Ian Sloan (UNSW Sydney)

# Table of contents

# Part I: Quasi-Monte Carlo methods

# High-dimensional numerical integration

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

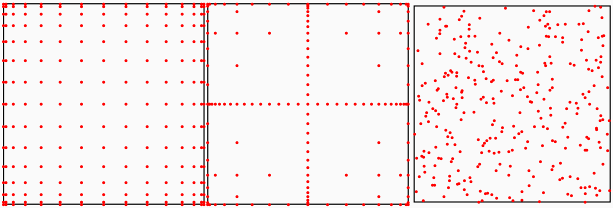


**Figure 1:** Tensor product grid, sparse grid, Monte Carlo nodes (not QMC rules)

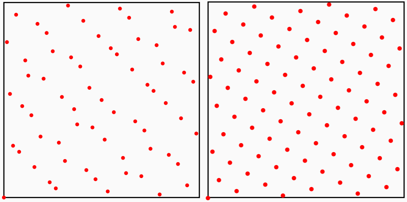


**Figure 2:** Sobol′ points, lattice rule (examples of QMC rules)

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

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

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.

# Lattice rules

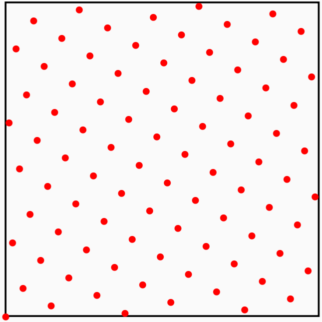
*Rank-1 lattice rules*

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

have the points

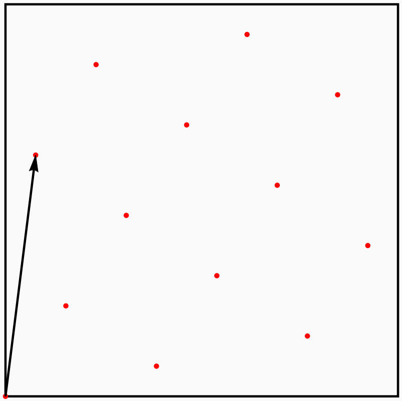
$$\boldsymbol{t}_i = \mathrm{mod}\left(\frac{i\boldsymbol{z}}{n}, 1\right), \quad i \in \{1, \ldots, n\},$$

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



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 $\boldsymbol{z}$.*



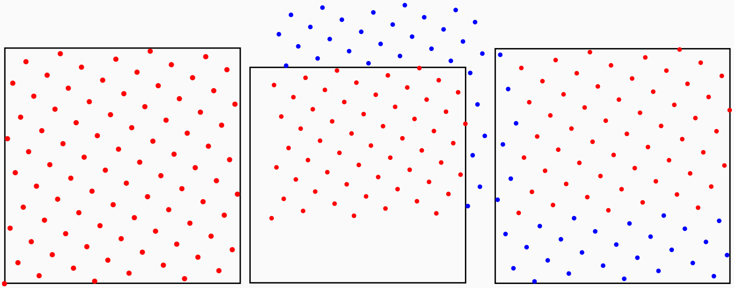
# Randomly shifted lattice rules

Shifted rank-1 lattice rules have points

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

$\Delta \in [0,1)^s$ is the *shift*

*Use a number of random shifts for error estimation.*



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

Let $\Delta^{(r)}$, $r = 1, \ldots, R$, be independent random shifts drawn from $U([0,1]^s)$ and define

$$Q_{s,n}^{(r)}(f) := \frac{1}{n} \sum_{i=1}^{n} f(\mathrm{mod}(\boldsymbol{t}_i + \Delta^{(r)}, 1)). \quad \text{(QMC rule with 1 random shift)}$$

Then

$$\overline{Q}_{s,n}(f) = \frac{1}{R} \sum_{r=1}^{R} Q_{s,n}^{(r)} f \qquad \text{(QMC rule with } R \text{ random shifts)}$$

is an unbiased estimator of $I_s(f)$.

Let $f : [0,1]^s \to \mathbb{R}$ be sufficiently smooth.

Error bound (one random shift):

$$|I_s(f) - Q_{s,n}^{\Delta}(f)| \le e_{s,n,\gamma}^{\Delta}(\boldsymbol{z}) \|f\|_{\gamma}.$$

R.M.S. error bound (shift-averaged):

$$\sqrt{\mathbb{E}_{\Delta}[|I_s(f) - \overline{Q}_{s,n}(f)|^2]} \le e_{s,n,\gamma}^{\mathrm{sh}}(\boldsymbol{z}) \|f\|_{\gamma}.$$

We consider weighted Sobolev spaces with dominating mixed smoothness, equipped with norm

$$\|f\|_{\gamma}^2 = \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[0,1]^{|\mathfrak{u}|}} \left( \int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} f}{\partial \boldsymbol{y}_{\mathfrak{u}}}(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}_{-\mathfrak{u}} \right)^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}}$$

and (squared) worst case error

$$P(\boldsymbol{z}) := e_{s,n,\gamma}^{\mathrm{sh}}(\boldsymbol{z})^2 = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\emptyset \ne \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} \omega\left( \left\{ \frac{k z_j}{n} \right\} \right)$$

where $\omega(x) = x^2 - x + \frac{1}{6}$.

# CBC algorithm (Sloan, Kuo, Joe 2002)

The idea of the *component-by-component* (CBC) algorithm is to find a good generating vector $\boldsymbol{z} = (z_1, \ldots, z_s)$ by proceeding as follows:

1. Set $z_1 = 1$ (this is a freebie since $P(1) = P(z)$ for all $z \in \mathbb{N}$);
2. With $z_1$ fixed, choose $z_2$ to minimize error criterion $P(z_1, z_2)$;
3. With $z_1$ and $z_2$ fixed, choose $z_3$ to minimize error criterion $P(z_1, z_2, z_3)$

   $\vdots$

Notes:

- The CBC algorithm is a *greedy algorithm*: in general, it will not find the generating vector $\boldsymbol{z}$ that minimizes $P(\boldsymbol{z})$. However, it can be shown that the generating vector obtained by the CBC algorithm satisfies an error bound (more on this later).
- For generic $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u} \subseteq \{1:s\}}$, evaluating $P(\boldsymbol{z}) = P(\boldsymbol{\gamma}, \boldsymbol{z})$ takes $\mathcal{O}(2^s)$ operations. For an efficient implementation, it is desirable that the weights $\boldsymbol{\gamma}$ can be characterized by an expression that does not contain too many degrees of freedom.

# CBC with POD weights

Suppose that we have QMC weights in *product and order dependent* (POD) form

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

for some positive scalars $(\Gamma_k)_{k \geq 1}$ and $(\gamma_j)_{j=1}^s$.

In this case, it turns out that the error criterion

$$P(\boldsymbol{z}) = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\varnothing \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}} \prod_{j \in \mathfrak{u}} \omega\left(\left\{\frac{kz_j}{n}\right\}\right)$$

can be written in the dimensionally recursive form

$$P(z_1, \ldots, z_s) = P(z_1, \ldots, z_{s-1}) + \frac{1}{n} \sum_{k=0}^{n-1} \sum_{\ell=1}^{s} p_{s,\ell}(k),$$

where $p_{s,\ell}(k) = p_{s-1,\ell}(k) + \frac{\Gamma_\ell}{\Gamma_{\ell-1}} \gamma_s \omega\left(\left\{\frac{kz_s}{n}\right\}\right) p_{s-1,\ell-1}(k)$ together with $p_{s,0}(k) = 1$ for all $k$.

For simplicity, let $n$ be prime. We note that the CBC algorithm can be implemented using the recurrence on the previous page, i.e.,

$$P(z_1,\ldots,z_d)=P(z_1,\ldots,z_{d-1})+\frac{1}{n}\sum_{k=0}^{n-1}\sum_{\ell=1}^{d}p_{d,\ell}(k),$$

where $p_{d,\ell}(k)=p_{d-1,\ell}(k)+\frac{\Gamma_\ell}{\Gamma_{\ell-1}}\gamma_d\omega\big(\big\{\frac{kz_d}{n}\big\}\big)p_{d-1,\ell-1}(k)$ together with $p_{d,0}(k)=1$ for all $k$, as follows:

1. Define the matrix $\Omega_n=\big[\omega\big(\big\{\frac{kz}{n}\big\}\big)\big]_{\substack{z\in\{1,\ldots,n-1\}\\ k\in\{0,\ldots,n-1\}}}$ and initialize vectors $\boldsymbol{p}_{0,\ell}=1_n$ for $\ell=1,\ldots,s$.

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

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

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

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

**end for**

# What makes fast CBC fast?

Note that 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 $\Omega_n$ can be permuted into circulant form $\rightarrow$ the matrix-vector product can be implemented in $\mathcal{O}(n \log n)$ operations using FFT.

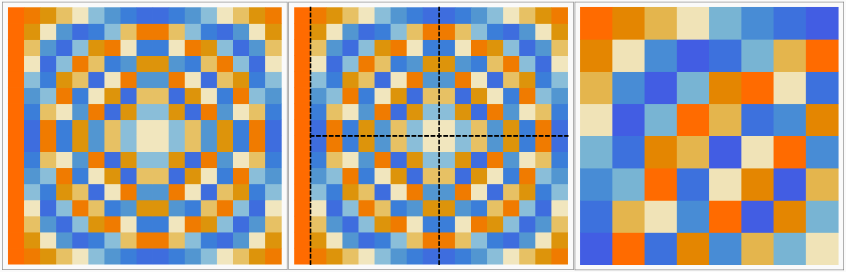


**Figure 3:** Example with $\Omega_{17}$. Note that the first column is a constant and can be left out (the components of $\Omega_n \boldsymbol{x}$ are shifted by a constant $\rightarrow$ the smallest component stays invariant). Noting the obvious symmetries in the remaining four blocks, we can focus on the top left block.

When $n$ is prime, it is possible to use the so-called Rader transformation to permute the block matrices into circulant form. (The permutation matrices can be easily generated by computing the "generator", i.e., primitive root modulo $n$.)

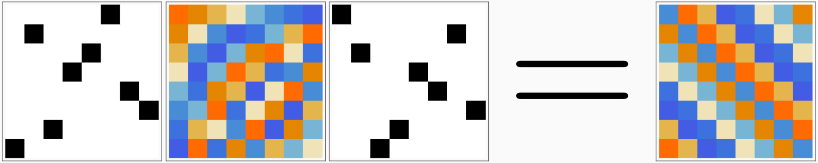


**Figure 4:** The original block matrix is multiplied from both sides by Rader permutation matrices (the black elements indicate the value 1 and white elements indicate the value 0) to obtain a circulant matrix.

# Example with $n = 1009$

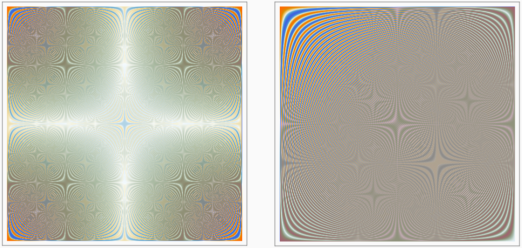


**Figure 5:** LHS: Original $\Omega_{1009}$. RHS: top left block of $\Omega_{1009}$ (sans first column).

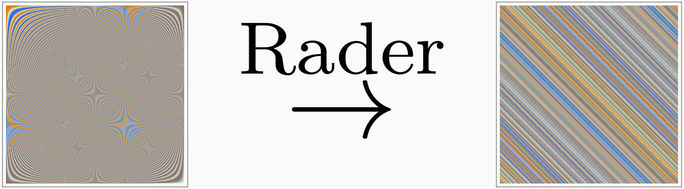



**Figure 6:** Rader transformation turns the top left block matrix circulant.

Remarks:

- Optimal rate of convergence $\mathcal{O}(n^{-1+\delta})$ in weighted Sobolev spaces, independently of $s$ under an appropriate condition on the weights.
- Cost of algorithm for POD weights is $\mathcal{O}(s\, n \log n + s^2\, n)$ using FFT.
- CBC works for any (composite) number $n \geq 2$, but the implementation is more involved when $n$ is not prime.

Application of QMC theory:

$$\sqrt{\mathbb{E}_\Delta[|I_s(f) - \overline{Q}_{s,n}(f)|^2]} \leq e^{\mathrm{sh}}_{s,n,\boldsymbol{\gamma}}(\boldsymbol{z})\|f\|_{\boldsymbol{\gamma}}.$$

- Estimate the norm (critical step)
- Choose the weights
- Weights as input to the CBC construction

# Part II: Optimal control problems subject to parabolic PDE constraints under uncertainty

# Optimal control

- Goal:
  - steer $u$ toward target $g$

$$\min_u \left( \frac{\alpha_1}{2} \int_0^T \|u(\cdot, t) - g(\cdot, t)\|^2 \,\mathrm{d}t + \frac{\alpha_2}{2} \|u(\cdot, T) - g(\cdot, T)\|^2 \right)$$

  - control $u$ through the source $z$ via the PDE

$$\partial_t u - \nabla \cdot (a \nabla u) = z \quad \text{in } D \times (0, T)$$

- Suppose $a^{\boldsymbol{y}}$ *uncertain*, $\boldsymbol{y}$ random parameter. Optimal control problem $\min_z J(z)$,

$$J(z) = \mathcal{R}\left( \frac{\alpha_1}{2} \int_0^T \|u(\cdot, t) - g(\cdot, t)\|^2 \,\mathrm{d}t + \frac{\alpha_2}{2} \|u(\cdot, T) - g(\cdot, T)\|^2 \right) + \frac{\alpha_3}{2} \int_0^T \|z(\cdot, t)\|^2 \,\mathrm{d}t$$

subject to

$$\partial_t u - \nabla \cdot (a^{\boldsymbol{y}} \nabla u) = z \quad \text{in } D \times (0, T)$$
$$+ \text{ initial-boundary values at } t = 0 \text{ and on } \partial D$$

- How do we handle uncertainty in $a^{\boldsymbol{y}}(\boldsymbol{x})$, parameterized by $\boldsymbol{y} \in U$?
- $\mathcal{R}$ risk measure
  - Expected value risk measure: $\mathcal{R}(\cdot) = \int_U \cdot \,\mathrm{d}\boldsymbol{y}$ (risk neutral)
  - Entropic risk measure: $\mathcal{R}(\cdot) = \frac{1}{\theta} \ln\left( \int_U \exp(\theta\, \cdot) \,\mathrm{d}\boldsymbol{y} \right)$ (risk averse)

# Optimal control problems

[Borzì, Schulz, Schillings, Winckel (2010); Ali, Ullmann, Hinze (2017)]

- **Mean-based control**: replace $a$ by $\int_U a^{\boldsymbol{y}} \, \mathrm{d}\boldsymbol{y}$
- **Pathwise control**: for several $\boldsymbol{y}$, find

$$z^*(\boldsymbol{y}) = \arg\min \left( \frac{\alpha_1}{2} \int_0^T \|u^{\boldsymbol{y}}(\cdot, t) - g(\cdot, t)\|^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|u^{\boldsymbol{y}}(\cdot, T) - g(\cdot, T)\|^2 \right) + \frac{\alpha_3}{2} \int_0^T \|z(\cdot, t)\|^2 \, \mathrm{d}t$$

  and compute the statistics of $z^*(\boldsymbol{y})$, e.g., $\mathbb{E}[z^*(\cdot)]$
- **Stochastic control**:

$$J(z) = \mathcal{R}\left( \frac{\alpha_1}{2} \int_0^T \|u^{\boldsymbol{y}}(\cdot, t) - g(\cdot, t)\|^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|u^{\boldsymbol{y}}(\cdot, T) - g(\cdot, T)\|^2 + \frac{\alpha_3}{2} \int_0^T \|z^{\boldsymbol{y}}(\cdot, t)\|^2 \, \mathrm{d}t \right)$$

  [Kouri et al. (2013); Kunoth, Schwab (2013); Ali, Ullmann, Hinze (2017); Chen, Ghattas (2018)]
- **Deterministic control**:

$$J(z) = \mathcal{R}\left( \frac{\alpha_1}{2} \int_0^T \|u^{\boldsymbol{y}}(\cdot, t) - g(\cdot, t)\|^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|u^{\boldsymbol{y}}(\cdot, T) - g(\cdot, T)\|^2 \right) + \frac{\alpha_3}{2} \int_0^T \|z(\cdot, t)\|^2 \, \mathrm{d}t$$

  • Taylor approximations [Chen, Villa, Ghattas (2019)], Sparse grids [Kouri (2014); Kouri Surowiec (2016)], Multilevel Monte Carlo [Van Barel, Vandewalle (2019)]

# Problem formulation

- Optimal control problem $\min_{z\in\mathcal{Z}} J(z)$, where

$$J(z) = \mathcal{R}\left(\frac{\alpha_1}{2}\int_0^T \|u^{\boldsymbol{y}}(\cdot,t)-g(\cdot,t)\|^2\,\mathrm{d}t + \frac{\alpha_2}{2}\|u^{\boldsymbol{y}}(\cdot,T)-g(\cdot,T)\|^2\right) + \frac{\alpha_3}{2}\int_0^T \|z(\cdot,t)\|^2\,\mathrm{d}t$$

  subject to

$$\begin{aligned} \partial_t u^{\boldsymbol{y}}(\boldsymbol{x},t) - \nabla\cdot(a^{\boldsymbol{y}}(\boldsymbol{x})\nabla u^{\boldsymbol{y}}(\boldsymbol{x},t)) &= z(\boldsymbol{x},t) && \boldsymbol{x}\in D,\ t\in(0,T),\ \boldsymbol{y}\in U,\\ u^{\boldsymbol{y}}(\boldsymbol{x},t) &= 0 && \boldsymbol{x}\in\partial D,\ t\in(0,T),\ \boldsymbol{y}\in U,\\ u^{\boldsymbol{y}}(\boldsymbol{x},0) &= u_0(\boldsymbol{x}) && \boldsymbol{x}\in D,\ \boldsymbol{y}\in U, \end{aligned}$$

- Look for optimal control $z$ in feasible set $\mathcal{Z}$, where we consider

$$\mathcal{Z} := L^2(D\times(0,T)) \qquad \text{(unconstrained)}$$

$$\mathcal{Z} := \{z\in L^2(D\times(0,T)) \mid z_{\min}(\boldsymbol{x},t)\le z(\boldsymbol{x},t)\le z_{\max}(\boldsymbol{x},t)\ a.e.\} \qquad \text{(constrained)}$$

- "**Uniform**" or "**affine**" model:

$$a^{\boldsymbol{y}}(\boldsymbol{x}) = \bar{a}(\boldsymbol{x}) + \sum_{j=1}^{\infty} y_j\psi_j(\boldsymbol{x}), \quad \boldsymbol{x}\in D,\ \boldsymbol{y}\in U := [-\tfrac{1}{2},\tfrac{1}{2}]^{\mathbb{N}}$$

  with assumptions

  - $0 < a_{\min} \le a^{\boldsymbol{y}}(\boldsymbol{x}) \le a_{\max} < \infty$
  - $\sum_{j=1}^{\infty}\|\psi_j\|_{L^\infty}^p < \infty$ for some $p\in(0,1)$
  - $\|\psi_1\|_{L^\infty} \ge \|\psi_2\|_{L^\infty} \ge \cdots$

# Reduced formulation

- Optimal control problem $\min_{z \in \mathcal{Z}} J(z)$, with

$$J(z) = \underbrace{\mathcal{R}\left( \frac{\alpha_1}{2} \int_0^T \|u^{\boldsymbol{y}}(\cdot, t) - g(\cdot, t)\|^2 \,\mathrm{d}t + \frac{\alpha_2}{2} \|u^{\boldsymbol{y}}(\cdot, T) - g(\cdot, T)\|^2 \right)}_{=:\Phi(\boldsymbol{y})} + \frac{\alpha_3}{2} \int_0^T \|z(\cdot, t)\|^2 \,\mathrm{d}t$$

- $u^{\boldsymbol{y}}$ is the "state" corresponding to the control $z$
- $\mathcal{Z}$ is the feasible set, which is bounded, closed, convex, nonempty
- **Theorem**: *There exists a unique optimal solution $z^*$.*
- The Fréchet derivative of $J$ is given by

$$J'(z) = \mathcal{R}(q) + \alpha_3 z \qquad \text{(linear risk measure } \mathcal{R}\text{)}$$

$$J'(z) = \frac{1}{\int_U \exp(\theta\Phi(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}} \int_U \exp(\theta\Phi(\boldsymbol{y}))\, q^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} + \alpha_3 z \qquad \text{(entropic risk measure)}$$

where $q$ is the adjoint state.

**Theorem**: Let $\mathcal{R}$ be the expected value or the entropic risk measure. A control $z = z^*$ is the unique minimizer of the problem if and only if it satisfies the KKT system

$$
\begin{cases}
\partial_t u^{\boldsymbol{y}}(\boldsymbol{x},t) - \nabla \cdot (a^{\boldsymbol{y}}(\boldsymbol{x}) \nabla u^{\boldsymbol{y}}(\boldsymbol{x},t)) = z(\boldsymbol{x}) & \boldsymbol{x} \in D,\ t \in (0,T), \\
u^{\boldsymbol{y}}(\boldsymbol{x},t) = 0 & \boldsymbol{x} \in \partial D,\ t \in (0,T), \\
u^{\boldsymbol{y}}(\boldsymbol{x},0) = u_0(\boldsymbol{x}) & \boldsymbol{x} \in D, \\
-\partial_t q^{\boldsymbol{y}}(\boldsymbol{x},t) - \nabla \cdot (a^{\boldsymbol{y}}(\boldsymbol{x}) \nabla q^{\boldsymbol{y}}(\boldsymbol{x},t)) & \\
\qquad = \alpha_1 (u^{\boldsymbol{y}}(\boldsymbol{x},t) - g(\boldsymbol{x},t)) & \boldsymbol{x} \in D,\ t \in (0,T), \\
q^{\boldsymbol{y}}(\boldsymbol{x},t) = 0 & \boldsymbol{x} \in \partial D,\ t \in (0,T), \\
q^{\boldsymbol{y}}(\boldsymbol{x},T) = \alpha_2 (u^{\boldsymbol{y}}(\boldsymbol{x},T) - g(\boldsymbol{x},T)) & \boldsymbol{x} \in D, \\
J'(z^*) - \mu_a + \mu_b = 0 & \boldsymbol{x} \in D,\ t \in (0,T), \\
z_{\min}(\boldsymbol{x},t) \leq z(\boldsymbol{x},t) \leq z_{\max}(\boldsymbol{x},t) & \boldsymbol{x} \in D,\ t \in (0,T) \\
\mu_a(\boldsymbol{x},t) \geq 0,\ \mu_b(\boldsymbol{x},t) \geq 0 & \boldsymbol{x} \in D,\ t \in (0,T) \\
(z(\boldsymbol{x},t) - z_{\min}(\boldsymbol{x},t)) \mu_a(\boldsymbol{x},t) & \\
\qquad = (z_{\max}(\boldsymbol{x},t) - z(\boldsymbol{x},t)) \mu_b(\boldsymbol{x},t) = 0 & \boldsymbol{x} \in D,\ t \in (0,T),
\end{cases}
$$

for all $\boldsymbol{y} \in U$. Can be solved, e.g., using projected gradient descent.

# Discretization of the problem (expected value risk measure $\mathcal{R} = \int_U \cdot \, \mathrm{d}\boldsymbol{y}$)

- Optimal control problem:

$$J(z) = \int_U \left( \frac{\alpha_1}{2} \int_0^T \|u^{\boldsymbol{y}}(\cdot,t) - g(\cdot,t)\|^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|u^{\boldsymbol{y}}(\cdot,T) - g(\cdot,T)\|^2 \right) \mathrm{d}\boldsymbol{y} + \frac{\alpha_3}{2} \int_0^T \|z(\cdot,t)\|^2 \, \mathrm{d}t$$

$$J'(z) = \int_U q^{\boldsymbol{y}} \, \mathrm{d}\boldsymbol{y} + \alpha_3 z$$

- Discretized problem:
  - Dimension truncation to $s$ terms
  - Quasi-Monte Carlo quadrature with $n$ points

$$J_{s,n}(z) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\alpha_1}{2} \int_0^T \|u_s^{\boldsymbol{y}^{(i)}}(\cdot,t) - g(\cdot,t)\|^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|u_s^{\boldsymbol{y}^{(i)}}(\cdot,T) - g(\cdot,T)\|^2 \right) + \frac{\alpha_3}{2} \int_0^T \|z(\cdot,t)\|^2 \, \mathrm{d}t$$

$$J'_{s,n}(z) = \frac{1}{n} \sum_{i=1}^{n} q_s^{\boldsymbol{y}^{(i)}} + \alpha_3 z$$

  - Positive (equal) weights of QMC preserve convexity of the problem

# Discretization of the problem (entropic risk measure)

- Optimal control problem:

$$J(z) = \frac{1}{\theta} \ln \left( \int_U \exp(\theta \Phi(\boldsymbol{y})) q^{\boldsymbol{y}} \, \mathrm{d}\boldsymbol{y} \right) + \frac{\alpha_3}{2} \int_0^T \|z(\cdot, t)\|^2 \, \mathrm{d}t$$

$$J'(z) = \frac{\int_U \exp(\theta \Phi(\boldsymbol{y})) q^{\boldsymbol{y}} \, \mathrm{d}\boldsymbol{y}}{\int_U \exp(\theta \Phi(\boldsymbol{y})) \, \mathrm{d}\boldsymbol{y}} + \alpha_3 z$$

$$\Phi(\boldsymbol{y}) = \frac{\alpha_1}{2} \int_0^T \|u^{\boldsymbol{y}}(\cdot, t) - g(\cdot, t)\|^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|u^{\boldsymbol{y}}(\cdot, T) - g(\cdot, T)\|^2$$

- Discretized problem:
  - Dimension truncation to $s$ terms
  - Quasi-Monte Carlo quadrature with $n$ points

$$J_{s,n}(z) = \frac{1}{\theta} \ln \left( \frac{1}{n} \sum_{i=1}^{n} \exp(\theta \Phi_s(\boldsymbol{y}^{(i)})) q_s^{\boldsymbol{y}^{(i)}} \right) + \frac{\alpha_3}{2} \int_0^T \|z(\cdot, t)\|^2 \, \mathrm{d}t$$

$$J'_{s,n}(z) = \frac{\sum_{i=1}^{n} \exp(\theta \Phi_s(\boldsymbol{y}^{(i)}) q_s^{\boldsymbol{y}^{(i)}})}{\sum_{i=1}^{n} \exp(\theta \Phi_s(\boldsymbol{y}^{(i)}))} + \alpha_3 z$$

# Error analysis

- **Theorem (expected value risk measure)** A randomly shifted lattice rule can be constructed using the fast CBC algorithm using POD weights such that

$$\sqrt{\mathbb{E}_\Delta \|z^* - z^*_{s,n}\|^2_{L^2(D\times(0,T))}} = \mathcal{O}(s^{-2/p+1} + n^{-\min\{1/p-1/2,1-\delta\}}),$$

where the implied coefficient is independent of $s$.

- **Theorem (entropic risk measure)** A randomly shifted lattice rule can be constructed using the fast CBC algorithm using POD weights such that

$$\sqrt{\mathbb{E}_\Delta \|z^* - z^*_{s,n}\|^2_{L^2(D\times(0,T))}} = \mathcal{O}(s^{-2/p+1} + n^{-\min\{1/p-1/2,1-\delta\}}),$$

where the implied coefficient is independent of $s$.

# Numerical experiments

Consider the coupled PDE system

$$\begin{cases} \partial_t u^{\boldsymbol{y}}(\boldsymbol{x},t) - \nabla\cdot(a^{\boldsymbol{y}}(\boldsymbol{x})\nabla u^{\boldsymbol{y}}(\boldsymbol{x},t)) = z(\boldsymbol{x}) & \boldsymbol{x}\in D\,, t\in(0,T)\,,\\ u^{\boldsymbol{y}}(\boldsymbol{x},t) = 0 & \boldsymbol{x}\in\partial D\,, t\in(0,T)\,,\\ u^{\boldsymbol{y}}(\boldsymbol{x},0) = u_0(\boldsymbol{x}) & \boldsymbol{x}\in D\,,\\ -\partial_t q^{\boldsymbol{y}}(\boldsymbol{x},t) - \nabla\cdot(a^{\boldsymbol{y}}(\boldsymbol{x})\nabla q^{\boldsymbol{y}}(\boldsymbol{x},t)) & \\ \qquad\qquad = \alpha_1(u^{\boldsymbol{y}}(\boldsymbol{x},t) - g(\boldsymbol{x})) & \boldsymbol{x}\in D\,, t\in(0,T),\\ q^{\boldsymbol{y}}(\boldsymbol{x},t) = 0 & \boldsymbol{x}\in\partial D\,, t\in(0,T)\\ q^{\boldsymbol{y}}(\boldsymbol{x},T) = \alpha_2(u^{\boldsymbol{y}}(\boldsymbol{x},T) - g(\boldsymbol{x})) & \boldsymbol{x}\in D\,, t\in(0,T), \end{cases}$$

for $\boldsymbol{y}\in[-\frac{1}{2},\frac{1}{2}]^s$ in $D=(0,1)^2$, equipped with

$$a(\boldsymbol{x},\boldsymbol{y}) = 1 + c\sum_{j=1}^{\infty} y_j \underbrace{\frac{1}{(k_j^2+\ell_j^2)^{\vartheta}}\sin(k_j\pi x_1)\sin(\ell_j\pi x_2)}_{=\psi_j}\,,$$

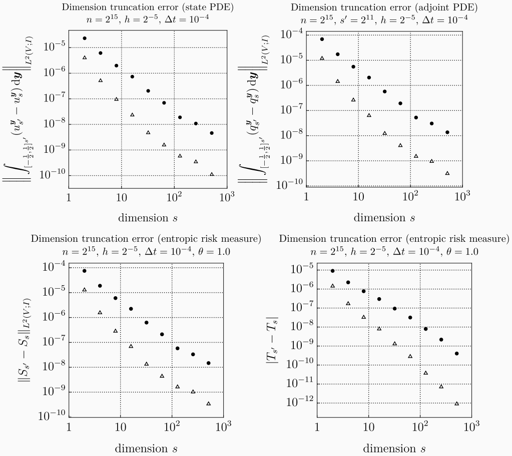
where $(k_j,\ell_j)_{j\geq 1}$ are ordered s.t. $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ is non-increasing.

Weyl asymptotics: $\|\psi_j\|_{L^\infty(D)} \sim j^{-\vartheta}$ as $j\to\infty$. We fix $z(\boldsymbol{x}) = x_2$.

# Numerical experiments

## Dimension truncation error

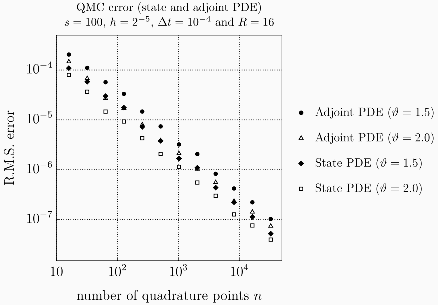
We use $s = 2^{11}$ as the reference solution.

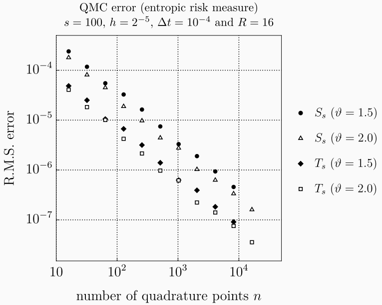


# Numerical experiments

## QMC error

We use $R = 16$ random shifts $\Delta$ drawn from $U([0,1]^s)$.




LHS: QMC errors corresponding to $\int_{[-\frac{1}{2},\frac{1}{2}]^{100}} u^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y}$ (state PDE) and $\int_{[-\frac{1}{2},\frac{1}{2}]^{100}} q^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y}$ (adjoint PDE).

RHS: QMC errors corresponding to $S_{100} = \int_{[-\frac{1}{2},\frac{1}{2}]^{100}} \exp(\Phi_s(\boldsymbol{y}))q^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y}$ and
$T_{100} = \int_{[-\frac{1}{2},\frac{1}{2}]^{100}} \exp(\Phi_s(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y}$

# Numerical experiments

Optimal control problem – unconstrained gradient descent $\min_{z\in L^2(D\times(0,T))} J(z)$,

$$J(z,u)=\mathcal{R}_{\boldsymbol{y}}\bigg(\frac{\alpha_1}{2}\int_0^T \|u^{\boldsymbol{y}}(\cdot,t)-g\|^2_{L^2(D\times(0,T))}\,\mathrm{d}t+\frac{\alpha_2}{2}\|u^{\boldsymbol{y}}(\cdot,T)-g(\cdot,T)\|^2_{L^2(D)}\bigg)+\frac{\alpha_3}{2}\|z\|^2_{L^2(D\times(0,T))}$$

We fix $\vartheta=1.2$, $h=2^{-4}$, $\Delta t=10^{-2}$, $n=2^{10}$, $s=100$, $T=1.0$, $u_0(\boldsymbol{x})=\sin(2\pi x_1)\sin(2\pi x_2)$, $\alpha_1=1$, $\alpha_2=\alpha_3=10^{-4}$, and $\theta=10^5$. As the diffusion coefficient, we take

$$a(\boldsymbol{x},\boldsymbol{y})=100+55\sum_{j=1}^{100} y_j j^{-1.2}\sin(j\pi x_1)\sin(j\pi x_2).$$

As the target, we choose

$$g(\boldsymbol{x},t):=\chi_{\|\boldsymbol{x}-(c_1(t),c_2(t))\|_\infty\le\frac{1}{10}}(\boldsymbol{x})g_1(\boldsymbol{x},t)+\chi_{\|\boldsymbol{x}+(c_1(t),c_2(t))-(1,1)\|_\infty\le\frac{1}{10}}(\boldsymbol{x})g_2(\boldsymbol{x},t),$$

where $g_1(\boldsymbol{x},t):=(4\sqrt{10})^4(x_1-c_1(t)-\frac{1}{10})(x_2-c_2(t)-\frac{1}{10})(x_1-c_1(t)+\frac{1}{10})(x_2-c_2(t)+\frac{1}{10})$,
$g_2(\boldsymbol{x},t):=(4\sqrt{10})^4(x_1+c_1(t)-\frac{11}{10})(x_2+c_2(t)-\frac{11}{10})(x_1+c_1(t)-\frac{9}{10})(x_2+c_2(t)-\frac{9}{10})$,
$c_1(t):=\frac{1}{2}+\frac{1}{4}(1-t^{10})\cos(4\pi t^2)$, and $c_2(t):=\frac{1}{2}+\frac{1}{4}(1-t^{10})\sin(4\pi t^2)$.

# Unconstrained optimization

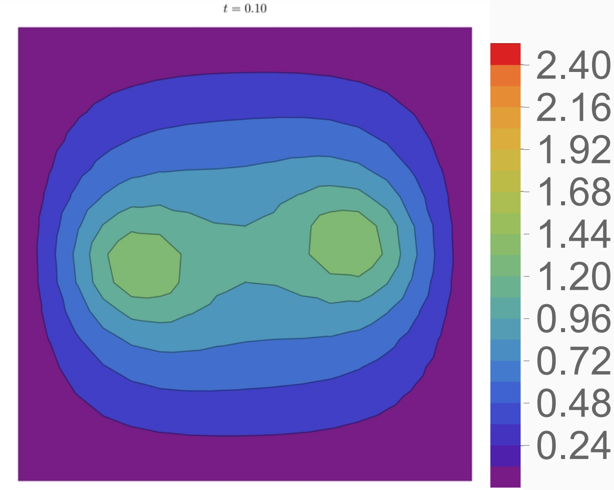



**Figure 7:** The reconstructed optimal control $z^*$ (without box constraints).

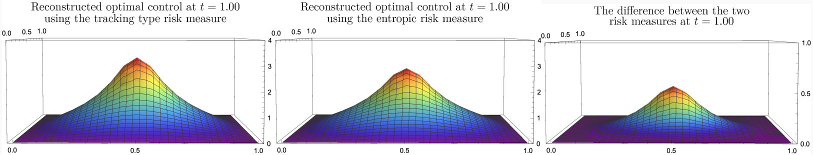



**Figure 8:** Left: the reconstructed optimal control using the tracking type risk measure at time $t = 1.00$. Middle: the reconstructed optimal control using the entropic risk measure at time $t = 1.00$. Right: The difference between the two reconstructed optimal controls.

# Conclusions

- We applied QMC to an optimal control problem with time-dependent PDE constraints under uncertainty. The QMC discretization preserves convexity, unlike sparse grids (which have negative weights).
- QMC approximations of high-dimensional integrals converge faster than MC.

References:


Guth, K., Kuo, Schillings, and Sloan. *Parabolic PDE-constrained optimal control under uncertainty with entropic risk measure using quasi-Monte Carlo integration*. Preprint 2022. arXiv:2208.02767

Guth, K., Kuo, Schillings, and Sloan. *A quasi-Monte Carlo method for optimal control under uncertainty*. SIAM/ASA Journal on Uncertainty Quantification **9**(2), 354–383, 2021.