Higher order QMC integration with scrambling for elliptic PDEs with random coefficients

Marcello Longo

MSc ETH in Mathematics

Supervisor: Prof. Christoph Schwab

Submitted: April 23, 2019
Revised: June 11, 2019

Introduction

Quasi-Monte Carlo (QMC) integration is the deterministic counterpart of Monte Carlo and it has been recently applied to the field of uncertainty quantification to approximate statistics of PDE solutions, with promising results (see for example [21, 9, 17] and the references mentioned therein). The first goal of this Thesis is to introduce the main concepts and tools to describe the error decay of some high order QMC integration rules, with respect to the number N of function evaluations. In particular, we will analyse the *interlaced scrambled polynomial lattice point sets* constructed in [13] and its component-by-component (CBC) construction, that does not suffer from the *curse of dimensionality* as classical integration rules. Next, we will apply these to some elliptic PDEs defined on a bounded interval or a polygon $D \subset \mathbb{R}^2$, where we allow countably many uncertain parameters; we aim at obtaining high order approximation rates independent of the parametric dimension of the problem. We will assume that the functions of the affine parametric expansion are locally supported: in [12, 11], it has been shown that this hypothesis yields dependence on the parameters that can be modeled by product weights, that is, positive coefficients $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1,2,\dots,s\}}$ in the form $\gamma_{\mathfrak{u}} := \prod_{j\in\mathfrak{u}} \gamma_j$ for a sequence $(\gamma_j)_{j\geq 1}$. The product weights can be exploited to generate a suitable QMC rule using the fast CBC construction in [13], that makes use of FFT. In order to reduce the overall computational cost of coupled QMC-FEM, the technique of Multi-Level QMC integration is studied, for example, in [22, 11, 15, 16]. Here, we will present a new result to bound the error in the Multi-Level case, which benefits from both high order QMC rules and high order FEM.

In Chapter 1 we review the preliminary definitions and results on some Higher Order QMC rules, in particular regarding the randomisation procedure of scrambling and the digit interlacing of points, that is the key to increase the order of the QMC integration rule. An important class of weighted Sobolev spaces will be introduced together with the Walsh decomposition and the ANOVA decomposition of a multivariate function.

In Chapter 2 we show the convergence of *interlaced scrambled polynomial lattice rules* to the exact integral, with order dependent on the smoothness of the integrand but independent of the dimension of the integration domain. Moreover, the proof is constructive in the sense that an explicit component-by-component algorithm can be applied to achieve said order of convergence.

Chapter 3 deals with the approximate evaluation of *Quantities of Interest* (QoI) of affine parametric, elliptic PDEs. We will identify three main sources of error: a *dimension truncation* error, a *Galerkin* error and a *QMC* error. All these sources will be discussed in detail. Next, we extend the analysis to Multi-Level QMC integration, which results in a reduction of the computational effort

needed to obtain approximations under some a priori error tolerance.

In Chapter 4 we outline some implementation aspects regarding the construction of the QMC integration points studied in the previous chapters, in particular the fast CBC algorithm for product weights. Finally, we present various numerical experiments and describe some observations on their results.

Basic Notation

Throughout we denote by $\mathbb{N} = \{1, 2, \ldots\}$ the set of natural numbers and we will write explicitly $\mathbb{N}_0$ to include 0. For any $n \in \mathbb{N}$ we write $\{1:n\} := \{1, 2, \ldots, n\}$ and $\{1:n\}^c := \mathbb{N}\backslash\{1:n\}$. For all $\mathfrak{u} \subseteq \{1:n\}$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, we write $\boldsymbol{y}_\mathfrak{u}$ to denote the subvector of $\boldsymbol{y}$ containing only the components in $\mathfrak{u}$. For a set $\Omega \in \mathbb{R}^n$, define $\partial\Omega$ its boundary and $f\big|_\Omega$ the restriction of a function f on Ω. Moreover, χ_Ω denotes the indicator function, i.e. $\chi_\Omega(x) = 1$ for all $x \in \Omega$ and $\chi_\Omega(x) = 0$ else; $\mathcal{B}(\Omega)$ and $\mathcal{L}(\Omega)$ denote the Borel and Lebesgue σ-algebra on Ω, respectively. For the Lebesgue measure on $\mathbb{R}^n$, we write $\mathrm{d}x$ or $\mathrm{d}\boldsymbol{x}$ when we want to emphasise the variable or, in certain cases, we write λ_n. We use the standard notation $L^p(\Omega)$ to denote the set of measurable functions f with $\int_\Omega |f(x)|^p \mathrm{d}x < \infty$, where functions agreeing λ_n-almost everywhere are identified. Similarly, we write $W^{m,p}(\Omega)$ for the Sobolev space of functions which weak derivatives up to (total) order m are in $L^p(\Omega)$. We often use the notation $H^m(\Omega) := W^{m,2}(\Omega)$ and $H_0^1(\Omega) := \{f \in H^1(\Omega) : f\big|_{\partial\Omega} = 0\}$, where $f\big|_{\partial\Omega} = 0$ is meant in the sense of traces.

AKNOWLEDGMENTS

I would like to thank my advisor Prof. Christoph Schwab and Dr. Lukas Herrmann for the valuable discussions and their constructive feedback on my work. I am also grateful to Prof. Josef Dick, for giving me a deep insight of his own research and to Prof. Takashi Goda, who made available a part of his software. My parents and my family deserve a special mention, for their unconditioned support (and patience) throughout my Swiss adventure. Finally, I would like to express my gratitude to Prof. Paolo Vasile, who played a decisive role in my decision to pursue the studies in Mathematics.

Contents

Chapter 1

Higher order QMC rules

In this chapter we will construct *interlaced scrambled polynomial lattice rules* using the algorithm described by Goda and Dick in [13], to evaluate numerically high-dimensional integrals. In particular, we consider the model problem of approximating the multivariate integral over a s-dimensional unit cube in (1.1). This will be done by sampling the integrand F on a set of nodes $P := \{\boldsymbol{y}_0, \dots, \boldsymbol{y}_{N-1}\}$ and averaging with uniform weights:

$$I_s(F) := \int_{[0,1]^s} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \approx \frac{1}{N} \sum_{i=0}^{N-1} F(\boldsymbol{y}_i). \tag{1.1}$$

One possibility is to use a Monte Carlo (MC) approximation, that relies on independent uniformly distributed nodes $\boldsymbol{y}_i$ in $[0,1)^s$. We denote by $I_s(F;P)$ the estimator on the right hand side of (1.1), which is well defined if and only if F is defined pointwise. The unit cube $[0,1]^s$ with its Lebesgue σ-algebra and the Lebesgue measure form a probability space. Then, uniform distribution of the samples and strong law of large numbers ensure that $\mathbb{E}[I_s(F;P)] = I_s(F)$, so that there holds, for $F \in L^2([0,1]^s) \cap C^0([0,1)^s)$,

$$\mathbb{E}\Big[\Big(I_s(F;P) - I_s(F)\Big)^2\Big] = \mathbb{E}\Big[\Big(I_s(F;P) - \mathbb{E}[I_s(F;P)]\Big)^2\Big] = \mathrm{Var}[I_s(F;P)]. \tag{1.2}$$

This in turn implies $\mathbb{E}\Big[\Big(I_s(F;P) - I_s(F)\Big)^2\Big] = \dfrac{1}{N^2} N \,\mathrm{Var}[F(\boldsymbol{y}_0)] = O(N^{-1})$. Equivalently, the convergence rate of the root mean squared error is only of $O(N^{-1/2})$, no matter how smooth F is. Said rate is not desirable, especially when each evaluation of F is very costly; this will be the case of our problem in later sections, since every $F(\boldsymbol{y}_i)$ will be the solution of a PDE, or a linear functional of such solution.

On the other hand, a Quasi-Monte Carlo (QMC in short) integration rule consists in a deterministic choice of the $\boldsymbol{y}_i$. This allows to exploit additional properties of an integrand like smoothness and decay of the derivatives. As a consequence, two problems need to be addressed to improve the convergence rate:

- find a suitable set of nodes $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{N-1}$;
- identify the set of functions F that realises the desired decay of the error.

The first point will be addressed in the Sections from 1.1 to 1.3 while the second will be discussed in Section 1.6.

Furthermore, the concepts of MC and QMC can be merged (see [8, 13]); the idea is to *randomise* a deterministic QMC rule in such a way that the structure and properties of the lattice are preserved. The result of this approach is a combination of the convergence rates of the two rules. In fact, some properly randomised QMC approximation can converge in L^2 to the exact integral with order $O(N^{-(\alpha+1/2)})$, provided that the corresponding high order QMC rule achieves rate $O(N^{-\alpha})$ with no scrambling. Here, the value of α is closely related to the smoothness of the integrand. These high order rules can be constructed by applying Owen's scrambling and digit interlacing to polynomial lattice rules.

1.1 Polynomial lattice rules

For any b prime, we denote by $\mathbb{Z}_b$ the field of integers modulo b, which can be identified with the set $\{0, 1, \dots, b-1\} \subset \mathbb{Z}$.

Let $\mathbb{Z}_b((x^{-1})) := \left\{ \sum_{i=l}^{\infty} t_i x^{-i} : l \in \mathbb{Z}, \quad t_i \in \mathbb{Z}_b \ \forall i \right\}$ be the field of formal Laurent series over $\mathbb{Z}_b$. Then for a fixed integer m, define the map

$$\begin{aligned} v_m : \mathbb{Z}_b((x^{-1})) &\longrightarrow [0,1) \\ \sum_{i=l}^{\infty} t_i x^{-i} &\longmapsto \sum_{i=max(1,l)}^{m} t_i b^{-i}. \end{aligned}$$

Example For $b = m = 2$, let $h(x) = x^2 + 1 + x^{-1} + x^{-2} + x^{-4} \in \mathbb{Z}_b((x^{-1}))$. Therefore, $l = -2$ and

$$v_m(h(x)) = v_m(\underbrace{x^2 + 1}_{\text{discard integer part}} + x^{-1} + x^{-2} + \underbrace{x^{-4}}_{\text{truncate}}) = 2^{-1} + 2^{-2} = (0.11)_2.$$

In the following, a natural number k is identified with the polynomial $k(x) \in \mathbb{Z}_b[x]$ by replacing b by x in its b-adic expansion. Moreover, the truncation of degree $m-1$ of $k(x)$ is denoted by $tr_m(k) \in \mathbb{Z}_b[x]$. For vectors $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}_0^s$, the function tr_m is applied component by component, that is

$$tr_m(\boldsymbol{k}) = (tr_m(k_1), \dots, tr_m(k_s)).$$

The following construction of QMC lattices was introduced by H. Niederreiter in [26].

Definition 1.1 (Polynomial lattice point set)**.** *Let $m, s \in \mathbb{N}, p \in \mathbb{Z}_b[x]$ be of degree m and $\boldsymbol{q} = (q_1, \dots, q_s) \in (\mathbb{Z}_b[x])^s$. A polynomial lattice point set is defined as $P(\boldsymbol{q}, p) := \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{b^m-1}\} \subset [0,1)^s$ where*

$$\boldsymbol{x}_n := \left(v_m\Big(\frac{n(x)q_1(x)}{p(x)}\Big), \dots, v_m\Big(\frac{n(x)q_s(x)}{p(x)}\Big) \right).$$

If a QMC rule uses $P(\boldsymbol{q}, p))$ as nodes, it is called *polynomial lattice rule* with generating vector $\boldsymbol{q}$ and modulus p.

Definition 1.2 (Dual polynomial lattice)**.** *For a polynomial lattice point set $P(\boldsymbol{q}, p)$, its dual polynomial lattice is given by*

$$P^{\perp}(\boldsymbol{q}, p) := \left\{ \boldsymbol{k} \in \mathbb{N}_0^s : tr_m(\boldsymbol{k}) \cdot \boldsymbol{q} \equiv 0 \pmod{p} \right\},$$

where the inner product is defined, $\forall \boldsymbol{g}, \boldsymbol{q} \in (\mathbb{Z}_b[x])^s$, as

$$\boldsymbol{g} \cdot \boldsymbol{q} := \sum_{j=1}^{s} g_j q_j \in \mathbb{Z}_b[x].$$

1.2 Scrambling algorithms

Digit scrambling was first introduced by A.B. Owen in [28] and it has proven to be a fundamental tool in QMC integration; in [8] J. Dick proved that it can be used to improve the convergence for smooth integrands. By a lower bound in [27, Section 2.2.4], it essentially realises the optimal rate achievable, up to a logarithmic factor to some power dependent linearly on the dimension of the domain. To illustrate how a random scrambling acts on our polynomial lattice, we consider for the moment only one point $\boldsymbol{x} = (x_1, \dots, x_s) \in [0,1)^s$. In fact, once we define scrambling for a single point in $[0,1)^s$, we can apply the same function to each element of a polynomial lattice to obtain a *scrambled polynomial lattice point set*.

Every x_j has b-adic expansion given by $x_j = \sum_{i=1}^{\infty} x_{j,i} b^{-i}$ and such expansion is unique except when $x_{j,i} = b - 1 \ \forall i > i_0$ for some i_0. We refer to the sequence $(x_{j,1}, x_{j,2} \dots)$ as to the *b-adic digits* of x_j. A scrambling algorithm consists in applying (random) permutations of the set $\{0, \dots, b-1\}$ to the digits of x_j, for each $1 \leq j \leq s$. The output will be a new point in $\boldsymbol{y}$ in $[0,1)^s$. The following definition due to Owen, has been introduced in [28].

Definition 1.3. *Let $1 \leq j \leq s$ and $k \in \mathbb{N}$. Define $\pi_{j, x_{j,1}, \dots, x_{j,k-1}}$ to be a random permutation of $\{0, \dots, b-1\}$, depending on j and on the first $k-1$ b-adic digits of x_j; we assume that these permutations are independent uniformly distributed. Then we construct the digits of y_j for each $j = 1, \dots, s$ by applying the permutations as follows: $y_{j,1} = \pi_j(x_{j,1}), y_{j,2} = \pi_{j,x_{j,1}}(x_{j,2}), \dots, y_{j,k} = \pi_{j,x_{j,1},\dots,x_{j,k-1}}(x_{j,k})$ and so on. We then write $\boldsymbol{y} = \boldsymbol{\Pi}(\boldsymbol{x})$ and we call this construction* Owen's scrambling algorithm*.*

The key property of Owen's scrambling is that the output $\boldsymbol{y}$ is a uniformly distributed point in $[0,1)^s$, as it is shown in the next proposition. As a consequence, equation (1.2) holds in the case of Owen's scrambled polynomial lattice point sets and we can again equivalently consider the variance of the estimator as a measure of the integration error.

Proposition 1.4. *[28, Proposition 2] Let $s \in \mathbb{N}$ and $\boldsymbol{y} = \boldsymbol{\Pi}(\boldsymbol{x})$ satisfy that $\boldsymbol{x} \in [0,1)^s$. Then $\boldsymbol{y} = (y_1, \dots, y_s)$ is uniformly distributed in $[0,1)^s$.*

Proof. Let λ_s be the s-dimensional Lebesgue measure. We prove that for all Lebesgue measurable $G \subseteq [0,1)^s$ there holds $\mathbb{P}(\boldsymbol{y} \in G) = \lambda_s(G)$. For $s = 1$, fix a positive integer k and consider the one dimensional b-adic intervals

$E = \left[\frac{a}{b^k}, \frac{a+1}{b^k}\right) \subseteq [0,1)$ where $0 \le a < b^k$. By independence and uniform distribution of the permutations, there holds

$$\mathbb{P}\left(\sum_{i>k} y_{j,i} b^{k-i} = 1\right) = \mathbb{P}(y_{j,i} = b-1,\ \forall i > k) = \prod_{i>k} \underbrace{\mathbb{P}(y_{j,i} = b-1)}_{1/b} = 0.$$

This implies that whether y_j is in E or not depends only on the first k b-adic digits of y_j, up to a set of zero probability. In particular, if a has b-adic expansion $\sum_{i=1}^{k} a_i b^{k-i}$,

$$\mathbb{P}(y_j \in E) = \mathbb{P}(y_{j,i} = a_i\ \forall\, i = 1, \dots, k) = b^{-k} = \lambda_1(E). \tag{1.3}$$

By additivity of λ_1, this can be extended to any $E = \left[\frac{a_1}{b^k}, \frac{a_2}{b^k}\right)$, where $0 \le a_1 \le a_2 < b^k$. Next, arguing by density of the extrema of such itervals, one can prove the equality for Borel measurable subsets of $[0,1)$ using arguments from [7]; moreover, (1.3) holds for all subsets of sets of zero measure. Thus (1.3) holds for any Lebesgue measurable set in $[0,1)$. Finally, independence of the scrambling permutations for $j = 1, \dots, s$ ensures that

$$\mathbb{P}(y_j \in E_j\ \forall\, j = 1, \dots, s) = \prod_{j=1}^{s} \lambda_1(E_j).$$

The claim then follows from the relation $\lambda_s = \otimes_{j=1}^{s} \lambda_1$. □

Even if Owen's scrambling produces uniformly distributed points, its numerical computation is generally not feasible. First of all, the algorithm is virtually infinite since there is no stopping criterion for k. This issue is not relevant in finite precision because $k_{\max}$ can be set to be the number of digits in single or double precision. On the other hand, we would need to store $s(b^{k_{\max}} - 1)/(b-1)$ independent permutations of $\mathbb{Z}_b$, but we are interested in large s and $k_{\max}$. Therefore, for practical purposes one prefers to employ more efficient scrambling schemes that have been studied in [25, 29]. Since we may lose uniform distribution in those cases, the aim of these methods is to control the *discrepancy* of a point set, which is a measure of how far is a point set from having a uniform distribution. We refer to [10, Chapter 3] for more details on discrepancy theory. In the following we introduce a simplified scrambling scheme called *Random linear scrambling*, due to J. Matoušek [25, pag. 540].

Definition 1.5. *Define $x = \sum_{i=1}^{\infty} x_i b^{-i} \in [0,1)$ and $g_i, h_{li} \in \mathbb{Z}_b$. Assume that h_{ii} are randomly chosen in $\{1, \dots, b-1\}$ and that g_i, h_{li} for $l < i$, are randomly chosen in $\{0, 1, \dots, b-1\}$. Moreover, assume that all the choices are independent. Then, we define the digits of $y \in [0,1)$ with the formula*

$$y_i := \tilde{\pi}_i(x) = \sum_{l=1}^{i} h_{li} \cdot x_l + g_i\ ,$$

where arithmetics is meant to be mod b. *In the multi-dimensional case, $\boldsymbol{y} = (y_1, \dots, y_s)$ is defined component by component, by means of scramblings $\tilde{\pi}_{j,i}$ as defined above, taken independently $\forall j = 1, \dots, s$. We then write $\boldsymbol{y} = \tilde{\boldsymbol{\Pi}}(\boldsymbol{x})$.*

Note that each permutation depends on all the previous digits as in Owen's scrambling, but the amount of randomness is considerably reduced by restricting the set of permutations allowed. In the following analysis, we will always consider Owen's scheme, while Matoušek scheme will be used in the implementation.

1.3 Digit interlacing

Definition 1.6. *Let d be a positive integer and $\boldsymbol{x} = (x_1, \ldots, x_d) \in [0,1)^d$. Each component of the vector has b-adic expansion given by $x_j = \sum_{i=1}^{\infty} x_{j,i} b^{-i}$. We define the* digit interlacing function *as*

$$\mathcal{D}_d : [0,1)^d \longrightarrow [0,1) \qquad \mathcal{D}_d(\boldsymbol{x}) := \sum_{i=1}^{\infty} \sum_{j=1}^{d} x_{j,i} b^{-d(i-1)-j}.$$

Analogously, for $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$ with b-adic expansions $k_j = \sum_{i=0}^{\infty} \kappa_{j,i} b^i$, the digit interlacing function is defined as

$$\mathcal{E}_d : \mathbb{N}_0^d \longrightarrow \mathbb{N}_0 \qquad \mathcal{E}_d(\boldsymbol{k}) := \sum_{i=0}^{\infty} \sum_{j=1}^{d} \kappa_{j,i} b^{di+j-1}.$$

For $\boldsymbol{x} \in [0,1)^{ds}$ and $\boldsymbol{k} \in \mathbb{N}_0^{ds}$ the same functions are defined by applying them to every consecutive d components, that is

$$\mathcal{D}_d(\boldsymbol{x}) := \big(\mathcal{D}_d(x_1, \ldots, x_d), \mathcal{D}_d(x_{d+1}, \ldots, x_{2d}), \ldots, \mathcal{D}_d(x_{d(s-1)+1}, \ldots, x_{ds})\big)$$

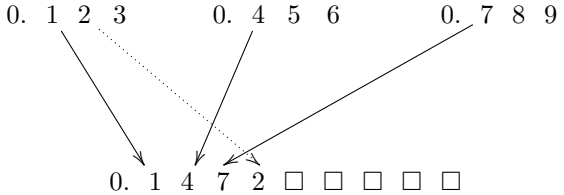
and

$$\mathcal{E}_d(\boldsymbol{k}) := \big(\mathcal{E}_d(k_1, \ldots, k_d), \mathcal{E}_d(k_{d+1}, \ldots, k_{2d}), \ldots, \mathcal{E}_d(k_{d(s-1)+1}, \ldots, k_{ds})\big).$$

The value d is called interlacing factor.

Observe that, for any $k_j \in \mathbb{N}$, there are finitely many non-zero digits $\kappa_{j,i}$. Therefore the sum in the definition of $\mathcal{E}_d$, contains only a finite number of non-zero terms.

Example Define $b = 10, d = 3$ and $\boldsymbol{x} = (0.123, 0.456, 0.789)$. The following diagram illustrates the operation of digit interlacing.



The outcome in this case will be $\mathcal{D}_3(\boldsymbol{x}) = 0.147258369$.

Remark Note that $\mathcal{E}_d$ is bijective while $\mathcal{D}_d$ is only injective. If we choose for example $s = 1$, $y = \mathcal{D}_d(x_1, \dots, x_d)$ then y cannot be of the form $\sum_{i=0}^{\infty} y_i b^{-i}$ with $y_{i_0+dk} = b - 1$ for all $k \geq k_0 \in \mathbb{N}_0$. In fact, this would force one of the x_j to end with an infinite sequence of $b - 1$, but in that case we would use instead the finite b-adic representation of x_j. However, the set of y realising $y_{i_0+dk} = b - 1 \; \forall k \geq k_0$ is only countable.

Since we wish to apply interlacing on a scrambled polynomial lattice, we now have to check that this operation preserves the uniform distribution of the points. Then, one can use again equation (1.2) and look for a bound of the variance of the estimator.

Proposition 1.7. *Let $\boldsymbol{x}$ be uniformly distributed in $[0,1)^{ds}$, then $\boldsymbol{y} := \mathcal{D}_d(\boldsymbol{x})$ is uniformly distributed in $[0,1)^s$.*

Proof. Let λ_s be the s-dimensional Lebesgue measure. We need to show that for any Lebesgue measurable G in $[0,1)^s$ there holds

$$\mathbb{P}(\boldsymbol{y} \in G) \overset{(i)}{=} \mathbb{P}(\boldsymbol{x} \in \mathcal{D}_d^{-1}(G)) \overset{(ii)}{=} \lambda_{ds}(\mathcal{D}_d^{-1}(G)) \overset{(iii)}{=} \lambda_s(G).$$

Item (i) follows from bijectivity of $\mathcal{D}_d : [0,1)^{ds} \to [0,1)^s \backslash N$ where $\lambda_s(N) = 0$ by the remark above.

For (ii) and (iii) we restrict to the case $s = 1$; the general case can be recovered from the relation $\lambda_s = \otimes_{j=1}^{s} \lambda_1$. First, we prove the claim for sets of the type $J := \left[\frac{c}{b^{d\nu}}, \frac{c+1}{b^{d\nu}}\right)$ with $0 \leq c < b^\nu$ for some $\nu \in \mathbb{N}$. Each number in J shares the first $d\nu$ digits with c; thus, the value of c uniquely determines the first ν digits of every component of $\boldsymbol{x} \in \mathcal{D}_d^{-1}(J)$. Moreover, the remaining digits can be chosen freely in J, so that there exist integers $0 \leq a_1, \dots, a_d < b^\nu$ satisfying

$$\mathcal{D}_d^{-1}(J) = \prod_{i=1}^{d} \left[\frac{a_i}{b^\nu}, \frac{a_i + 1}{b^\nu}\right).$$

Then $\mathcal{D}_d^{-1}(J)$ is Lebesgue measurable and $(ii) - (iii)$ hold in this case. Note that the intervals of the type J generate the Borel σ-algebra by density of the endpoints in $[0,1)$ and then $\mathcal{D}_d^{-1}(B) \subseteq [0,1)^d$ is Borel measurable for any Borel set $B \subseteq [0,1)$. Hence, arguing as in Proposition 1.4, we get that $(ii) - (iii)$ hold also for any Borel set $B \subseteq [0,1)$. Let $Z \subseteq [0,1)$ be a λ_1-nullset, then for every $n \in \mathbb{N}$ there exist a family of intervals $(I_k^n)_{k \in \mathbb{N}}$ with $I_k^n \subseteq [0,1)$ such that

$$\sum_k \lambda_1(I_k^n) \leq \frac{1}{n} \qquad \text{and} \qquad Z \subseteq \bigcup_k I_k^n$$

and we can assume with no loss of generality that $I_k^n \subseteq I_k^m$ for all $m \leq n$. Thus, $Z \subseteq \bigcap_n \bigcup_k I_k^n$ and

$$\lambda_1\left(\bigcap_n \bigcup_k I_k^n\right) = \lim_{n \to \infty} \lambda_1\left(\bigcup_k I_k^n\right) \leq \lim_{n \to \infty} \sum_k \lambda_1(I_k^n) = 0.$$

Therefore, completeness of the Lebesgue measure gives that $\mathcal{D}_d^{-1}(Z)$ is measurable and $\lambda_d(\mathcal{D}_d^{-1}(Z)) = 0$, that is $(ii) - (iii)$ hold for Z. Since every Lebesgue measurable G can be written as disjoint union of a Borel set and a λ_1-nullset, the claim follows. □

Now we have all the ingredients to construct our lattice in the following definition.

Definition 1.8. *Let b be a prime number, let $p \in \mathbb{Z}_b[x]$ be of degree m and define $\boldsymbol{q} \in (\mathbb{Z}_b[x])^{ds}$. Let $\boldsymbol{\Pi}$ be a scrambling algorithm and $\mathcal{D}_d$ be the digit interlacing function; then the set*

$$\mathcal{D}_d(\boldsymbol{\Pi}(P(\boldsymbol{q}, p))) := \{\boldsymbol{y}_i = \mathcal{D}_d(\boldsymbol{\Pi}(\boldsymbol{x}_i)) \quad : \boldsymbol{x}_i \in P(\boldsymbol{q}, p)\} \tag{1.4}$$

consists of b^m points in $[0,1)^s$ and it is called interlaced scrambled polynomial lattice point set *of order d. A QMC approximation that uses it as nodes is called* interlaced scrambled polynomial lattice rule *of order d.*

1.4 Walsh decomposition

Walsh functions were introduced in [30] by J.L. Walsh in the context of harmonic analysis, following the construction of the Haar basis; in the same paper, Walsh proved that they form a complete system for continuous functions with bounded variation. However, only several years later, in [23], the Walsh decomposition has been exploited for the first time to study the convergence of QMC integration. More recently, J. Dick used the same idea in combination with scrambled digital nets, (see [8]). In the rest of the chapter we will always assume that b is a prime number.

Definition 1.9 (Walsh functions)**.** *Let $\omega_b := e^{\frac{2\pi i}{b}}$ be a primitive b-th root of unity and $k = \sum_{i=0}^{m-1} \kappa_i b^i \in \mathbb{N}_0$. In one dimension, let $(x_i)_i$ be the b-adic digits of x. The k-th b-adic Walsh function is defined as*

$$\mathrm{wal}_k : [0,1) \longrightarrow \{1, \omega_b, \dots, \omega_b^{b-1}\} \qquad \mathrm{wal}_k(x) := \omega_b^{\kappa_0 x_1 + \dots + \kappa_{m-1} x_m}.$$

Furthermore, for vector valued $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}_0^s$ and $\boldsymbol{x} = (x_1, \dots, x_s) \in [0,1)^s$, define

$$\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) := \prod_{j=1}^{s} \mathrm{wal}_{k_j}(x_j).$$

In his paper, Walsh considered only the case $b = 2$ and often, in the literature, the name Walsh functions only refers to this case. In the following proposition we collect some properties of Walsh functions that will be useful in the ensuing analysis.

Proposition 1.10.

P1. For given $k < b^m$, $\mathrm{wal}_k(x)$ depends only on the first m digits of x. Hence, the functions are piecewise constant.

P2. The set $\{\mathrm{wal}_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{N}_0^s\}$ is a complete orthonormal system of $L^2([0,1]^s)$. In particular, there holds the equality in $L^2([0,1]^s)$ given by

$$f = \sum_{\boldsymbol{k} \in \mathbb{N}_0^s} \hat{f}(\boldsymbol{k}) \, \mathrm{wal}_{\boldsymbol{k}}, \qquad \hat{f}(\boldsymbol{k}) := \int_{[0,1]^s} f(\boldsymbol{x}) \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}.$$

Moreover, for $f \in C^0([0,1)^s)$, the equality above holds pointwise.

P3. Define $\boldsymbol{k} \in \mathbb{N}_0^{ds}$ *and* $\boldsymbol{y} = \mathcal{D}_d(\boldsymbol{x}) \in [0,1)^s$*. Then, there holds*

$$\mathrm{wal}_{\mathcal{E}_d(\boldsymbol{k})}(\boldsymbol{y}) = \prod_{i=1}^{ds} \mathrm{wal}_{k_i}(x_i).$$

P4. $\forall\ \boldsymbol{y}, \boldsymbol{y}'$ *uniformly distributed in* $[0,1)^s$ *and for all* $\boldsymbol{k} \neq \boldsymbol{k}'$ *in* $\mathbb{N}_0^s$ *there holds*

$$\mathbb{E}[\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y})\overline{\mathrm{wal}_{\boldsymbol{k}'}(\boldsymbol{y}')}] = 0.$$

P5. Let $P^\perp(\boldsymbol{q}, p) = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{b^m-1}\}$ *be a polynomial lattice point set. Then*

$$\frac{1}{b^m}\sum_{n=0}^{b^m-1} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n) = \begin{cases} 1 & \text{if } \boldsymbol{k} \in P^\perp(\boldsymbol{q}, p) \\ 0 & \text{else} \end{cases}.$$

P6. $\forall\ y, y'$ *uniformly distributed in* $[0,1)$ *and* $\forall\ k \in \mathbb{N}_0$*, let* $l \in \mathbb{N}_0$ *be such that* $\lfloor b^{l-1} \rfloor \leq k < b^l$*. Then*

$$\mathbb{E}[\mathrm{wal}_k(y)\overline{\mathrm{wal}_k(y')}] = \frac{b}{b-1}\chi_{\lfloor b^l y \rfloor = \lfloor b^l y' \rfloor} - \frac{1}{b-1}\chi_{\lfloor b^{l-1} y \rfloor = \lfloor b^{l-1} y' \rfloor}.$$

P1 and *P3* follow from the definition. *P2* is, for example, proved in [10, Appendix A]. A proof of *P4* and *P6* can be found in [10, Lemma 13.3]. Finally, *P5* is shown in [10, Lemma 4.75].

Proposition 1.11. *Let* P^{IS} *be an interlaced scrambled polynomial lattice point set and* $I_s(F; P^{IS})$ *the corresponding QMC rule approximating the integral of* $F \in L^2([0,1]^s) \cap C^0([0,1)^s)$*. Then there exist quantities* $\sigma_{\boldsymbol{l}}(F)$ *independent of the lattice and* $\Gamma_{\boldsymbol{l}}(\boldsymbol{q}, p)$ *independent of the integrand such that*

$$\mathrm{Var}[I_s(F; P^{IS})] = \sum_{\boldsymbol{l} \in \mathbb{N}_0^{ds} \setminus \{\boldsymbol{0}\}} \sigma_{\boldsymbol{l}}^2(F)\Gamma_{\boldsymbol{l}}(\boldsymbol{q}, p). \tag{1.5}$$

Proof. The proof of this result follows closely [8, Lemma 7]. First decompose F using its Walsh expansion and note that $\hat{F}(\boldsymbol{0})$ is the integral of F. Thus we get

$$\begin{aligned}\left(\frac{1}{b^m}\sum_{n=0}^{b^m-1} F(\boldsymbol{y}_n) - \int_{[0,1]^s} F\right)^2 &= \left(\frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\boldsymbol{k}\in\mathbb{N}_0^s} \hat{F}(\boldsymbol{k})\,\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y}_n) - \int_{[0,1]^s} F\right)^2 \\ &= \left(\frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\boldsymbol{0}\neq\boldsymbol{k}\in\mathbb{N}_0^s} \hat{F}(\boldsymbol{k})\,\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y}_n)\right)^2.\end{aligned}$$

Using equation (1.2) we then obtain

$$\mathrm{Var}\left[\frac{1}{b^m}\sum_{n=0}^{b^m-1} F(\boldsymbol{y}_n)\right] = \frac{1}{b^{2m}}\mathbb{E}\left[\left(\sum_{\boldsymbol{0}\neq\boldsymbol{k}\in\mathbb{N}_0^s} \hat{F}(\boldsymbol{k})\sum_{n=0}^{b^m-1}\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y}_n)\right)^2\right].$$

By the property *P4*, the cross terms corresponding to $\boldsymbol{k} \neq \boldsymbol{k}'$ vanish, thus

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{b^m}\sum_{n=0}^{b^m-1} F(\boldsymbol{y}_n)\right] &= \frac{1}{b^{2m}}\mathbb{E}\left[\sum_{\boldsymbol{0}\neq\boldsymbol{k}\in\mathbb{N}_0^s} |\hat{F}(\boldsymbol{k})|^2 \left(\sum_{n=0}^{b^m-1} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y}_n)\right)^2\right] \\
&= \frac{1}{b^{2m}}\sum_{\boldsymbol{0}\neq\boldsymbol{k}\in\mathbb{N}_0^s} |\hat{F}(\boldsymbol{k})|^2 \sum_{n,n'=0}^{b^m-1} \mathbb{E}\left[\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y}_n)\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y}_{n'})}\right] \\
&= \frac{1}{b^{2m}}\sum_{\boldsymbol{0}\neq\boldsymbol{k}\in\mathbb{N}_0^{ds}} |\hat{F}(\mathcal{E}_d(\boldsymbol{k}))|^2 \sum_{n,n'=0}^{b^m-1} \mathbb{E}\left[\mathrm{wal}_{\mathcal{E}_d(\boldsymbol{k})}(\boldsymbol{y}_n)\overline{\mathrm{wal}_{\mathcal{E}_d(\boldsymbol{k})}(\boldsymbol{y}_{n'})}\right] \\
&= \frac{1}{b^{2m}}\sum_{\boldsymbol{0}\neq\boldsymbol{k}\in\mathbb{N}_0^{ds}} |\hat{F}(\mathcal{E}_d(\boldsymbol{k}))|^2 \sum_{n,n'=0}^{b^m-1} \prod_{i=1}^{ds} \mathbb{E}\left[\mathrm{wal}_{k_i}(x_{n,i})\overline{\mathrm{wal}_{k_i}(x_{n',i})}\right],
\end{aligned}$$

where in the last equality we used property *P3* and independence of the different components of the scrambled point $\boldsymbol{x}_n := \mathcal{D}_d^{-1}(\boldsymbol{y}_n)$. Finally, for any $\boldsymbol{l} \in \mathbb{N}_0^{ds}$, define $\mathcal{B}_{\boldsymbol{l},ds} = \{(k_1,\ldots,k_{ds}) \in \mathbb{N}_0^{ds} : \lfloor b^{l_i-1}\rfloor \leq k_i < b^{l_i} \text{ for } 1 \leq i \leq ds\}$. By *P6*, it follows that $\mathbb{E}\left[\mathrm{wal}_{k_i}(x_{n,i})\overline{\mathrm{wal}_{k_i}(x_{n',i})}\right]$ has the same value for all $\boldsymbol{k}$ in $\mathcal{B}_{\boldsymbol{l}}$ once $\boldsymbol{l}$ is fixed. As a consequence, the claim follows if we define

$$\sigma_{\boldsymbol{l}}^2(F) := \sum_{\boldsymbol{k}\in\mathcal{B}_{\boldsymbol{l},ds}} |\hat{F}(\mathcal{E}_d(\boldsymbol{k}))|^2, \tag{1.6}$$

$$\Gamma_{\boldsymbol{l}}(\boldsymbol{q},p) := \frac{1}{b^{2m}}\sum_{n,n'=0}^{b^m-1}\prod_{i=1}^{ds}\mathbb{E}\left[\mathrm{wal}_{k_i}(x_{n,i})\overline{\mathrm{wal}_{k_i}(x_{n',i})}\right]. \tag{1.7}$$

The dependence of $\Gamma_{\boldsymbol{l}}$ on $\boldsymbol{q}$ and p comes from the definition $\boldsymbol{x}_n := \boldsymbol{\Pi}(\boldsymbol{z}_n)$ for some $\boldsymbol{z}_n \in P(\boldsymbol{q},p)$ (cf. (1.4)). □

1.5 ANOVA decomposition

To develop some theoretical aspects of quasi-Monte Carlo integration, it is sometimes useful to write a function $F : [0,1]^s \to \mathbb{R}$ as combination of functions that depend each on a subset of variables $\mathfrak{u} \subseteq \{1:s\}$. One simple way of proceeding is to freeze a group of variables while letting the others vary, leading to the so called anchored decomposition. Here, we introduce instead the so called ANOVA (ANalysis Of VAriance) decomposition, which dates back to an idea of Hoeffding in [18].

Definition 1.12. *Let F be a function satisfying $F \in L^2([0,1]^s)$ for some $s \in \mathbb{N}$. We define iteratively $F_{\emptyset}^* = \int_{[0,1]^s} F(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$ and*

$$F_{\mathfrak{u}}^*(\boldsymbol{y}_{\mathfrak{u}}) := \int_{[0,1]^{s-|\mathfrak{u}|}} F(\boldsymbol{y})\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} - \sum_{\mathfrak{v}\subset\mathfrak{u}} F_{\mathfrak{v}}^*(\boldsymbol{y}_{\mathfrak{v}}),$$

where for $\mathfrak{u} = \{1:s\}$ we use the convention the integral over the empty set is $F(\boldsymbol{y})$. Then F can be written as

$$F(\boldsymbol{y}) = \sum_{\mathfrak{u}\subseteq\{1:s\}} F_{\mathfrak{u}}^*(\boldsymbol{y}_{\mathfrak{u}}).$$

Finally, we briefly mention some useful properties of the ANOVA decomposition, for more details we refer to [24]. We can easily find from the recurrence relation that the component $F^*_{\mathfrak{u}}$ is independent of y_j whenever $j \notin \mathfrak{u}$. Next, as it is shown in [20, Theorem 2.1] in a more general framework, there holds the annihilating property $\int_0^1 F^*_{\mathfrak{u}}(\boldsymbol{y}_{\mathfrak{u}})\mathrm{d}y_j = 0$ for all $j \in \mathfrak{u}$. Also, ANOVA decomposition is the unique decomposition of the form $F = \sum_{\mathfrak{u}\subseteq\{1:s\}} G_{\mathfrak{u}}$ where the $G_{\mathfrak{u}}$ satisfy these two properties. More generally, the components $(F^*_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ are orthogonal in $L^2([0,1]^s)$; hence

$$\mathrm{Var}(F) = \sum_{\mathfrak{u}\subseteq\{1:s\}} \mathrm{Var}(F^*_{\mathfrak{u}}).$$

This justifies that ANOVA decomposition is an optimal choice to control some kind of L^2 norm of a function.

1.6 A weighted space of differentiable functions

So far, we worked under weak smoothness assumptions of the integrand, as we only required $F \in L^2([0,1]^s) \cap C^0([0,1)^s)$. Under this general hypothesis, the Monte Carlo integration achieves already the best error decay. The goal of this section is to determine a space of smooth functions such that it is possible to control the coefficients $\sigma_{\boldsymbol{l}}(F)$ defined in (1.6). First, we introduce a set of positive parameters $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ called weights. These will play the role of weighting the dependence of F on the different components: in particular, small values of $\gamma_{\mathfrak{u}}$ denote that the components in $\mathfrak{u}$ are less relevant and vice versa. In [21], the authors considered a weighted and unanchored Sobolev space of functions to control first order derivatives. As a result, the convergence of the QMC approximation was capped at $O(N^{-1+\delta})$ for $\delta > 0$ as $N \to \infty$, with constant independent of s. The key idea is that, in order to achieve higher convergence rate, we need to provide a bound on higher order derivatives of F. This stems from the papers [8, 13], where high order QMC analysis is carried out, and from [9, 12], where some applications to parametric PDEs have been studied. For $\boldsymbol{y} = (y_1, \dots, y_s)$ and all $\boldsymbol{\nu} = (\nu_1, \dots, \nu_s) \in \mathbb{N}_0^s$, define the multi-index notation $\partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F = \dfrac{\partial^{\nu_1+\dots+\nu_s} F}{\partial y_1^{\nu_1} \cdots \partial y_s^{\nu_s}}$. For all $\mathfrak{u} \subset \{1:s\}$, we can integrate out the coordinates that are not in $\mathfrak{u}$ to obtain a $|\mathfrak{u}|$-dimensional function and define

$$F_{\mathfrak{u}}(\boldsymbol{y}_{\mathfrak{u}}) := \int_{[0,1]^{s-|\mathfrak{u}|}} F(\boldsymbol{y})\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \ .$$

We also use the convention that for $\mathfrak{u} = \{1:s\}$, $F_{\mathfrak{u}}(\boldsymbol{y}_{\mathfrak{u}}) := F(\boldsymbol{y})$.

Definition 1.13. *Fix an order $\alpha \in \mathbb{N}$ and a positive sequence $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$. The* weighted and unanchored Sobolev space $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ *is defined as the completion of the space $C^\infty([0,1]^s)$ with respect to the norm $\|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}$, where*

$$\|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)} := \sup_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \sum_{\boldsymbol{\nu}\in\{1:\alpha\}^{|\mathfrak{u}|}} \sup_{\boldsymbol{y}_{\mathfrak{u}}\in[0,1]^{|\mathfrak{u}|}} \left| \int_{[0,1]^{s-|\mathfrak{u}|}} \partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F(\boldsymbol{y})\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \right|^2 .$$

Here, if $\mathfrak{u} = \{1:s\}$*, the inner integral is replaced by* $\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} F(\boldsymbol{y})$ *and if* $\mathfrak{u} = \emptyset$*, we use the convention*

$$\sup_{\emptyset} \sum_{\boldsymbol{\nu} \in \{1:\alpha\}^0} \left| \int_{[0,1]^s} \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}_{\{1:s\}} \right|^2 \mathrm{d}\boldsymbol{y}_u := \left| \int_{[0,1]^s} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \right|^2 .$$

As a consequence of the next result, the space $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ is a well defined Banach space.

Proposition 1.14. *The functional* $\|\cdot\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}$ *is a norm.*

Proof. For all $c \in \mathbb{R}$, there holds $\|cF\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)} = |c| \, \|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}$ and the triangular inequality follows from a multiple application of Minkowski inequality with exponent 2 and ∞. To show definiteness, we assume that $\|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)} = 0$. Therefore, we have that $\int_{[0,1]^s} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = 0$ and that for all $\emptyset \neq \mathfrak{u} \subseteq \{1:s\}$, $\boldsymbol{\nu} \in \{1:\alpha\}^{|\mathfrak{u}|}$ and $\boldsymbol{y}_{\mathfrak{u}} \in [0,1]^{|\mathfrak{u}|}$, there holds

$$\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} \int_{[0,1]^{s-|\mathfrak{u}|}} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} = \int_{[0,1]^{s-|\mathfrak{u}|}} \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} F(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} = 0.$$

From $\mathfrak{u} = \{1:s\}$ and $\boldsymbol{\nu} = (1,1,\ldots,1) \in \mathbb{N}^s$, we can deduce that $\partial_{y_1,\ldots,y_s} F \equiv 0$, implying that F is independent of (at least) one variable. Without loss of generality F is independent of the last; therefore, the case $\mathfrak{u} = \{1:s-1\}$ and $\boldsymbol{\nu} = (1,1,\ldots,1) \in \mathbb{N}^{s-1}$ implies

$$\partial_{y_1,\ldots,y_{s-1}} \int_0^1 F(\boldsymbol{y}) \mathrm{d}y_s = \partial_{y_1,\ldots,y_{s-1}} F(\boldsymbol{y}) \equiv 0$$

so that F is also independent of another variable. Iterating s times, we get that F is constant and since F has also vanishing average, we conclude that $F \equiv 0$. □

There is clearly a relation between the ANOVA decomposition and the $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ norm: from the the definition of the functions $F_{\mathfrak{u}}$, there holds

$$F_{\mathfrak{u}}(\boldsymbol{y}_{\mathfrak{u}}) = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} F_{\mathfrak{v}}^*(\boldsymbol{y}_{\mathfrak{v}})$$

and

$$\|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}^2 = \sup_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \sum_{\boldsymbol{\nu} \in \{1:\alpha\}^{|\mathfrak{u}|}} \sup_{\boldsymbol{y}_{\mathfrak{u}} \in [0,1]^{|\mathfrak{u}|}} \left| \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} F_{\mathfrak{u}}(\boldsymbol{y}_{\mathfrak{u}}) \right|^2 .$$

We will sometimes work with a discrete version of the $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ norm, which entails the substitution of derivatives with (multivariate) finite differences. For simplicity we start with univariate functions F: given a point $y \in [0,1]$ and a sequence $(z_j)_{j\geq 1} \subset (-1,1)$, we define iteratively $\Delta_0(y)F = F(y)$ and, for all $\nu \in \mathbb{N}$,

$$\Delta_\nu(y; z_1, \ldots, z_\nu)F = \Delta_{\nu-1}(y + z_\nu; z_1, \ldots, z_{\nu-1})F - \Delta_{\nu-1}(y; z_1, \ldots, z_{\nu-1})F.$$

Here we assume that, for our choice of $(z_j)_{j\geq 1}$, we evaluate F only at points in its domain $[0,1]$. The generalisation to the multivariate case is a tensor product

$$\Delta_{\boldsymbol{\nu}}\big(\boldsymbol{y}; (z_{1,1}, \ldots, z_{1,\nu_1}), \ldots, (z_{s,1}, \ldots, z_{s,\nu_s})\big) := \bigotimes_{i=1}^{s} \Delta_{\nu_i}(y_i; z_{i,1}, \ldots, z_{i,\nu_i}) \ ,$$

where each Δ_{ν_i} is applied only to the i-th component of F while keeping the others fixed. It turns out that we can also define the same difference operator by a closed formula as follows.

Definition 1.15. *Let $\boldsymbol{\nu} \in \{1:\alpha\}^s$ be satisfied for some $\alpha, s \in \mathbb{N}$ and $F : [0,1]^s \to \mathbb{R}$. Then, for all $\boldsymbol{y} = (y_1, \dots, y_s)$ in the domain of F and for all $\boldsymbol{z}_i = (z_{i,1}, \dots, z_{i,\nu_i})$, $i \in \{1:s\}$, with $z_{i,j} \in (-1,1)$ the difference operator $\Delta_{\boldsymbol{\nu}}$ is defined as*

$$\Delta_{\boldsymbol{\nu}}(\boldsymbol{y}; \boldsymbol{z}_1, \dots, \boldsymbol{z}_s)F :=$$

$$= \sum_{\mathfrak{v}_1 \subseteq \{1:\nu_1\}} \cdots \sum_{\mathfrak{v}_s \subseteq \{1:\nu_s\}} (-1)^{|\mathfrak{v}_1|+\dots+|\mathfrak{v}_s|} F\left(y_1 + \sum_{j_1 \in \mathfrak{v}_1} z_{1,j_1}, \dots, y_s + \sum_{j_s \in \mathfrak{v}_s} z_{s,j_s}\right),$$

provided that $y_i + \sum_{j \in \mathfrak{v}_i} z_{i,j} \in [0,1]$ for all $i \in \{1:s\}$ and for all $\mathfrak{v}_i \subseteq \{1:\nu_i\}$.

We now define the generalized weighted Hardy and Krause variation, using the given characterisation of the difference operator and a structure that is similar to the $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ norm. The following definition can be found in [8] and is here slightly modified to include weights.

Definition 1.16. *Fix $\alpha \in \mathbb{N}$ and let $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u} \subseteq \{1:s\}}$ be a sequence of weights. Denote by $\lambda_{|\mathfrak{u}|}$ the Lebesgue measure over $\bigotimes_{k \in \mathfrak{u}} [0,1)$. Define $\Xi_{\mathfrak{u}}$ as the set of all partitions of $\bigotimes_{k \in \mathfrak{u}} [0,1)^{\alpha}$ into subcubes of the form*

$$J = \prod_{\substack{i=1 \\ \lceil i/\alpha \rceil \in \mathfrak{u}}}^{\alpha s} \left[\frac{a_i}{b^{l_i}}, \frac{a_i+1}{b^{l_i}}\right) \qquad \text{where } l_i \in \mathbb{N} \text{ and } 0 \le a_i < b^{l_i},\ \forall i.$$

For all functions $F : [0,1]^s \to \mathbb{R}$ we define the generalised weighted Hardy and Krause variation as

$$V_{s,\boldsymbol{\gamma},\alpha}(F) :=$$

$$= \left(\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \sum_{\boldsymbol{\nu} \in \{1:\alpha\}^{|\mathfrak{u}|}} \sup_{\mathcal{P} \in \Xi_{\mathfrak{u}}} \sum_{J \in \mathcal{P}} \lambda_{|\mathfrak{u}|}(\mathcal{D}_\alpha(J)) \sup_{\mathcal{T}_J} \left| \frac{\Delta_{\boldsymbol{\nu}}(\boldsymbol{y}; \boldsymbol{z}_1, \dots, \boldsymbol{z}_{|\mathfrak{u}|}) F_{\mathfrak{u}}}{\prod_{i=1}^{|\mathfrak{u}|} \prod_{j=1}^{\nu_i} z_{i,j}} \right|^2 \right)^{\frac{1}{2}},$$

where $\mathcal{T}_J$ is the set of choices $(\boldsymbol{y}; \boldsymbol{z}_1, \dots, \boldsymbol{z}_{|\mathfrak{u}|})$, $\boldsymbol{z}_i = (z_{i,1}, \dots, z_{i,\nu_i})$ satisfying:

1. *$\boldsymbol{y} \in \mathcal{D}_\alpha(J)$ for $a_1, \dots, a_{\alpha|\mathfrak{u}|}$ and $l_1, \dots, l_{\alpha|\mathfrak{u}|}$ determined by J as above;*
2. *the operator $\Delta_{\boldsymbol{\nu}}(\boldsymbol{y}; \boldsymbol{z}_1, \dots, \boldsymbol{z}_{|\mathfrak{u}|})$ only takes values of the function in the set $\mathcal{D}_\alpha\left(\prod_{i=1}^{\alpha|\mathfrak{u}|} [\lfloor a_i/b \rfloor b^{-l_i+1}, (\lfloor a_i/b \rfloor + 1) b^{-l_i+1})\right)$;*
3. *for all $i \in \{1:|\mathfrak{u}|\}, j \in \{1:\nu_i\}$, there holds $z_{i,j} = \tau_{i,j} b^{-\alpha(l_i-1)-j}$ for some $\tau_{i,j} \in \{1-b, \dots, b-1\} \backslash \{0\}$.*

Here, we use the convention that the term corresponding to $\mathfrak{u} = \emptyset$ is equal to $\frac{1}{\gamma_\emptyset} \left| \int_{[0,1]^s} F \right|^2$.

We will apply this variation to the terms in the ANOVA decomposition of a function. A crucial observation is that, for a fixed $\mathfrak{v} \subseteq \{1:s\}$, the only non-vanishing term of $V^2_{s,\boldsymbol{\gamma},\alpha}(F^*_\mathfrak{v})$ corresponds to the summand with $\mathfrak{u} = \mathfrak{v}$. In fact, if $\mathfrak{u} \subset \mathfrak{v}$ we have $(F^*_\mathfrak{v})_\mathfrak{u} := \int_{[0,1]^{s-\mathfrak{u}}} F^*_\mathfrak{v} = 0$ by the annihilating property; on the other hand, if $\mathfrak{v} \subset \mathfrak{u}$, $(F^*_\mathfrak{v})_\mathfrak{u}$ is independent of y_j for $j \in \mathfrak{u}\backslash\mathfrak{v}$, so that the difference operator vanishes. A direct application of Definition 1.12 and again the annihilating property imply that $(F^*_\mathfrak{v})_\mathfrak{v} = F_\mathfrak{v}$. Therefore,

$$V^2_{s,\boldsymbol{\gamma},\alpha}(F^*_\mathfrak{v}) = \frac{1}{\gamma_\mathfrak{v}} \sum_{\boldsymbol{\nu}\in\{1:\alpha\}^{|\mathfrak{v}|}} \sup_{\mathcal{P}\in\Xi_\mathfrak{v}} \sum_{J\in\mathcal{P}} \lambda_{|\mathfrak{v}|}(\mathcal{D}_\alpha(J)) \sup_{\mathcal{T}_J} \left| \frac{\Delta_{\boldsymbol{\nu}}\big(\boldsymbol{y}; \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{|\mathfrak{v}|}\big) F_\mathfrak{v}}{\prod_{i=1}^{|\mathfrak{v}|} \prod_{j=1}^{\nu_i} z_{i,j}} \right|^2 . \tag{1.8}$$

Proposition 1.17. *Let $\mathfrak{v} \subseteq \{1:s\}$ and $F : [0,1]^s \to \mathbb{R}$ be a continuous function. If $\partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F_\mathfrak{v} \in C^0([0,1]^s)$ for all $\boldsymbol{\nu} \in \{1:\alpha\}^{|\mathfrak{v}|}$. Then, for every choice of positive weights $\boldsymbol{\gamma}$ there holds*

$$V_{s,\boldsymbol{\gamma},\alpha}(F^*_\mathfrak{v}) \leq \|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)} .$$

Proof. In the definition of $V_{s,\boldsymbol{\gamma},\alpha}(F)$, the expression inside the absolute value is a divided difference of order $\boldsymbol{\nu}$ with $\nu_i \leq \alpha$. Thus, by the mean value theorem, for all $\boldsymbol{\nu}$ and $(\boldsymbol{y}; \boldsymbol{z}_i) \in \mathcal{T}_J$ there exists $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{|\mathfrak{v}|}) \in [0,1)^{|\mathfrak{v}|}$ satisfying

$$\frac{\Delta_{\boldsymbol{\nu}}\big(\boldsymbol{y}; \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{|\mathfrak{v}|}\big) F_\mathfrak{v}}{\prod_{i=1}^{|\mathfrak{v}|} \prod_{j=1}^{\nu_i} z_{i,j}} = \partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F_\mathfrak{v}(\boldsymbol{\xi}).$$

Moreover, for all $\emptyset \neq \mathfrak{v} \subseteq \{1:s\}$ and all $\boldsymbol{\nu} \in \{1:\alpha\}^{|\mathfrak{v}|}$ there holds

$$\sum_{J\in\mathcal{P}} \lambda_{|\mathfrak{v}|}(\mathcal{D}_\alpha(J)) \sup_{(\boldsymbol{y};\boldsymbol{z}_i)\in\mathcal{T}_J} \left|\partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F_\mathfrak{v}(\boldsymbol{\xi})\right|^2 \leq \sup_{\boldsymbol{y}_\mathfrak{v}\in[0,1]^{|\mathfrak{v}|}} \left|\partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F_\mathfrak{v}(\boldsymbol{y}_\mathfrak{v})\right|^2 ,$$

which is a bound independent of the partition. Hence, equation (1.8) implies

$$V_{s,\boldsymbol{\gamma},\alpha}(F^*_\mathfrak{v}) \leq \frac{1}{\gamma_\mathfrak{v}} \sum_{\boldsymbol{\nu}\in\{1:\alpha\}^{|\mathfrak{v}|}} \sup_{\boldsymbol{y}_\mathfrak{v}\in[0,1]^{|\mathfrak{v}|}} \left|\partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F_\mathfrak{v}(\boldsymbol{y}_\mathfrak{v})\right|^2 \leq \|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1])^s} .$$

□

Chapter 2

Bounds on the variance of the estimator

The triple $([0,1]^s, \mathcal{L}([0,1]^s), \otimes_{j=1}^s \mathrm{d}y_j)$ forms a probability space. As discussed in the previous chapters, we look for a bound on the $L^2([0,1]^s)$ error of the randomised QMC approximation, considering the (root of) the variance of the estimator. The upper bound that will be presented in this chapter, based on the work by Goda and Dick [8, 13], takes the form

$$\mathrm{Var}(I_s(F; P^{IS})) \leq B(\boldsymbol{q}, p) \, \|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)} \,, \tag{2.1}$$

where $P^{IS} := \mathcal{D}_d(\boldsymbol{\Pi}(P(\boldsymbol{q}, p)))$ as in equation (1.4). Here $B(\boldsymbol{q}, p)$ is a computable deterministic quantity independent of the dimension s and the integrand. Therefore, it can be used as a quality criterion to determine a good generating polynomial $\boldsymbol{q}$ and modulus p. Then, $\boldsymbol{q}$ will be constructed by a component-by-component (CBC) algorithm, that is using induction over the dimension. Such CBC-constructed polynomial is not necessarily the one minimising $B(\boldsymbol{q}, p)$; however, it is a good approximation that can be found without involving the solution of a multivariate optimisation problem and, additionally, can be implemented using FFT with computational cost $O(dsN \log(N))$.

2.1 Decay of Walsh coefficients

We have seen in (1.5) how the Walsh basis decomposition allows to separate the error due to the integand from the error due to the lattice. The goal of this section is to provide an upper bound for the Walsh coefficients $\sigma_{\boldsymbol{l}}(F)$ defined in (1.6), under the assumption that F is in the weighted space $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ for some positive weights $\boldsymbol{\gamma}$ and $\alpha \in \mathbb{N}$. To ease the notation, we partition the vector $\boldsymbol{l} = (\boldsymbol{l}_{\mathfrak{u}}, \boldsymbol{0}) \in \mathbb{N}_0^{ds}$ so that $\mathfrak{u}$ contains all non-zero components, i.e. $\boldsymbol{l}_{\mathfrak{u}} \in \mathbb{N}^{|\mathfrak{u}|}$. Moreover, for any $\mathfrak{u} \subseteq \{1 : ds\}$, we denote by $\mathfrak{v}(\mathfrak{u}) \subseteq \{1 : s\}$ the set of active dimensions, that is

$$i \in \mathfrak{v}(\mathfrak{u}) \iff \mathfrak{u}_i := \mathfrak{u} \cap \{(i-1)d + 1 : id\} \neq \emptyset$$

(cf. Definition 1.6 of digit interlacing). Moreover, analogously to the proof of Proposition 1.11, we define

$$\mathcal{B}_{(\boldsymbol{l}_{\mathfrak{v}},\mathbf{0}),s} := \{(k_1,\dots,k_s) \in \mathbb{N}_0^s \,:\, b^{l_i-1} \le k_i < b^{l_i} \text{ for } i \in \mathfrak{v} \text{ , } k_i = 0 \text{ for } i \notin \mathfrak{v}\}$$
$$\mathcal{B}_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0}),ds} := \{(k_1,\dots,k_{ds}) \in \mathbb{N}_0^{ds} \,:\, b^{l_i-1} \le k_i < b^{l_i} \text{ for } i \in \mathfrak{u} \text{ , } k_i = 0 \text{ for } i \notin \mathfrak{u}\}.$$

Proposition 2.1. *Let $\boldsymbol{l} = (\boldsymbol{l}_{\mathfrak{u}}, \mathbf{0}) \in \mathbb{N}_0^{ds}$, $d \in \mathbb{N}$ and F with bounded generalized weighted Hardy and Krause variation of order α. Then for the constant $D := 4^{\max(d-\alpha,0)} b^{(2d-1)\alpha}$, there holds*

$$\sigma^2_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(F) \le V^2_{s,\boldsymbol{\gamma},\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})})\gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \frac{(b-1)^{2|\mathfrak{u}|}}{b^{2\min(\alpha,d)|\boldsymbol{l}_{\mathfrak{u}}|+\alpha|\mathfrak{u}|}}. \tag{2.2}$$

The proof of this proposition is a consequence of the two lemmas below. First we define the coefficients $\beta'_k := (b-1)b^{-k+(i-1)d-(l_k-1)d}$ where $k \in \mathfrak{u}_i, i \in \mathfrak{v}(\mathfrak{u})$. Then we reorder the sets $\{\beta'_k \,:\, k \in \mathfrak{u}_i\} = \{\beta_{i,j}(\boldsymbol{l}_{\mathfrak{u}}) \,:\, 1 \le j \le |\mathfrak{u}_i|\}$ so that $\beta_{i,1}(\boldsymbol{l}_{\mathfrak{u}}) < \ldots < \beta_{i,|\mathfrak{u}_i|}(\boldsymbol{l}_{\mathfrak{u}})$ for all $i \in \mathfrak{v}(\mathfrak{u})$. Hence, we write

$$\beta(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0}) := \prod_{i\in\mathfrak{v}(\mathfrak{u})} \prod_{j=1}^{\min(\alpha,|\mathfrak{u}_i|)} \beta_{i,j}(\boldsymbol{l}_{\mathfrak{u}}).$$

Lemma 2.2. *For all $\boldsymbol{l} \in \mathbb{N}_0^{ds}$, let $\mathfrak{u} \subseteq \{1:ds\}$ satisfy that $\boldsymbol{l}_{\mathfrak{u}} \in \mathbb{N}^{|\mathfrak{u}|}$. Let F be of bounded generalized weighted Hardy and Krause variation of order α. Then there holds*

$$\sigma_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(F) \le 2^{|\mathfrak{v}(\mathfrak{u})|\max(d-\alpha,0)} \beta(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0}) \sqrt{\gamma_{\mathfrak{v}(\mathfrak{u})}} V_{s,\boldsymbol{\gamma},\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})}). \tag{2.3}$$

Proof. By definition of the interlacing function, $\mathcal{E}_d(\boldsymbol{\xi}) = \mathbf{0} \iff \boldsymbol{\xi} = \mathbf{0} \in \mathbb{R}^d$, so that $\boldsymbol{k} \in \mathcal{B}_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0}),ds}$ is equivalent to $\mathcal{E}_d(\boldsymbol{k}) \in \mathcal{B}_{(\boldsymbol{l}_{\mathfrak{v}(\mathfrak{u})},\mathbf{0}),s}$. Thus

$$\sigma^2_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(F) := \sum_{\boldsymbol{k}\in\mathcal{B}_{(\boldsymbol{l}_{\mathfrak{v}(\mathfrak{u})},\mathbf{0}),s}} |\hat{F}(\boldsymbol{k})|^2.$$

Each Walsh coefficient, using the ANOVA decomposition of F, satisfies

$$F(\boldsymbol{y}) = \sum_{\mathfrak{v}\subseteq\{1:s\}} F^*_{\mathfrak{v}}(\boldsymbol{y}_{\mathfrak{v}}) \implies \hat{F}(\boldsymbol{k}) = \sum_{\mathfrak{v}\subseteq\{1:s\}} \hat{F}^*_{\mathfrak{v}}(\boldsymbol{k}),$$

where $\hat{F}^*_{\mathfrak{v}}$ denote the Walsh coefficients of $F^*_{\mathfrak{v}}$. We claim that $\forall \boldsymbol{k} \in \mathcal{B}_{(\boldsymbol{l}_{\mathfrak{v}(\mathfrak{u})},\mathbf{0}),s}$ and $\forall \mathfrak{v} \ne \mathfrak{v}(\mathfrak{u})$, there holds $\hat{F}^*_{\mathfrak{v}}(\boldsymbol{k}) = 0$. To show this, fix an arbitrary $\boldsymbol{k} \in \mathcal{B}_{(\boldsymbol{l}_{\mathfrak{v}(\mathfrak{u})},\mathbf{0}),s}$. By Fubini's theorem we get

$$\begin{aligned}
\hat{F}^*_{\mathfrak{v}}(\boldsymbol{k}) &= \int_{[0,1]^s} F^*_{\mathfrak{v}}(\boldsymbol{y}_{\mathfrak{v}}) \prod_{i\in\mathfrak{v}(\mathfrak{u})} \overline{\mathrm{wal}_{k_i}(y_i)} \mathrm{d}\boldsymbol{y} \\
&= \int_{[0,1]^{|v|}} F^*_{\mathfrak{v}}(\boldsymbol{y}_{\mathfrak{v}}) \prod_{i\in\mathfrak{v}(\mathfrak{u})\cap\mathfrak{v}} \overline{\mathrm{wal}_{k_i}(y_i)} \int_{[0,1]^{s-|\mathfrak{v}|}} \prod_{i\in v(\mathfrak{u})\setminus\mathfrak{v}} \overline{\mathrm{wal}_{k_i}(y_i)} \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{v}} \mathrm{d}\boldsymbol{y}_{\mathfrak{v}}
\end{aligned}$$

where by convention the inner integral is defined to be equal to 1 if $\mathfrak{v} = \{1:s\}$. If $\exists i \in \mathfrak{v}(\mathfrak{u})\setminus\mathfrak{v}$, then $k_i \ne 0$ and $\int_0^1 \overline{\mathrm{wal}_{k_i}(y_i)} \mathrm{d}y_i = 0$ so that $\hat{F}^*_{\mathfrak{v}}(\boldsymbol{k}) = 0$. Else, if

$\exists i \in \mathfrak{v}\backslash\mathfrak{v}(\mathfrak{u})$, the above can be rewritten as

$$\begin{aligned}\hat{F}^*_{\mathfrak{v}}(\boldsymbol{k}) &= \int_{[0,1]^{|\mathfrak{v}|}} F^*_{\mathfrak{v}}(\boldsymbol{y}_{\mathfrak{v}}) \prod_{i\in\mathfrak{v}(\mathfrak{u})} \overline{\mathrm{wal}_{k_i}(y_i)}\mathrm{d}\boldsymbol{y}_{\mathfrak{v}} \\ &= \int_{[0,1]^{|\mathfrak{v}(\mathfrak{u})|}} \prod_{i\in\mathfrak{v}(\mathfrak{u})} \overline{\mathrm{wal}_{k_i}(y_i)}\mathrm{d}\boldsymbol{y}_{\mathfrak{v}} \int_{[0,1]^{|\mathfrak{v}|-|\mathfrak{v}(\mathfrak{u})|}} F^*_{\mathfrak{v}}(\boldsymbol{y}_{\mathfrak{v}})\mathrm{d}\boldsymbol{y}_{\mathfrak{v}(\mathfrak{u})}\mathrm{d}\boldsymbol{y}_{\mathfrak{v}\backslash\mathfrak{v}(\mathfrak{u})}\end{aligned}$$

and the inner integral vanishes by the annihilating property of ANOVA decomposition. Thus, $\hat{F}(\boldsymbol{k}) = \hat{F}^*_{\mathfrak{v}(\mathfrak{u})}(\boldsymbol{k})$ and we obtain $\sigma_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}(F) = \sigma_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}(F^*_{\mathfrak{v}(\mathfrak{u})})$. Applying [8, Lemma 9], we get

$$\sigma_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}(F) \le 2^{|\mathfrak{v}(\mathfrak{u})|\max(d-\alpha,0)}\beta(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})V_{s,\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})}),$$

where $V_{s,\alpha} = V_{s,\boldsymbol{1},\alpha}$ is the unweighted counterpart of $V_{s,\boldsymbol{\gamma},\alpha}$. Finally, observe that the annihilating property and being $F^*_{\mathfrak{v}(\mathfrak{u})}$ independent of y_j for $j \notin \mathfrak{v}(\mathfrak{u})$ imply that the only non-zero summand in the generalised Hardy and Krause variation is for the subset $\mathfrak{v}(\mathfrak{u}) \subseteq \{1:s\}$ (cp. equation (1.8)), so that

$$V^2_{s,\boldsymbol{\gamma},\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})}) = \gamma^{-1}_{\mathfrak{v}(\mathfrak{u})}V^2_{s,\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})})$$

and the claim follows. □

Lemma 2.3. *[13, Lemma 3] For $\boldsymbol{l} \in \mathbb{N}_0^{ds}$, let $\emptyset \ne \mathfrak{u} \subseteq \{1:ds\}$ satisfy that $\boldsymbol{l}_{\mathfrak{u}} \in \mathbb{N}^{|\mathfrak{u}|}$ and $\alpha, d, s \in \mathbb{N}$. There holds*

$$\beta(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0}) \le b^{(2d-1)\alpha|\mathfrak{v}(\mathfrak{u})|/2}\frac{(b-1)^{|\mathfrak{u}|}}{b^{\min(\alpha,d)|\boldsymbol{l}_{\mathfrak{u}}|+\alpha|\mathfrak{u}|/2}}.$$

2.2 Quality criterion of a lattice

In this section we will prove an error bound for the interlaced scrambled polynomial lattice as in (2.1). This is similar to a Koksma–Hlawka inequality in the sense that splits the error into two parts, one due to the integrand and the other due to the quality of the lattice. The following lemma, proved in [13, Lemma 2] allows to rewrite (1.5) and makes the gain coefficients $\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}$ more explicit.

Proposition 2.4. *Let $d \in \mathbb{N}$ and $F \in L^2([0,1]^s) \cap C^0([0,1)^s)$. Let $I_s(F;P^{IS})$ be an interlaced scrambled polynomial lattice rule with generating polynomial $\boldsymbol{q} \in \mathbb{Z}_b[x]^{ds}$ and modulus $p \in \mathbb{Z}_b[x]$, approximating $\int_{[0,1]^s} F$. Then there holds*

$$\mathrm{Var}[I_s(F;P^{IS})] = \sum_{\emptyset\ne\mathfrak{u}\subseteq\{1:ds\}} \frac{b^{|\mathfrak{u}|}}{(b-1)^{|\mathfrak{u}|}} \sum_{\boldsymbol{l}_{\mathfrak{u}}\in\mathbb{N}^{|\mathfrak{u}|}} \frac{\sigma^2_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}(F)}{b^{|\boldsymbol{l}_{\mathfrak{u}}|}} |\mathcal{B}_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0}),ds} \cap P^{\perp}(\boldsymbol{q},p)| \tag{2.4}$$

where $P^{\perp}(\boldsymbol{q},p)$ is the dual lattice of Definition 1.2 .

Proof. Recall that by Proposition 1.11, with the notation $\boldsymbol{l} := (\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})$,

$$\mathrm{Var}[I_s(F;P^{IS})] = \sum_{\emptyset\ne\mathfrak{u}\subseteq\{1:ds\}} \sum_{\boldsymbol{l}_{\mathfrak{u}}\in\mathbb{N}^{|\mathfrak{u}|}} \sigma^2_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}(F)\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\boldsymbol{0})}(\boldsymbol{q},p).$$

Hence, it is sufficient to prove that $\forall\ \boldsymbol{l} := (\boldsymbol{l}_{\mathfrak{u}}, \mathbf{0})$ there holds

$$\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(\boldsymbol{q},p) = \frac{b^{|\mathfrak{u}|-|\boldsymbol{l}_{\mathfrak{u}}|}}{(b-1)^{|\mathfrak{u}|}} |\mathcal{B}_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0}),ds} \cap P^{\perp}(\boldsymbol{q},p)|, \tag{2.5}$$

where $\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}$ was defined in (1.7). Note that $k_i = 0 \implies l_i = 0$ so that the property *P6* of Proposition 1.10 gives

$$\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(\boldsymbol{q},p) = \frac{1}{b^{2m}} \sum_{n,n'=0}^{b^m-1} \prod_{i\in\mathfrak{u}} \Big[\frac{b}{b-1} \chi_{\lfloor b^{l_i} x_{n,i} \rfloor = \lfloor b^{l_i} x_{n',i} \rfloor} - \frac{1}{b-1} \chi_{\lfloor b^{l_i-1} x_{n,i} \rfloor = \lfloor b^{l_i-1} x_{n',i} \rfloor} \Big].$$

Let $S_{\boldsymbol{l}_{\mathfrak{u}}}(\mathfrak{v})$ be the set of points $(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) \in [0,1)^{ds} \times [0,1)^{ds}$ such that $x_{n,i}$ and $x_{n',i}$ share the first l_i b-adic digits for all $i \in \mathfrak{v}$ and the first $l_i - 1$ b-adic digits for all $i \in \mathfrak{u}\backslash\mathfrak{v}$. Using the identity $\prod_{i\in\mathfrak{u}}(x_i + y_i) = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (\prod_{i\in\mathfrak{v}} x_i)(\prod_{i\in\mathfrak{u}\backslash\mathfrak{v}} y_i)$, the product above gives

$$\begin{aligned}
&\prod_{i\in\mathfrak{u}} \Big[\frac{b}{b-1} \chi_{\lfloor b^{l_i} x_{n,i} \rfloor = \lfloor b^{l_i} x_{n',i} \rfloor} - \frac{1}{b-1} \chi_{\lfloor b^{l_i-1} x_{n,i} \rfloor = \lfloor b^{l_i-1} x_{n',i} \rfloor} \Big] = \\
&\quad = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \frac{(-1)^{|\mathfrak{u}|-|\mathfrak{v}|} b^{|\mathfrak{v}|}}{(b-1)^{|\mathfrak{u}|}} \prod_{i\in\mathfrak{v}} \chi_{\lfloor b^{l_i} x_{n,i} \rfloor = \lfloor b^{l_i} x_{n',i} \rfloor} \prod_{i\in\mathfrak{u}\backslash\mathfrak{v}} \chi_{\lfloor b^{l_i-1} x_{n,i} \rfloor = \lfloor b^{l_i-1} x_{n',i} \rfloor} \\
&\quad = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \frac{(-1)^{|\mathfrak{u}|-|\mathfrak{v}|} b^{|\mathfrak{v}|}}{(b-1)^{|\mathfrak{u}|}} \chi_{(\boldsymbol{x}_n,\boldsymbol{x}_{n'}) \in S_{\boldsymbol{l}_{\mathfrak{u}}}(\mathfrak{v})}\ .
\end{aligned}$$

If we define

$$\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}} := \{\boldsymbol{k} \in \mathbb{N}_0^{ds} :\ k_i < b^{l_i}\ \forall i \in \mathfrak{v},\ k_i < b^{l_i-1}\ \forall i \in \mathfrak{u}\backslash\mathfrak{v} \text{ and } k_i = 0 \text{ else}\}$$

and if we apply *P1* from Proposition 1.10, we obtain $\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n)\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_{n'})} = 1$ for all $\boldsymbol{k} \in \mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}$ and $(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) \in S_{\boldsymbol{l}_{\mathfrak{u}}}(\mathfrak{v})$. On the other hand, if $(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) \notin S_{\boldsymbol{l}_{\mathfrak{u}}}(\mathfrak{v})$, then $\sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n)\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_{n'})} = 0$, because of the definition of Walsh functions and the identity $\sum_{k=0}^{b^l-1} (\omega_b^{\xi})^k = 0$ valid for all $\xi \in \{1, \ldots, b-1\}$. Thus,

$$\begin{aligned}
\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(\boldsymbol{q},p) &= \frac{1}{b^{2m}} \sum_{n,n'=0}^{b^m-1} \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \frac{(-1)^{|\mathfrak{u}|-|\mathfrak{v}|} b^{|\mathfrak{v}|}}{(b-1)^{|u|}} \frac{1}{|\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}|} \sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n)\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_{n'})} \\
&= \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \frac{(-1)^{|\mathfrak{u}|-|\mathfrak{v}|} b^{|\mathfrak{v}|}}{(b-1)^{|\mathfrak{u}|}} \frac{1}{|\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}|} \sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}} \Big| \frac{1}{b^m} \sum_{n=0}^{b^m-1} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n) \Big|^2 \\
&= \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \frac{(-1)^{|\mathfrak{u}|-|\mathfrak{v}|} b^{|\mathfrak{v}|}}{(b-1)^{|\mathfrak{u}|}} \frac{1}{|\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}|} \sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}} \cap P^{\perp}(\boldsymbol{q},p)} 1
\end{aligned}$$

where in the last equality we used the property *P5*. Note that $|\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}}| = b^{|\boldsymbol{l}_{\mathfrak{u}}|-|\mathfrak{u}|+|\mathfrak{v}|}$ so that

$$\Gamma_{(\boldsymbol{l}_{\mathfrak{u}},\mathbf{0})}(\boldsymbol{q},p) = \frac{b^{|\mathfrak{u}|-|\boldsymbol{l}_{\mathfrak{u}}|}}{(b-1)^{|\mathfrak{u}|}} \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} \sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_{\mathfrak{u}},\mathfrak{v}} \cap P^{\perp}(\boldsymbol{q},p)} 1.$$

Finally, since $\mathfrak{v}_1 \subseteq \mathfrak{v}_2 \iff \mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{v}_1} \subseteq \mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{v}_2}$, then for each $\boldsymbol{k} \in \mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{u}}$ there is a minimal $\mathfrak{v}(\boldsymbol{k}) \subseteq \mathfrak{u}$ with $\boldsymbol{k} \in \mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{v}(\boldsymbol{k})}$. Moreover, $\mathfrak{v}(\boldsymbol{k}) = \mathfrak{u} \iff \boldsymbol{k} \in \mathcal{B}_{(\boldsymbol{l}_\mathfrak{u},\boldsymbol{0}),ds}$ and this implies that

$$\begin{aligned}\sum_{\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{u}|-|\mathfrak{v}|}\sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{v}}\cap P^\perp(\boldsymbol{q},p)} 1 &= \sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{u}}\cap P^\perp(\boldsymbol{q},p)}\ \sum_{\mathfrak{v}(\boldsymbol{k})\subseteq\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{u}|-|\mathfrak{v}|}\\ &= \sum_{\boldsymbol{k}\in\mathcal{R}_{\boldsymbol{l}_\mathfrak{u},\mathfrak{u}}\cap P^\perp(\boldsymbol{q},p)}\chi_{\mathfrak{v}(\boldsymbol{k})=\mathfrak{u}}\\ &= |\mathcal{B}_{(\boldsymbol{l}_u,\boldsymbol{0}),ds}\cap P^\perp(\boldsymbol{q},p)|.\end{aligned}$$

Thus, (2.5) holds and the proof is complete. □

Corollary 2.5. *Let $d \in \mathbb{N}$ and $F \in \mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ with $\partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} F \in C^0([0,1]^s)$ for all $\boldsymbol{\nu} \in \{0:\alpha\}^s$. Let $I_s(F;P^{IS})$ be an interlaced scrambled polynomial lattice rule with generating polynomial $\boldsymbol{q} \in \mathbb{Z}_b[x]^{ds}$ and modulus $p \in \mathbb{Z}_b[x]$, approximating $\int_{[0,1]^s} F$. Then there holds*

$$\mathrm{Var}[I_s(F;P^{IS})] \le \|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}\sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\gamma_{\mathfrak{v}(\mathfrak{u})}D^{|\mathfrak{v}(\mathfrak{u})|}\sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0})\in P^\perp(\boldsymbol{q},p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0}). \tag{2.6}$$

Here D is the constant of Proposition 2.1 and for all $(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0}) \in \mathcal{B}_{(\boldsymbol{l}_\mathfrak{u},\boldsymbol{0})}$ we define

$$r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0}) := \prod_{j\in\mathfrak{u}}\frac{b-1}{b^{\alpha-1}}b^{-(2\min(\alpha,d)+1)l_j} = \frac{(b-1)^{|\mathfrak{u}|}}{b^{(\alpha-1)|\mathfrak{u}|+(2\min(\alpha,d)+1)|\boldsymbol{l}_\mathfrak{u}|}}.$$

Proof. First, from (2.4) we get

$$\begin{aligned}\mathrm{Var}[I_s(F;P^{IS})] &= \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\frac{b^{|\mathfrak{u}|}}{(b-1)^{|\mathfrak{u}|}}\sum_{\boldsymbol{l}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}}\ \sum_{(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0})\in\mathcal{B}_{(\boldsymbol{l}_\mathfrak{u},\boldsymbol{0})}\cap P^\perp(\boldsymbol{q},p)}\frac{\sigma^2_{(\boldsymbol{l}_\mathfrak{u},\boldsymbol{0})}(F)}{b^{|\boldsymbol{l}_\mathfrak{u}|}}\\ &= \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\frac{b^{|\mathfrak{u}|}}{(b-1)^{|\mathfrak{u}|}}\sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0})\in P^\perp(\boldsymbol{q},p)}}\frac{\sigma^2_{(\boldsymbol{l}_\mathfrak{u},\boldsymbol{0})}(F)}{b^{|\boldsymbol{l}_\mathfrak{u}|}}.\end{aligned}$$

We showed in Proposition 1.17 that $V_{s,\boldsymbol{\gamma},\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})}) \le \|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}$. Then, using the upper bound of Proposition 2.1 one obtains

$$\begin{aligned}\mathrm{Var}[I_s(F;P^{IS})] &\le \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\ \sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\(\boldsymbol{k}_u,\boldsymbol{0})\in P^\perp(\boldsymbol{q},p)}}\frac{V^2_{s,\alpha}(F^*_{\mathfrak{v}(\mathfrak{u})})\gamma_{\mathfrak{v}(\mathfrak{u})}D^{|\mathfrak{v}(\mathfrak{u})|}(b-1)^{|\mathfrak{u}|}}{b^{(2\min(\alpha,d)+1)|\boldsymbol{l}_\mathfrak{u}|+(\alpha-1)|\mathfrak{u}|}}\\ &\le \|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}\sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\gamma_{\mathfrak{v}(\mathfrak{u})}D^{|\mathfrak{v}(\mathfrak{u})|}\sum_{\substack{\boldsymbol{k}_u\in\mathbb{N}^{|\mathfrak{u}|}\\(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0})\in P^\perp(\boldsymbol{q},p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0}).\end{aligned}$$

□

This last corollary gives the bound as in (2.1), once we define

$$B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q},p) := \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\gamma_{\mathfrak{v}(\mathfrak{u})}D^{|\mathfrak{v}(\mathfrak{u})|}\sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0})\in P^\perp(\boldsymbol{q},p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u},\boldsymbol{0}). \tag{2.7}$$

We conclude this section with a remark on the dependence of this quality criterion on the parameters $\alpha, d, \boldsymbol{\gamma}$. We see that $B_{\alpha,d,\boldsymbol{\gamma}}$ is formally independent of F, but it still varies with the parameters determining the space $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$. In particular, the corresponding QMC rule will be also influenced by those parameters and remains valid, provided that $F \in \mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ for the same fixed values of $\alpha, d, \boldsymbol{\gamma}$.

2.3 CBC error analysis

The previous sections provide us a quality criterion for the QMC lattice; thus we can proceed with its construction. The first step is to determine the generating vector $\boldsymbol{q}$. Fix the smoothness $\alpha \in \mathbb{N}$, the sequence of positive weights $\boldsymbol{\gamma}$ and the order of the QMC rule $d \in \mathbb{N}$. Given the form of the error estimates of the previous section, a natural way to construct a good generating vector $\boldsymbol{q}$ for a QMC rule is to minimize the function $B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p)$. This can be done efficiently using inductively the components of $\boldsymbol{q}$ already computed. For any $1 \leq \tau \leq ds$, define $\boldsymbol{q}_\tau := (q_1, \ldots, q_\tau)$, $\beta := \lceil \tau/d \rceil$ and generalise $B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p)$ for $\boldsymbol{q}_\tau$ as follows

$$B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}_\tau, p) := \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1:\tau\}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \sum_{\substack{\boldsymbol{k}_\mathfrak{u} \in \mathbb{N}^{|\mathfrak{u}|} \\ (\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \in P^\perp(\boldsymbol{q}_\tau, p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u}, \mathbf{0}). \tag{2.8}$$

Remark For $\tau = ds$ this corresponds to the previous $B_{\alpha,d,\boldsymbol{\gamma}}(\mathbf{q}, p)$ defined in (2.7). Moreover, here the extension of $\boldsymbol{k}_\mathfrak{u}$ with zero components is such that $(\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \in \mathbb{N}^\tau$.

Algorithm - CBC construction of the generating vector. Let $R_m := \{q \in \mathbb{Z}_b[x],\ deg(q) < m,\ q \neq 0\}$ be the set of admissible choices for $\boldsymbol{q}$, that is $|R_m| = b^m - 1$. The CBC algorithm is given by the following steps:

1. Choose an irreducible polynomial $p \in \mathbb{Z}_b[x]$ of degree m;
2. Set $q_1 = 1$;
3. For $\tau = 2, \ldots, ds$ choose $q_\tau := \operatorname{argmin}_{q \in R_m} B_{\alpha,d,\boldsymbol{\gamma}}\big((\boldsymbol{q}_{\tau-1}, q), p\big)$.

The next proposition is the key step to estimate the error of the CBC construction and its proof was first presented in [13, Theorem 1].

Proposition 2.6. *Let $b \in \mathbb{N}$ be prime and $j_0, d_0 \in \mathbb{N}$ satisfy $\tau = (j_0 - 1)d + d_0$ with $0 < d_0 \leq d$. If $\boldsymbol{q}, p$ are constructed with the CBC algorithm above, then for all $\lambda \in \left(\frac{1}{2\min(\alpha, d) + 1}, 1\right]$ and all $\tau = 1, \ldots, ds$, there holds for a positive constant $C_{\lambda,d} := C(b, \alpha, d, \lambda)$,*

$$B_{\alpha,d,\boldsymbol{\gamma}}(\mathbf{q}_\tau, p) \leq \frac{1}{(b^m - 1)^{\frac{1}{\lambda}}} \left[\sum_{\emptyset \neq \mathfrak{v} \subseteq \{1:j_0-1\}} \gamma_\mathfrak{v}^\lambda C_{\lambda,d}^{|\mathfrak{v}|} + C_{\lambda,d_0,d} \sum_{\mathfrak{v} \subseteq \{1:j_0-1\}} \gamma_{\mathfrak{v} \cup \{j_0\}}^\lambda C_{\lambda,d}^{|\mathfrak{v}|} \right]^{\frac{1}{\lambda}}$$

where, for all $d_0 \leq d$, there holds $C_{\lambda,d_0,d} \leq C_{\lambda,d,d} =: C_{\lambda,d}$. Moreover, $C_{\lambda,d}$ satisfies

$$\lim_{\lambda \searrow (2\min(\alpha,d)+1)^{-1}} C_{\lambda,d} = \infty.$$

Proof. We proceed by induction over τ. First, note that there holds

$$\begin{aligned}\sum_{\substack{k=1\\ b^m|k}}^{\infty} r_{\alpha,d}(k) &= \frac{b-1}{b^{\alpha-1}} \sum_{l=1}^{\infty} b^{-(2\min(\alpha,d)+1)l} \sum_{\substack{k=b^{l-1}\\ b^m|k}}^{b^l-1} 1 \\ &= \frac{b-1}{b^{\alpha}} \sum_{l=m+1}^{\infty} \frac{b^{-2\min(\alpha,d)l}}{b^{l-1}} \left(\frac{b^l - b^{l-1}}{b^m}\right) \\ &= \frac{(b-1)^2}{b^{\alpha}} \frac{b^{-2\min(\alpha,d)m}}{b^m(b^{2\min(\alpha,d)}-1)}.\end{aligned}$$

For $\tau = 1$, we have $j_0 = d_0 = 1$ and $q_1 \equiv 1$. Thus, for all $\lambda > \frac{1}{2\min(\alpha,d)+1}$ we get

$$\begin{aligned}B_{\alpha,d,\boldsymbol{\gamma}}(1,p) &= \gamma_{\{1\}} D \sum_{\substack{k=1\\ b^m|k}}^{\infty} r_{\alpha,d}(k) \\ &= \frac{1}{b^{(2\min(\alpha,d)+1)m}} \gamma_{\{1\}} D \frac{(b-1)^2}{b^{\alpha}(b^{2\min(\alpha,d)}-1)} \\ &\le \frac{1}{(b^m-1)^{1/\lambda}} \left[\gamma_{\{1\}}^{\lambda} \left(\frac{D(b-1)^2}{b^{\alpha}(b^{2\min(\alpha,d)}-1)}\right)^{\lambda}\right]^{1/\lambda}\end{aligned}$$

and the claim holds if $C_{\lambda,d} \ge D^{\lambda} \left(\frac{(b-1)^2}{b^{\alpha}(b^{2\min(\alpha,d)}-1)}\right)^{\lambda} =: D^{\lambda} \breve{C}_{\lambda,d}$. Assume now that the result is true for some $1 \le \tau < ds$. In equation (2.8) we separate summands including the $(\tau+1)$-component as follows

$$\begin{aligned}B_{\alpha,d,\boldsymbol{\gamma}}((\boldsymbol{q}_\tau, q), p) &= \sum_{\emptyset \ne \mathfrak{u} \subseteq \{1:\tau+1\}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \sum_{\substack{\boldsymbol{k}_\mathfrak{u} \in \mathbb{N}^{|\mathfrak{u}|} \\ (\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \in P^{\perp}((\boldsymbol{q}_\tau, q), p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \\ &= \sum_{\emptyset \ne \mathfrak{u} \subseteq \{1:\tau\}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \sum_{\substack{\boldsymbol{k}_\mathfrak{u} \in \mathbb{N}^{|\mathfrak{u}|} \\ (\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \in P^{\perp}((\boldsymbol{q}_\tau, q), p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \\ &\quad + \sum_{\substack{\mathfrak{u} \subseteq \{1:\tau+1\} \\ \tau+1 \in \mathfrak{u}}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \sum_{\substack{\boldsymbol{k}_\mathfrak{u} \in \mathbb{N}^{|\mathfrak{u}|} \\ (\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \in P^{\perp}((\boldsymbol{q}_\tau, q), p)}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u}, \mathbf{0}) \\ &=: B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}_\tau, p) + \theta(q).\end{aligned}$$

Throughout the rest of the proof, to simplify the notation, we avoid to explicit the dependency on p unless necessary. The only term that depends on the last component q of the generating vector is $\theta(q)$, so that the choice $q_{\tau+1}$ made in the CBC algorithm satisfies $\theta(q_{\tau+1}) \le \theta(q)$ for all $q \in R_m$. This implies that for all $\lambda > 0$, one has $\theta^{\lambda}(q_{\tau+1}) \le (b^m-1)^{-1} \sum_{q \in R_m} \theta^{\lambda}(q)$.

Jensen's inequality $(\sum c_k)^{\lambda} \le \sum c_k^{\lambda}$ valid for a positive sequence $(c_k)_k$ and

$0 < \lambda \leq 1$, ensures that

$$\theta^\lambda(q_{\tau+1}) \leq \frac{1}{b^m-1} \sum_{q\in R_m} \sum_{\substack{\mathfrak{u}\subseteq\{1:\tau+1\}\\ \tau+1\in\mathfrak{u}}} \gamma_{\mathfrak{v}(\mathfrak{u})}^\lambda D^{\lambda|\mathfrak{v}(\mathfrak{u})|} \sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\ (\boldsymbol{k}_\mathfrak{u},\mathbf{0})\in P^\perp((\boldsymbol{q}_\tau,q))}} r_{\alpha,d}^\lambda(\boldsymbol{k}_\mathfrak{u},\mathbf{0})$$

$$= \sum_{\substack{\mathfrak{u}\subseteq\{1:\tau+1\}\\ \tau+1\in\mathfrak{u}}} \gamma_{\mathfrak{v}(\mathfrak{u})}^\lambda D^{\lambda|\mathfrak{v}(\mathfrak{u})|} \frac{1}{b^m-1} \sum_{q\in R_m} \sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\ (\boldsymbol{k}_\mathfrak{u},\mathbf{0})\in P^\perp((\boldsymbol{q}_\tau,q))}} r_{\alpha,d}^\lambda(\boldsymbol{k}_\mathfrak{u},\mathbf{0}).$$

At this point we distinguish two cases in the inner sum, depending on whether $k_{\tau+1}$ is a multiple of b^m or not. In the first case, $tr_m(k_{\tau+1}) = 0$ and we get

$$(\boldsymbol{k}_\mathfrak{u},\mathbf{0}) \in P^\perp((\boldsymbol{q}_\tau,q)) \iff (\boldsymbol{k}_{\mathfrak{u}\setminus\{\tau+1\}},\mathbf{0}) \in P^\perp(\boldsymbol{q}_\tau)$$

that is a relation independent of q. If instead $tr_m(k_{\tau+1}) \neq 0$, at most one $q \in R_m$ satisfies the equation

$$tr_m(\boldsymbol{k}_\mathfrak{u},\mathbf{0})\cdot(\boldsymbol{q}_\tau,q) = tr_m(k_{\tau+1})q + \sum_{j=1}^{\tau} tr_m(k_j)q_j \equiv 0 \pmod p.$$

In fact, assume by contradiction that $\exists q,\bar{q} \in R_m$ distinct solutions, then subtracting the corresponding equations, the hypothesis that p is irreducible yields $p \mid tr_m(k_{\tau+1}) \ \vee \ p \mid (q-\bar{q})$. This gives a contradiction because p has degree strictly larger than $tr_m(k_{\tau+1})$ and $q-\bar{q}$, and neither of them vanish. The same equation and the assumption that $q \neq 0$ imply $(\boldsymbol{k}_{\mathfrak{u}\setminus\{\tau+1\}},\mathbf{0}) \notin P^\perp(\boldsymbol{q}_\tau)$ in this case. Thus, defining $\bar{\mathfrak{u}} := \mathfrak{u}\setminus\{\tau+1\}$ we obtain

$$\frac{1}{b^m-1} \sum_{q\in R_m} \sum_{\substack{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}\\ (\boldsymbol{k}_\mathfrak{u},\mathbf{0})\in P^\perp((\boldsymbol{q}_\tau,q))}} r_{\alpha,d}^\lambda(\boldsymbol{k}_\mathfrak{u},\mathbf{0}) \tag{2.9}$$

$$= \left(\frac{1}{b^m-1} \sum_{k_{\tau+1}:\ b^m\nmid k_{\tau+1}} r_{\alpha,d}^\lambda(k_{\tau+1})\right) \sum_{\substack{\boldsymbol{k}_{\bar{\mathfrak{u}}}\in\mathbb{N}^{|\mathfrak{u}|-1}\\ (\boldsymbol{k}_{\bar{\mathfrak{u}}},\mathbf{0})\notin P^\perp(\boldsymbol{q}_\tau)}} r_{\alpha,d}^\lambda(\boldsymbol{k}_{\bar{\mathfrak{u}}},\mathbf{0}) \tag{2.10}$$

$$+ \left(\sum_{k_{\tau+1}:\ b^m\mid k_{\tau+1}} r_{\alpha,d}^\lambda(k_{\tau+1})\right) \sum_{\substack{\boldsymbol{k}_{\bar{\mathfrak{u}}}\in\mathbb{N}^{|\mathfrak{u}|-1}\\ (\boldsymbol{k}_{\bar{\mathfrak{u}}},\mathbf{0})\in P^\perp(\boldsymbol{q}_\tau)}} r_{\alpha,d}^\lambda(\boldsymbol{k}_{\bar{\mathfrak{u}}},\mathbf{0}). \tag{2.11}$$

Now we rewrite the term in the parenthesis of (2.10) and (2.11). Note that if $l \leq m$ there are no multiples of b^m in $\{b^{l-1} : b^l-1\}$, while if $l \geq m+1$, there are $\frac{b^l-b^{l-1}}{b^m}$ many multiples. First, for (2.11) there holds

$$\sum_{k:\ b^m\mid k} r_{\alpha,d}^\lambda(k) = \frac{(b-1)^\lambda}{b^{\lambda(\alpha-1)}} \sum_{l=1}^{\infty} b^{-(2\min(\alpha,d)+1)\lambda l} \sum_{\substack{k=b^{l-1}\\ b^m\mid k_{\tau+1}}}^{b^l-1} 1$$

$$= \frac{(b-1)^{1+\lambda}}{b^{m+1+\lambda(\alpha-1)}} \sum_{l=m+1}^{\infty} b^{(1-(2\min(\alpha,d)+1)\lambda)l}.$$

On the other hand, the parenthesis in (2.10) becomes

$$\begin{aligned}\frac{1}{b^m-1}\sum_{k:\ b^m\nmid k} r^\lambda_{\alpha,d}(k) &= \frac{(b-1)^\lambda}{(b^m-1)b^{\lambda(\alpha-1)}}\sum_{l=1}^{\infty} b^{-(2\min(\alpha,d)+1)\lambda l}\sum_{\substack{k=b^{l-1}\\ b^m\nmid k}}^{b^l-1} 1\\ &= \frac{(b-1)^{1+\lambda}}{(b^m-1)b^{1+\lambda(\alpha-1)}}\sum_{l=1}^{m} b^{(1-(2\min(\alpha,d)+1)\lambda)l}\\ &\quad+\frac{(b-1)^{1+\lambda}}{b^{m+1+\lambda(\alpha-1)}}\sum_{l=m+1}^{\infty} b^{(1-(2\min(\alpha,d)+1)\lambda)l}.\end{aligned}$$

Moreover, observe that $\sum_{\boldsymbol{k}_{\mathfrak{u}}\in\mathbb{N}^{|\mathfrak{u}|}} r^\lambda_{\alpha,d}(\boldsymbol{k}_{\mathfrak{u}},\boldsymbol{0}) = \prod_{j\in\mathfrak{u}}\sum_{k_j=1}^{\infty} r^\lambda_{\alpha,d}(k_j)$ by a multinomial identity, so that using similar arguments we also have

$$\sum_{\boldsymbol{k}_{\bar{\mathfrak{u}}}\in\mathbb{N}^{|\mathfrak{u}|-1}} r^\lambda_{\alpha,d}(\boldsymbol{k}_{\bar{\mathfrak{u}}},\boldsymbol{0}) = \prod_{j\in\bar{\mathfrak{u}}}\frac{(b-1)^{1+\lambda}}{b^{1+\lambda(\alpha-1)}}\sum_{l_j=1}^{\infty} b^{(1-(2\min(\alpha,d)+1)\lambda)l_j}.$$

Therefore, we obtain from (2.9) that for all $\mathfrak{u}\subseteq\{1:\tau+1\}$ with $\tau+1\in\mathfrak{u}$,

$$\begin{aligned}&\frac{1}{b^m-1}\sum_{q\in R_m}\sum_{\substack{\boldsymbol{k}_{\mathfrak{u}}\in\mathbb{N}^{|\mathfrak{u}|}\\ (\boldsymbol{k}_{\mathfrak{u}},\boldsymbol{0})\in P^\perp((\boldsymbol{q}_\tau,q))}} r^\lambda_{\alpha,d}(\boldsymbol{k}_{\mathfrak{u}},\boldsymbol{0})\\ &\quad= \frac{(b-1)^{1+\lambda}}{b^{m+1+\lambda(\alpha-1)}}\sum_{l=m+1}^{\infty} b^{(1-(2\min(\alpha,d)+1)\lambda)l}\sum_{\boldsymbol{k}_{\bar{\mathfrak{u}}}\in\mathbb{N}^{|\mathfrak{u}|-1}} r^\lambda_{\alpha,d}(\boldsymbol{k}_{\bar{\mathfrak{u}}},\boldsymbol{0})\\ &\qquad+\frac{(b-1)^{1+\lambda}}{(b^m-1)b^{1+\lambda(\alpha-1)}}\sum_{l=1}^{m} b^{(1-(2\min(\alpha,d)+1)\lambda)l}\sum_{\substack{\boldsymbol{k}_{\bar{\mathfrak{u}}}\in\mathbb{N}^{|\mathfrak{u}|-1}\\ (\boldsymbol{k}_{\bar{\mathfrak{u}}},\boldsymbol{0})\notin P^\perp(\boldsymbol{q}_\tau)}} r^\lambda_{\alpha,d}(\boldsymbol{k}_{\bar{\mathfrak{u}}},\boldsymbol{0})\\ &\quad\le \frac{(b-1)^{1+\lambda}}{(b^m-1)b^{1+\lambda(\alpha-1)}}\sum_{l=1}^{\infty} b^{(1-(2\min(\alpha,d)+1)\lambda)l}\sum_{\boldsymbol{k}_{\bar{\mathfrak{u}}}\in\mathbb{N}^{|\mathfrak{u}|-1}} r^\lambda_{\alpha,d}(\boldsymbol{k}_{\bar{\mathfrak{u}}},\boldsymbol{0})\\ &\quad= \frac{1}{b^m-1}\left[\frac{(b-1)^{1+\lambda}}{b^{1+\lambda(\alpha-1)}}\sum_{l=1}^{\infty} b^{(1-(2\min(\alpha,d)+1)\lambda)l}\right]^{|\mathfrak{u}|} = \frac{1}{b^m-1}\tilde{C}^{|\mathfrak{u}|}_{\lambda,d},\end{aligned}$$

where $\tilde{C}_{\lambda,d} := \dfrac{(b-1)^{1+\lambda}}{b^{\lambda(\alpha-1)}(b^{(2\min(\alpha,d)+1)\lambda}-b)} < \infty$. This in turn implies that

$$\theta^\lambda(q_{\tau+1}) \le \frac{1}{b^m-1}\sum_{\substack{\mathfrak{u}\subseteq\{1:\tau+1\}\\ \tau+1\in\mathfrak{u}}}\gamma^\lambda_{\mathfrak{v}(\mathfrak{u})}D^{\lambda|\mathfrak{v}(\mathfrak{u})|}\tilde{C}^{|\mathfrak{u}|}_{\lambda,d}.$$

We write $\tau+1=(j_1-1)d+d_1$, for some $j_1,d_1\in\mathbb{N}$ such that $d_1\le d$. Define the partition of $\{1:\tau\}$ given by $\bar{S}\cup S$, where $S_j := \{(j-1)d+1:jd\}$ for all $1\le j<j_1$, $S:=\bigcup_{j=1}^{j_1-1}S_j$ and $\bar{S}:=\{(j_1-1)d+1:(j_1-1)d+d_1-1\}$, that is the empty set when $d_1=1$. Then, each $\mathfrak{u}\subseteq\{1:\tau+1\}$ containing $\tau+1$ can be partitioned by the singleton $\{\tau+1\}$, some $\bar{\mathfrak{w}}\subseteq\bar{S}$ and some $\emptyset\ne\mathfrak{w}_j\subseteq S_j$ for all

$j \in \mathfrak{v}(\mathfrak{u})$.

$$\begin{aligned}
\sum_{\substack{\mathfrak{u}\subseteq\{1:\tau+1\}\\ \tau+1\in\mathfrak{u}}} \gamma^{\lambda}_{\mathfrak{v}(\mathfrak{u})} D^{\lambda|\mathfrak{v}(\mathfrak{u})|}\tilde{C}^{|\mathfrak{u}|}_{\lambda,d} &= \sum_{\bar{\mathfrak{w}}\subseteq\bar{S}}\sum_{\mathfrak{w}\subseteq S} \gamma_{\mathfrak{v}(\mathfrak{w})\cup\{j_1\}} D^{\lambda(|\mathfrak{v}(\mathfrak{w})|+1)}\tilde{C}^{|\mathfrak{w}|+|\bar{\mathfrak{w}}|+1}_{\lambda,d} \\
&= \tilde{C}_{\lambda,d} D^{\lambda} \sum_{\bar{\mathfrak{w}}\subseteq\bar{S}} \tilde{C}^{|\bar{\mathfrak{w}}|}_{\lambda,d} \sum_{\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma_{\mathfrak{v}\cup\{j_1\}} D^{\lambda|\mathfrak{v}|} \sum_{\substack{\mathfrak{w}_j\subseteq S_j,\forall j\in\mathfrak{v}\\ \mathfrak{v}(\bigcup \mathfrak{w}_j)=\mathfrak{v}}} \prod_{j\in\mathfrak{v}} \tilde{C}^{|\mathfrak{w}_j|}_{\lambda,d} \\
&= \tilde{C}_{\lambda,d} D^{\lambda} \left(\sum_{\bar{\mathfrak{w}}\subseteq\bar{S}} \tilde{C}^{|\bar{\mathfrak{w}}|}_{\lambda,d}\right) \sum_{\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma_{\mathfrak{v}\cup\{j_1\}} D^{\lambda|\mathfrak{v}|} \prod_{j\in\mathfrak{v}} \sum_{\substack{\mathfrak{w}_j\subseteq S_j\\ \mathfrak{w}_j\neq\emptyset}} \tilde{C}^{|\mathfrak{w}_j|}_{\lambda,d} \\
&= \tilde{C}_{\lambda,d} D^{\lambda} (1+\tilde{C}_{\lambda,d})^{d_1-1} \sum_{\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma_{\mathfrak{v}\cup\{j_1\}} D^{\lambda|\mathfrak{v}|} \Big((1+\tilde{C}_{\lambda,d})^{d}-1\Big)^{|\mathfrak{v}|}.
\end{aligned}$$

In particular, defining $C_{\lambda,a,d} := D^{\lambda}\Big((1+\max\{\tilde{C}_{\lambda,d}, \breve{C}_{\lambda,d}\})^{a}-1\Big)$, to include the case $\tau = 1$, we get

$$\theta^{\lambda}(q_{\tau+1}) \leq \frac{1}{b^m-1}(C_{\lambda,d_1,d} - C_{\lambda,d_1-1,d}) \sum_{\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma_{\mathfrak{v}\cup\{j_1\}} C^{|\mathfrak{v}|}_{\lambda,d,d}.$$

Finally, by Jensen's inequality and inductive hypothesis

$$\begin{aligned}
B^{\lambda}_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}_{\tau+1},p) &\leq B^{\lambda}_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}_{\tau},p) + \theta^{\lambda}(q_{\tau+1}) \\
&\leq \frac{1}{b^m-1}\left[\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:j_0-1\}} \gamma^{\lambda}_{\mathfrak{v}} C^{|\mathfrak{v}|}_{\lambda,d} + C_{\lambda,d_0,d} \sum_{\mathfrak{v}\subseteq\{1:j_0-1\}} \gamma^{\lambda}_{\mathfrak{v}\cup\{j_0\}} C^{|\mathfrak{v}|}_{\lambda,d}\right] \\
&\quad + \frac{1}{b^m-1}(C_{\lambda,d_1,d} - C_{\lambda,d_1-1,d}) \sum_{\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma_{\mathfrak{v}\cup\{j_1\}} C^{|\mathfrak{v}|}_{\lambda,d,d} \\
&= \frac{1}{b^m-1}\left[\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma^{\lambda}_{\mathfrak{v}} C^{|\mathfrak{v}|}_{\lambda,d} + C_{\lambda,d_1,d} \sum_{\mathfrak{v}\subseteq\{1:j_1-1\}} \gamma^{\lambda}_{\mathfrak{v}\cup\{j_1\}} C^{|\mathfrak{v}|}_{\lambda,d}\right],
\end{aligned}$$

where in the last step we used that either $j_1 = j_0$ and $d_1 = d_0 + 1$, so that the equality is trivial, or $j_1 = j_0 + 1$ and $d_1 = 1$, hence $C_{\lambda,d_1-1,d} = 0$ and the result follows. □

Theorem 2.7. *Let $s, m, \alpha, d \in \mathbb{N}$ and b be a prime number. Let $\boldsymbol{\gamma} = (\gamma_{\mathfrak{v}})_{\mathfrak{v}\subseteq\{1:s\}}$ be a sequence of positive weights. Then there exist $p \in \mathbb{Z}_b[x], \boldsymbol{q} \in \mathbb{Z}_b[x]^{ds}$ constructed with a CBC algorithm such that $\mathcal{D}_d(\mathbf{\Pi}(P(\boldsymbol{q},p))) := \{\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{b^m-1}\}$ has the property that, for all $\lambda \in \left(\frac{1}{2\min(\alpha,d)+1}, 1\right]$, there is a positive constant $C := C(b,\alpha,d,\lambda)$ such that for all $F \in \mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ satisfying that*

$\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} F \in C^0([0,1]^s)$ *for all* $\boldsymbol{\nu} \in \{0:\alpha\}^s$*, there holds*

$$\mathbb{E}\left[\left(\frac{1}{b^m}\sum_{i=0}^{b^m-1} F(\boldsymbol{y}_i) - I_s(F)\right)^2\right] \leq \frac{1}{(b^m-1)^{\frac{1}{\lambda}}}\left[\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:s\}} \gamma_{\mathfrak{v}}^{\lambda} C^{|\mathfrak{v}|}\right]^{\frac{1}{\lambda}} \|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)}.$$

Proof. By Proposition 2.6 applied with $\tau = ds$, that is $j_0 = s$ and $d_0 = d$, there holds

$$B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q},p) \leq \frac{1}{(b^m-1)^{\frac{1}{\lambda}}}\left[\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:s\}} \gamma_{\mathfrak{v}}^{\lambda} C^{|\mathfrak{v}|}\right]^{\frac{1}{\lambda}},$$

where C is the constant $C_{\lambda,d}$ of the proposition. The claim then follows by combining the above inequality with (1.2) and Corollary 2.5. □

Corollary 2.8. *Under the assumptions of Theorem 2.7, let* $N = b^m$ *be the number of nodes of the corresponding QMC rule. If we choose* $d \geq \alpha$ *and the weights in product form*

$$\gamma_{\mathfrak{v}} = \prod_{j\in\mathfrak{v}} \gamma_j, \ \textit{with} \quad \gamma_j \sim j^{-(2\alpha+1)} \ \forall j \geq 1,$$

then for all s *and for any* $\delta > 0$*, the* $L^2(\Omega)$ *error of the QMC approximation decays as* $O(N^{-(\alpha+\frac{1}{2})+\delta})$ *as* $N \to \infty$*, with constants independent of* s*.*

Proof. For product weights and for all $\lambda \in \left(\frac{1}{2\min(\alpha,d)+1}, 1\right]$, there holds

$$\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:s\}} \gamma_{\mathfrak{v}}^{\lambda} C^{|\mathfrak{v}|} \leq \exp\left(C\sum_{j\geq 1}\gamma_j^{\lambda}\right) < \infty$$

and the claim follows. □

All the results of this chapter are still valid when $V_{s,\boldsymbol{\gamma},\alpha}(F) < \infty$, as shown in [13]. The key difference here is that the space $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}$ is not based on an ℓ^2 norm over $\mathfrak{u} \subseteq \{1:s\}$, but it only entails a weaker ℓ^∞ criterion, which is more indicated for QMC applications to PDEs. Furthermore, all the integrands in the following chapters will always be smooth, so that this weighted Sobolev space can be used with no loss of information.

Chapter 3

QMC-FEM for affine parametric, elliptic PDEs

Let $D \subset \mathbb{R}^d$ be a bounded interval if $d = 1$, or a polygon if $d = 2$; we wish to study an elliptic partial differential equation on D with uncertain diffusion coefficient $a(x, \boldsymbol{y})$. We distinguish between the space variable $x \in D$ and the parameters $\boldsymbol{y} \in U := \left[-\frac{1}{2}, \frac{1}{2}\right]^{\mathbb{N}}$ and we assume that for $(y_j)_{j\geq 1} \in U$, the y_j are independent and identically uniformly distributed. These parameters model the uncertainty of the following problem: given smooth functions $a(\cdot, \boldsymbol{y}), f$, find $u(x, \boldsymbol{y})$ solving the parametric equation

$$\begin{cases} -\operatorname{div}\Big(a(x, \boldsymbol{y})\nabla u(x, \boldsymbol{y})\Big) = f(x) & x \in D, \\ u(x, \boldsymbol{y}) = 0 & x \in \partial D. \end{cases} \tag{3.1}$$

The operators div and ∇ are only with respect to x and we assume that f is independent of $\boldsymbol{y}$. In the subsequent sections, we will only consider a weaker formulation of this elliptic PDE. Hence, in Section 3.1 we will describe in detail convenient hypothesis on $a(\cdot, \boldsymbol{y})$ and f for the existence and uniqueness of a weak solution. For now, we only mention the assumption that the dependence of the diffusion coefficients on the parameters is affine, that is

$$a(x, \boldsymbol{y}) = \bar{a}(x) + \sum_{j\geq 1} y_j \psi_j(x) \tag{3.2}$$

for some $\bar{a} \in L^\infty(D)$ and a suitable sequence $(\psi_j)_{j\geq 1} \subset L^\infty(D)$. These functions can be interpreted as fluctuations of the diffusion coefficient around the *nominal* value given by $a(x, \mathbf{0}) = \bar{a}(x)$. The triple $(U, \otimes_{j\geq 1}\mathcal{L}([-1/2, 1/2]), \otimes_{j\geq 1}\mathrm{d}y_j)$ is a probability space that reflects independence and uniform distribution of the uncertainties. Our goal is to approximate numerically ensemble averages of (functionals of) the solution u on this probability space, that is

$$I(G(u)) := \int_U G(u(\cdot, \boldsymbol{y}))\mathrm{d}\boldsymbol{y}. \tag{3.3}$$

The functional G is often referred to as *quantity of interest* in the literature. We further assume that G is not influenced by the uncertainty of the equation and

that we are able to evaluate it exactly. Since the problem depends on infinitely many variables, the first step is to reduce it to a finite number s of dimensions; the truncated integral will be denoted by $I_s(G(u_s))$. Next, the integral has to be approximated by a quadrature rule, that we choose to be an interlaced scrambled QMC rule and that we denote by $I_s(\cdot; P^{IS})$. Finally, each evaluation of $u_s(\cdot, \boldsymbol{y})$ consists of the solution of a PDE, so that we have to take into account the corresponding discretisation error and replace u_s with the Galerkin solution $u_{s,h}$

$$\begin{aligned}
\left\|I(G(u)) - I_s(G(u_{s,h}); P^{IS})\right\|_{L^2(U)} \leq & \|I(G(u)) - I_s(G(u_s))\|_{L^2(U)} \\
& + \left\|I_s(G(u_s)) - I_s(G(u_s); P^{IS})\right\|_{L^2(U)} \\
& + \left\|I_s(G(u_s - u_{s,h}); P^{IS})\right\|_{L^2(U)} \\
\leq & |I(G(u)) - I_s(G(u_s))| \qquad (3.4) \\
& + \left\|I_s(G(u_s)) - I_s(G(u_s); P^{IS})\right\|_{L^2(U)} \qquad (3.5) \\
& + \sup_{\boldsymbol{y} \in U} \left|G(u_s(\cdot, \boldsymbol{y}) - u_{s,h}(\cdot, \boldsymbol{y}))\right|. \qquad (3.6)
\end{aligned}$$

These sources of error will be discussed in Section 3.2, Section 3.3 and Section 3.4 respectively.

3.1 Well-posedness analysis

In this section the definition of the problem will be made formal by introducing some assumptions that will ensure well-posedness. We follow the description of [12] based on locally supported uncertainties. The first step is to consider the weak formulation of our problem. We denote by X the Hilbert space $H_0^1(D)$ and by X' its (topological) dual $H^{-1}(D)$; by multipling (3.1) by a test function $v \in X$ and integrating by parts we obtain

$$\int_D a(x, \boldsymbol{y}) \nabla u(x, \boldsymbol{y}) \cdot \nabla v(x) \mathrm{d}x = \int_D f(x) v(x) \mathrm{d}x \qquad \forall v \in X.$$

With a slight abuse of notation motivated by the Riesz representation theorem, we identify with f the functional in X' given by the right hand side. Moreover, we denote the left hand side by the bilinear form $B_{\boldsymbol{y}}(u(\cdot, \boldsymbol{y}), v)$, so that we are left with the following equation: given $f \in X'$ and $\boldsymbol{y} \in U$, find $u(\cdot, \boldsymbol{y}) \in X$ such that

$$B_{\boldsymbol{y}}(u(\cdot, \boldsymbol{y}), v) = f(v) \qquad \forall v \in X. \tag{3.7}$$

In order to ensure existence and uniqueness of the solution of this problem, we introduce the assumptions that for some constants $\bar{a}_{\min}, \bar{a}_{\max} \in \mathbb{R}$, there holds

$$0 < \bar{a}_{\min} \leq \bar{a}(x) \leq \bar{a}_{\max}, \qquad \text{a.e. } x \in D \tag{3.8}$$

and that for a sequence $(b_j)_{j \geq 1} \subseteq (0, 1]$ and a $\kappa \in (0, 1)$,

$$\left\| \frac{\sum_{j \geq 1} |\psi_j| / b_j}{2\bar{a}} \right\|_{L^\infty(D)} \leq \kappa. \tag{3.9}$$

In particular, (3.8) is sufficient to have uniform ellipticity of the nominal equation, while (3.9) states that the fluctuations are relatively small with respect to the nominal diffusion coefficient. At this stage, we do not require any extra hypothesis on $(b_j)_{j\geq 1}$; however, for the QMC error bound we will also impose that this sequence belongs to $\ell^p(\mathbb{N})$ for some $p \in (0,1)$. In [21], the assumptions were similar, but imposed stronger bounds on the ψ_j. In fact, there it was required smallness (compared to $\bar{a}$) of the sum $\sum_{j\geq 1} \|\psi_j\|_{L^\infty(D)}$. While this is easier to work with by means of abstract operator equations, it prevents from exploiting local support of the ψ_j. In turn, (3.9) controls the overlap of the ψ_j and it is more indicated to analyse the case of B-Splines, NURBS, Wavelets or other locally supported fluctuation functions.

To prove well-posedness we will make use of a weaker version of (3.9), obtained by setting $b_j = 1$ for all j. Then we get, for some $0 < \bar{\kappa} < 1$,

$$\left\| \frac{\sum_{j\geq 1} |\psi_j|}{2\bar{a}} \right\|_{L^\infty(D)} \leq \bar{\kappa}. \tag{3.10}$$

Proposition 3.1. *Under the assumptions in* (3.8) *and* (3.10) *for* $\bar{\kappa} \in (0,1)$*, the problem in* (3.7) *has unique solution for all* $\boldsymbol{y} \in U$*. Moreover, there holds the following bound, uniformly in* $\boldsymbol{y}$*:*

$$\|u(\cdot, \boldsymbol{y})\|_X \leq \frac{\|f\|_{X'}}{\bar{a}_{\min}(1-\bar{\kappa})}.$$

Proof. Following [12, Section 2], the proof is an application of the Lax-Milgram lemma; we only have to check that $B_{\boldsymbol{y}}$ is a continuous and coercive bilinear form on $X \times X$, with bounds independent of $\boldsymbol{y}$. Exploiting the affine parametric structure in (3.2), we get

$$\begin{aligned} a(x, \boldsymbol{y}) &\geq \bar{a}_{\min} + \bar{a}_{\min} \frac{\sum_{j\geq 1} y_j \psi_j(x)}{\bar{a}(x)} \\ &\geq \bar{a}_{\min} \left(1 - \frac{\sum_{j\geq 1} |\psi_j(x)|}{2\bar{a}(x)}\right) \quad \text{a.e. } x \in D \end{aligned}$$

and similarly

$$a(x, \boldsymbol{y}) \leq \bar{a}_{\max} \left(1 + \frac{\sum_{j\geq 1} |\psi_j(x)|}{2\bar{a}(x)}\right) \quad \text{a.e. } x \in D.$$

Thus,

$$\begin{aligned} B_{\boldsymbol{y}}(w,w) &\geq \bar{a}_{\min}(1-\bar{\kappa}) \|w\|_X^2 > 0 && \forall w \in X \backslash \{0\}, \\ |B_{\boldsymbol{y}}(w,v)| &\leq \bar{a}_{\max}(1+\bar{\kappa}) \|w\|_X \|v\|_X && \forall w, v \in X \end{aligned}$$

and the claim follows. □

We sometimes identify a continuous bilinear form $B : X \times X \to \mathbb{R}$ with a bounded linear operator $A \in \mathcal{L}(X, X')$ via $B(w,v) = \langle Aw, v \rangle$, where the angled brackets denote the duality product in X, X'. Hence, $B_{\boldsymbol{y}}$ induces a linear operator $A_{\boldsymbol{y}}$ and coercivity of the former translates into bounded invertibility of the latter with the relation $\left\|A_{\boldsymbol{y}}^{-1}\right\|_{\mathcal{L}(X',X)} \leq \dfrac{1}{\bar{a}_{\min}(1-\bar{\kappa})}$.

3.2 Dimension truncation

Fix a finite $s \in \mathbb{N}$. The infinite sum in (3.2) has to be truncated to the first s terms, which is equivalent to setting $y_j = 0 \ \forall j > s$. As a consequence, Proposition 3.1 applies and, for all $\boldsymbol{y} \in U$, we get weak solutions $u_s(\cdot, \boldsymbol{y}) := u(\cdot, (y_1, \ldots, y_s, 0, 0, \ldots)) \in X$ of the equation

$$\int_D \left(\bar{a}(x) + \sum_{j \le s} y_j \psi_j(x) \right) \nabla u_s(x, \boldsymbol{y}) \cdot \nabla v(x) \mathrm{d}x = f(v) \qquad \forall v \in X.$$

We denote the right hand side above by $B_{s,\boldsymbol{y}}(u_s(\cdot, \boldsymbol{y}), v)$ and the corresponding induced linear operator by $A_{s,\boldsymbol{y}}$. The coercivity estimate $B_{s,\boldsymbol{y}}(w, w) \ge \bar{a}_{\min}(1 - \bar{\kappa}) \|w\|_X^2$ is valid for all $w \in X \backslash \{0\}$ following the same lines as in Proposition 3.1, thus $\left\| A_{s,\boldsymbol{y}}^{-1} \right\|_{\mathcal{L}(X',X)} \le \dfrac{1}{\bar{a}_{\min}(1 - \bar{\kappa})}$. Moreover, the integral in (3.3) is over an infinite dimensional set, hence we also restrict it to the subset $\left[-\frac{1}{2}, \frac{1}{2}\right]^s$. We then define for $F : U \to \mathbb{R}$,

$$I_s(F) := \int_{\left[-\frac{1}{2}, \frac{1}{2}\right]^s} F(\boldsymbol{y}) \mathrm{d}y_1 \ldots \mathrm{d}y_s.$$

Note that formally $I_s(G(u_s)) \to I(G(u))$ as $s \to \infty$. The following proposition has been proved in [12, Proposition 5.1] and allows to control the truncation error of the QMC approximation in (3.4).

Proposition 3.2. *Assume the conditions* (3.8), (3.9) *and* (3.10) *for* $\kappa, \bar{\kappa} \in (0, 1)$ *and* $\bar{\kappa} \le \kappa$. *If* $(b_j)_{j \ge 1} \subseteq (0, 1]$, *then for all* $s \in \mathbb{N}$ *and for all* $\boldsymbol{y} \in U$ *there holds*

$$\|u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y})\|_X \le \frac{\bar{a}_{\max} \|f\|_{X'}}{\bar{a}_{\min}^2 (1 - \bar{\kappa})^2} \sup_{j \ge s+1} b_j. \tag{3.11}$$

Moreover, if

$$\frac{\kappa \bar{a}_{\max}}{\bar{a}_{\min}(1 - \bar{\kappa})} \sup_{j \ge s+1} b_j < 1 \ ,$$

then for every $G \in X'$ *holds*

$$|I(G(u)) - I_s(G(u_s))| \le \frac{\|G\|_{X'} \|f\|_{X'}}{(1 - \bar{\kappa})\bar{a}_{\min} - \bar{a}_{\max} \kappa \sup_{j \ge s+1} \{b_j\}} \frac{\kappa^2 \bar{a}_{\max}^2}{(1 - \bar{\kappa})^2 \bar{a}_{\min}^2} \left(\sup_{j \ge s+1} b_j \right)^2. \tag{3.12}$$

Proof. To simplify the notation we omit the dependence of u, u_s on $\boldsymbol{y}$. By definition, $A_{s,\boldsymbol{y}} u_s = f = A_{\boldsymbol{y}} u$; therefore,

$$0 = A_{s,\boldsymbol{y}}(u - u_s) + (A_{\boldsymbol{y}} - A_{s,\boldsymbol{y}})u \implies u - u_s = -A_{s,\boldsymbol{y}}^{-1}(A_{\boldsymbol{y}} - A_{s,\boldsymbol{y}})u.$$

The tail operator $\Delta_{s,\boldsymbol{y}} := A_{\boldsymbol{y}} - A_{s,\boldsymbol{y}}$ has the following expression

$$\Delta_{s,\boldsymbol{y}} w : \ v \mapsto \int_D \sum_{j > s} y_j \psi_j \nabla w \nabla v. \tag{3.13}$$

Thus, the condition (3.9) gives

$$\begin{aligned}\left\|A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}}\right\|_{\mathcal{L}(X,X)} &\leq \frac{1}{\bar{a}_{\min}(1-\bar{\kappa})}\left\|\sum_{j>s}|y_j||\psi_j|\right\|_{L^\infty(D)}\\ &\leq \frac{\bar{a}_{\max}}{\bar{a}_{\min}(1-\bar{\kappa})}\left\|\frac{\sum_{j>s}|\psi_j|/b_j}{2\bar{a}}\right\|_{L^\infty(D)}\sup_{j\geq s+1} b_j\\ &\leq \frac{\kappa\bar{a}_{\max}}{\bar{a}_{\min}(1-\bar{\kappa})}\sup_{j\geq s+1} b_j. \end{aligned} \tag{3.14}$$

With the estimate from Proposition 3.1 we can conclude

$$\|u-u_s\|_X \leq \frac{\bar{a}_{\max}\|f\|_{X'}}{\bar{a}_{\min}^2(1-\bar{\kappa})^2}\sup_{j\geq s+1} b_j$$

and the first part of the theorem is proved. The assumption and (3.14) gives $\left\|A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}}\right\|_{\mathcal{L}(X,X)} < 1$. Denoting by $\mathcal{I}$ the identity operator in X, the Neumann series below converges absolutely:

$$A_{\boldsymbol{y}}^{-1} = (\mathcal{I}+A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}})^{-1}A_{s,\boldsymbol{y}}^{-1} = \sum_{k\geq 0}(A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}})^k A_{s,\boldsymbol{y}}^{-1}.$$

Since $\left[-\frac{1}{2},\frac{1}{2}\right]$ has Lebesgue measure 1 and $u_s = u(\cdot,(y_1,\ldots,y_s,0,0,\ldots))$ for every finite s, then for all $G\in X'$ there holds $I(G(u_s)) = I_s(G(u_s))$ by Fubini's theorem. Therefore, using linearity of G,

$$\begin{aligned}|I(G(u)) - I_s(G(u_s))| &= |I(G(u-u_s))| = |I(G(A_{\boldsymbol{y}}^{-1}f - A_{s,\boldsymbol{y}}^{-1}f))|\\ &\leq \sum_{k\geq 1}\left|G\circ I\left((A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}})^k A_{s,\boldsymbol{y}}^{-1}f\right)\right|\\ &\leq \|G\|_{X'}\sum_{k\geq 1}\left|I\left((A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}})^k u_s\right)\right|\end{aligned}$$

where, for all $u\in X$, in the last two expressions we interpret $I(u)$ as the Bochner integral with respect to the measure $\otimes_{j\geq 1}\mathrm{d}y_j =: \mathrm{d}\boldsymbol{y}$. We now show that the term for $k=1$ vanishes. Let $U_s := \left[-\frac{1}{2},\frac{1}{2}\right]^{\mathbb{N}\setminus\{1:s\}}$. Clearly, $A_{s,\boldsymbol{y}}$ and u_s are independent of y_j for $j>s$ and are measurable with respect to the product σ-algebra $\mathcal{B}(\left[-\frac{1}{2},\frac{1}{2}\right]^s)\otimes\{\emptyset,U_s\}$, then applying Fubini's theorem,

$$\int_U A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}}u_s\mathrm{d}\boldsymbol{y} = \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^s} A_{s,\boldsymbol{y}}^{-1}\int_{U_s}\Delta_{s,\boldsymbol{y}}\mathrm{d}\boldsymbol{y}_{\{1:s\}^c}u_s\mathrm{d}\boldsymbol{y}_{\{1:s\}}.$$

Recall the definition (3.13) of the tail operator $\Delta_{s,\boldsymbol{y}}$: using the identity $\int_{-\frac{1}{2}}^{\frac{1}{2}} y_j\mathrm{d}y_j = 0$ for $j>s$, the inner integral above vanishes. By a geometric series argument

we can finally deduce

$$
\begin{aligned}
|I(G(u)) - I_s(G(u_s))| &\le \|G\|_{X'} \sum_{k\ge 2} \left| I\left((A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}})^k u_s \right) \right| \\
&\le \|G\|_{X'} \sup_{\boldsymbol{y}\in U} \sum_{k\ge 2} \left\| A_{s,\boldsymbol{y}}^{-1}\Delta_{s,\boldsymbol{y}} \right\|_{\mathcal{L}(X,X)}^k \|u_s\|_X \\
&\le \frac{\|G\|_{X'}\|u_s\|_X}{1 - \frac{\kappa\bar{a}_{\max}\sup_{j\ge s+1} b_j}{\bar{a}_{\min}(1-\bar{\kappa})}} \left(\frac{\kappa\bar{a}_{\max}}{\bar{a}_{\min}(1-\bar{\kappa})} \sup_{j\ge s+1} b_j \right)^2 \\
&\le \frac{\|G\|_{X'}\|f\|_{X'}}{(1-\bar{\kappa})\bar{a}_{\min} - \bar{a}_{\max}\kappa \sup_{j\ge s+1}\{b_j\}} \frac{\kappa^2\bar{a}_{\max}^2}{(1-\bar{\kappa})^2\bar{a}_{\min}^2} \left(\sup_{j\ge s+1} b_j \right)^2 .
\end{aligned}
$$

□

3.3 Parametric regularity

With the definition of the norm for $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}\left([-\frac{1}{2},\frac{1}{2}]^s\right)$ as in Section 1.6 in mind, we investigate the behaviour of the derivatives of $u(x,\boldsymbol{y})$ with respect to the parameters $\boldsymbol{y}$. For simplicity, we will make use of the following multiindex notation. For $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots) \in \mathbb{N}_0^{\mathbb{N}}$, define $|\boldsymbol{\nu}| := \sum_{j\ge 1}\nu_j$, $\boldsymbol{\nu}! := \prod_{j\in\mathrm{supp}(\boldsymbol{\nu})}\nu_j!$ and $\mathrm{supp}(\boldsymbol{\nu}) = \{j\in\mathbb{N} : \ \nu_j \neq 0\}$. For sequences $\boldsymbol{b} = (b_j)_{j\ge 1}$, we also write $\boldsymbol{b}^{\boldsymbol{\nu}} := \prod_{j\in\mathrm{supp}(\boldsymbol{\nu})} b_j^{\nu_j}$. Let $\mathcal{F} := \{\boldsymbol{\nu}\in\mathbb{N}_0^{\mathbb{N}} : \ |\boldsymbol{\nu}| < \infty\}$ be the countable set of finitely supported multiindices.

Let $\tilde{U} := [-1,1]^{\mathbb{N}}$ be an auxiliary domain and fix a scaling factor $\eta \in (\kappa, 1)$. For every fixed $\boldsymbol{y}\in U$, we introduce a new diffusion coefficient parametrised by $\boldsymbol{z}\in\tilde{U}$.

$$
\tilde{a}_{\boldsymbol{y}}(x,\boldsymbol{z}) := \bar{a}(x) + \sum_{j\ge 1} y_j\psi_j(x) + \sum_{j\ge 1} z_j \frac{\eta^{-1} - 2|y_j|}{2b_j}\psi_j(x).
$$

Define the shorthand notation $T_{\boldsymbol{y}}(\boldsymbol{z}) := \left(y_j + \frac{\eta^{-1}-2|y_j|}{2b_j} z_j \right)_{j\ge 1}$, so that for all $\boldsymbol{y},\boldsymbol{z}$, we have $\tilde{a}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) = a(\cdot, T_{\boldsymbol{y}}(\boldsymbol{z}))$. This allows to extend the set U of admissible parameters in such a way that the value $\boldsymbol{z} = 0$ gives the equation in (3.1); we will then consider the solution $\tilde{u}_{\boldsymbol{y}}(x,\boldsymbol{z})$ of the equation below and analyse its Taylor expansion with a real variable approach:

$$
\begin{cases}
-\operatorname{div}\left(\tilde{a}_{\boldsymbol{y}}(x,\boldsymbol{z})\nabla\tilde{u}_{\boldsymbol{y}}(x,\boldsymbol{z}) \right) = f(x) & x\in D, \\
\tilde{u}_{\boldsymbol{y}}(x,\boldsymbol{z}) = 0 & x\in\partial D.
\end{cases}
\tag{3.15}
$$

Note the affine parametric structure with respect to $\boldsymbol{z}$ of $\tilde{a}_{\boldsymbol{y}}(x,\boldsymbol{z})$; we can then use similar arguments as the previous sections to prove well-posedness. In particular, by (3.8) and (3.10), for the new nominal operator $\bar{a}_{\boldsymbol{y}}(x) := \bar{a}(x) + \sum_{j\ge 1} y_j\psi_j(x)$ one obtains

$$
0 < (1-\bar{\kappa})\bar{a}_{\min} \le \bar{a}(x) + \sum_{j\ge 1} y_j\psi_j(x) \le (1+\bar{\kappa})\bar{a}_{\max} \quad \text{a.e. } x\in D \tag{3.16}
$$

and by (3.9), together with the assumption that $b_j \leq 1$, the new fluctuations $\psi_{\boldsymbol{y},j} := \frac{\eta^{-1}-2|y_j|}{2b_j}\psi_j$ satisfy

$$\left\| \frac{\sum_{j\geq 1} \frac{\eta^{-1}-2|y_j|}{2b_j} |\psi_j|}{\bar{a} + \sum_{j\geq 1} y_j \psi_j} \right\|_{L^\infty(D)} \leq \left\| \frac{\sum_{j\geq 1} |\psi_j|/(2\eta b_j) - \sum_{j\geq 1} |y_j||\psi_j|/b_j}{\bar{a} - \sum_{j\geq 1} |y_j||\psi_j|} \right\|_{L^\infty(D)}$$
$$\leq \left\| \frac{\sum_{j\geq 1} |\psi_j|/b_j}{2\eta\bar{a}} \right\|_{L^\infty(D)} \leq \frac{\kappa}{\eta} < 1. \tag{3.17}$$

These two inequalities allow to apply Proposition 3.1, so that the weak formulation of (3.15) has a unique weak solution satisfying $\forall \boldsymbol{y} \in U, \boldsymbol{z} \in \tilde{U}$ that

$$\|\tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})\|_X \leq \frac{\|f\|_{X'}}{(1-\bar{\kappa})\bar{a}_{\min}(1-\frac{\kappa}{\eta})}$$

and $\forall \boldsymbol{y} \in U, \boldsymbol{z} \in \tilde{U}$ there holds $\tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) = u(\cdot, T_{\boldsymbol{y}}(\boldsymbol{z})) \in X$. For all $\boldsymbol{y} \in U, \boldsymbol{\nu} \in \mathcal{F}$, define the Taylor coefficients

$$t_{\boldsymbol{y},\boldsymbol{\nu}} := \frac{1}{\boldsymbol{\nu}!} \partial_{\boldsymbol{z}}^{\boldsymbol{\nu}} \tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})\big|_{\boldsymbol{z}=0}.$$

In particular, there holds that $(\|t_{\boldsymbol{y},\boldsymbol{\nu}}\|_X)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^2(\mathcal{F})$ uniformly in $\boldsymbol{y} \in U$. We now present the proof of this summability property, given in [12, Lemma 4.1].

Lemma 3.3. *Assume the conditions* (3.8), (3.9) *and* (3.10) *for* $\kappa, \bar{\kappa} \in (0,1)$ *and* $\bar{\kappa} \leq \kappa$. *Let* $\eta \in (\kappa, 1)$ *be arbitrary, fixed. Then, for every* $\boldsymbol{y} \in U$, *it holds that*

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|_X^2 \leq \frac{\eta(1+\bar{\kappa})}{(\eta-\kappa)(1-\bar{\kappa})^3} \frac{\bar{a}_{\max}}{\bar{a}_{\min}^3} \|f\|_{X'}^2 < \infty. \tag{3.18}$$

Proof. Recall the multivariate product rule for sufficiently regular functions g, h.

$$\partial_{\boldsymbol{z}}^{\boldsymbol{\nu}}(g(\boldsymbol{z})h(\boldsymbol{z})) := \sum_{\boldsymbol{\tau}\leq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{\tau}} \partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} g(\boldsymbol{z}) \partial_{\boldsymbol{z}}^{\boldsymbol{\nu}-\boldsymbol{\tau}} h(\boldsymbol{z})$$

where $\boldsymbol{\tau} \leq \boldsymbol{\nu}$ means that $\tau_j \leq \nu_j \ \forall j \in \mathbb{N}$. Consider the weak formulation of (3.15) and differentiate it with respect to $\boldsymbol{z}$; since f is independent of $\boldsymbol{z}$, we obtain

$$\sum_{\boldsymbol{\tau}\leq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{\tau}} \int_D \partial_{\boldsymbol{z}}^{\boldsymbol{\tau}} \tilde{a}_{\boldsymbol{y}}(x, \boldsymbol{z}) \nabla \partial_{\boldsymbol{z}}^{\boldsymbol{\nu}-\boldsymbol{\tau}} \tilde{u}_{\boldsymbol{y}}(x, \boldsymbol{z}) \cdot \nabla v(x) \mathrm{d}x = 0 \qquad \forall v \in X.$$

The coefficient $\tilde{a}_{\boldsymbol{y}}(x, \boldsymbol{z})$ is affine parametric for the nominal operator $\bar{a}_{\boldsymbol{y}}(x)$ and the fluctuations $\psi_{\boldsymbol{y},j}(x)$. Thus, the only non-zero terms correspond to $\boldsymbol{\tau} = 0$ or $\boldsymbol{\tau} = \boldsymbol{e}_j$ with $j \in \mathrm{supp}(\boldsymbol{\nu})$, where $\boldsymbol{e}_j$ denotes the sequence with 1 in the j-th position and zeros elsewhere. Setting $\boldsymbol{z} = 0$ and dividing by $\boldsymbol{\nu}!$ we get a recurrence relation for the Taylor coefficients:

$$\int_D \bar{a}_{\boldsymbol{y}}(x) \nabla t_{\boldsymbol{y},\boldsymbol{\nu}}(x) \cdot \nabla v(x) \mathrm{d}x = - \sum_{j\in\mathrm{supp}(\boldsymbol{\nu})} \int_D \psi_{\boldsymbol{y},j} \nabla t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}(x) \cdot \nabla v(x) \mathrm{d}x.$$

If we choose $v = t_{\boldsymbol{y},\boldsymbol{\nu}}$, Young inequality and (3.17) give

$$\begin{aligned}\int_D \bar{a}_{\boldsymbol{y}} |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \mathrm{d}x &\leq \frac{1}{2} \sum_{j\in\mathrm{supp}(\boldsymbol{\nu})} \int_D |\psi_{\boldsymbol{y},j}| \left(|\nabla t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 + |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \right) \mathrm{d}x \\ &\leq \frac{\kappa}{2\eta} \int_D \bar{a}_{\boldsymbol{y}}(x) |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 + \frac{1}{2} \sum_{j\in\mathrm{supp}(\boldsymbol{\nu})} \int_D |\psi_{\boldsymbol{y},j}| |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \mathrm{d}x.\end{aligned}$$

Define the energy norm $\|u\|^2_{\bar{a}_{\boldsymbol{y}}} := \int_D \bar{a}_{\boldsymbol{y}} |\nabla u|^2$ for all $u \in X$, thus we proved that

$$\left(1 - \frac{\kappa}{2\eta}\right) \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}} \leq \frac{1}{2} \sum_{j\in\mathrm{supp}(\boldsymbol{\nu})} \int_D |\psi_{\boldsymbol{y},j}| |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \mathrm{d}x.$$

Fix $k \in \mathbb{N}$; summing over all the $\boldsymbol{\nu} \in \mathcal{F}$ such that $|\boldsymbol{\nu}| = k$ leads to

$$\begin{aligned}\sum_{\substack{\boldsymbol{\nu}\in\mathcal{F}\\ |\boldsymbol{\nu}|=k}} \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}} &\leq \frac{\eta}{2\eta-\kappa} \sum_{\substack{\boldsymbol{\nu}\in\mathcal{F}\\ |\boldsymbol{\nu}|=k}} \sum_{j\in\mathrm{supp}(\boldsymbol{\nu})} \int_D |\psi_{\boldsymbol{y},j}| |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \mathrm{d}x \\ &= \frac{\eta}{2\eta-\kappa} \sum_{\substack{\boldsymbol{\nu}\in\mathcal{F}\\ |\boldsymbol{\nu}|=k-1}} \sum_{j\geq 1} \int_D |\psi_{\boldsymbol{y},j}| |\nabla t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \mathrm{d}x \\ &\leq \frac{\kappa}{2\eta-\kappa} \sum_{\substack{\boldsymbol{\nu}\in\mathcal{F}\\ |\boldsymbol{\nu}|=k-1}} \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}}\end{aligned}$$

where in the last inequality we applied again (3.17). Iteration of this argument implies that

$$\begin{aligned}\sum_{k=0}^{\infty} \sum_{\substack{\boldsymbol{\nu}\in\mathcal{F}\\ |\boldsymbol{\nu}|=k}} \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}} &\leq \sum_{k=0}^{\infty} \left(\frac{\kappa}{2\eta-\kappa}\right)^k \|\tilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{0})\|^2_{\bar{a}_{\boldsymbol{y}}} \\ &= \frac{2\eta-\kappa}{2\eta-2\kappa} \|u(\cdot,\boldsymbol{y})\|^2_{\bar{a}_{\boldsymbol{y}}} \\ &\leq \frac{\eta}{\eta-\kappa} \|u(\cdot,\boldsymbol{y})\|^2_{\bar{a}_{\boldsymbol{y}}}.\end{aligned}$$

Observe that by (3.16) the two norms $\|\cdot\|_X$ and $\|\cdot\|_{\bar{a}_{\boldsymbol{y}}}$ are equivalent since

$$(1-\bar{\kappa})\bar{a}_{\min} \|u\|^2_X \leq \|u\|^2_{\bar{a}_{\boldsymbol{y}}} \leq (1+\bar{\kappa})\bar{a}_{\max} \|u\|^2_X.$$

Hence, recalling the estimate from Proposition 3.1 we conclude

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_X \leq \frac{1}{(1-\bar{\kappa})\bar{a}_{\min}} \sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}} \leq \frac{\eta(1+\bar{\kappa})}{(\eta-\kappa)(1-\bar{\kappa})^3} \frac{\bar{a}_{\max}}{\bar{a}^3_{\min}} \|f\|^2_{X'}$$

and the proof is complete. □

In the next step we find a sequence of weights that implies finiteness of the $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}\left(\left[-\frac{1}{2},\frac{1}{2}\right]^s\right)$ norm. As in [12], the locality of the fluctuations leads to weights in product form, that is

$$\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j \qquad \text{for some} \qquad (\gamma_j)_{j\geq 1} \subset (0,+\infty).$$

In the following proof we will use the observation that, since $T_{\boldsymbol{y}}(\boldsymbol{z})$ is affine and $\tilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) = u(\cdot, T_{\boldsymbol{y}}(\boldsymbol{z}))$, the chain rule gives the relation

$$t_{\boldsymbol{y},\boldsymbol{\nu}} = \frac{1}{\boldsymbol{\nu}!}\left(\prod_{j\in\mathrm{supp}(\boldsymbol{\nu})}\left(\frac{\eta^{-1}-2|y_j|}{2b_j}\right)^{\nu_j}\right)\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}u(\cdot,\boldsymbol{y}) \tag{3.19}$$

that links directly the Taylor coefficients with the partial derivatives of u.

Proposition 3.4. *Fix $s,\alpha\in\mathbb{N}$. Assume the conditions* (3.8), (3.9) *and* (3.10) *for $\kappa,\bar{\kappa}\in(0,1)$ and $\bar{\kappa}\le\kappa$. Let $\eta\in(\kappa,1)$ be arbitrary, fixed. Then, if we choose the weights $\boldsymbol{\gamma}=(\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ in the product form*

$$\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}}\sum_{\nu=1}^{\alpha}\left[\left(\frac{2b_j}{1-\eta}\right)^{\nu}\nu!\right]^2 \tag{3.20}$$

there exist a real constant $K_1<\infty$ independent of s such that

$$\|G(u_s)\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([-\frac{1}{2},\frac{1}{2}]^s)} \le K_1.$$

Proof. Define $F := G(u_s)$, then by (3.19) and Lemma 3.3, for all $\boldsymbol{y}\in U$ and $\boldsymbol{\nu}$ with $\mathrm{supp}(\boldsymbol{\nu})=\mathfrak{u}$, there holds

$$\begin{aligned}
\left|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}F(\boldsymbol{y})\right| &\le \|G\|_{X'}\,\boldsymbol{\nu}!\left(\prod_{j\in\mathfrak{u}}\left(\frac{2b_j}{\eta^{-1}-2|y_j|}\right)^{\nu_j}\right)\|t_{\boldsymbol{y},\boldsymbol{\nu}}\|_X \\
&\le \|G\|_{X'}\,\boldsymbol{\nu}!\left(\prod_{j\in\mathfrak{u}}\left(\frac{2b_j}{1-\eta}\right)^{\nu_j}\right)\|t_{\boldsymbol{y},\boldsymbol{\nu}}\|_X \\
&\le K_1\left(\prod_{j\in\mathfrak{u}}\left(\frac{2b_j}{1-\eta}\right)^{\nu_j}\nu_j!\right),
\end{aligned}$$

where

$$K_1 := \sqrt{\frac{\eta(1+\bar{\kappa})}{(\eta-\kappa)(1-\bar{\kappa})^3}\frac{\bar{a}_{\max}}{\bar{a}_{\min}^3}}\,\|f\|_{X'}\,\|G\|_{X'}.$$

Hence, by Jensen's integral inequality and the estimate above we get

$$\begin{aligned}
\|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([-\frac{1}{2},\frac{1}{2}]^s)} &\le \sup_{\mathfrak{u}\subseteq\{1:s\}}\frac{1}{\gamma_{\mathfrak{u}}}\sum_{\boldsymbol{\nu}\in\{1:\alpha\}^{|\mathfrak{u}|}}\sup_{\boldsymbol{y}\in[-\frac{1}{2},\frac{1}{2}]^s}\left|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}F(\boldsymbol{y})\right|^2 \\
&\le K_1^2\sup_{\mathfrak{u}\subseteq\{1:s\}}\frac{1}{\gamma_{\mathfrak{u}}}\sum_{\boldsymbol{\nu}\in\{1:\alpha\}^{|\mathfrak{u}|}}\prod_{j\in\mathfrak{u}}\left[\left(\frac{2b_j}{1-\eta}\right)^{\nu_j}\nu_j!\right]^2 \\
&\le K_1^2\sup_{\mathfrak{u}\subseteq\{1:s\}}\frac{1}{\gamma_{\mathfrak{u}}}\prod_{j\in\mathfrak{u}}\sum_{\nu=1}^{\alpha}\left[\left(\frac{2b_j}{1-\eta}\right)^{\nu}\nu!\right]^2.
\end{aligned}$$

The claim then follows from the definition of the weights, upon taking square roots. □

Observe that the summability of the Taylor coefficients is not necessary in the proof above, since we only used uniform boundedness of those, independently of the dimension s. The definition of the product weights given in the previous proposition is independent of s; this is consistent with our goal, since we aim to find a dimension independent upper bound on the approximation error $\left\|I(G(u)) - I_s(G(u_{s,h}); P^{IS})\right\|_{L^2(U)}$. Therefore, the definition of $(\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ in (3.20) can be extended with no modifications to every set $\mathfrak{u}$ with $|\mathfrak{u}| < \infty$ due to the arbitrariety of s. We make use of this observation to bound the QMC error.

Proposition 3.5. *Fix $s, \alpha \in \mathbb{N}$ and $C \in (0, +\infty)$. Assume that $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, that $b_j \in (0,+\infty)$ for all $j \in \mathbb{N}$ and choose $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1:s\}}$ as in (3.20). Then, for all $\lambda \in [p/2, 1]$ there exists a constant K_2 independent of s such that*

$$\sum_{\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} C^{|\mathfrak{u}|} \leq K_2. \tag{3.21}$$

Proof. The claim is true if and only if for all $\lambda \in [p/2, 1]$ there holds

$$\sum_{\substack{\mathfrak{u}\subset\mathbb{N}\\ |\mathfrak{u}|<\infty}} \gamma_{\mathfrak{u}}^{\lambda} C^{|\mathfrak{u}|} < \infty.$$

Observe that if $x_j > 0$ for all $j \in \mathbb{N}$, then

$$\sum_{|\mathfrak{u}|<\infty} \prod_{j\in\mathfrak{u}} x_j = \prod_{j\geq 1}(1+x_j) = \exp\left[\sum_{j\geq 1}\log(1+x_j)\right] \leq \exp\left[\sum_{j\geq 1} x_j\right].$$

This and the Jensen's inequality $(\sum c_k)^{\lambda} \leq \sum c_k^{\lambda}$, valid for a positive sequence $(c_k)_k$ and $0 < \lambda \leq 1$, ensure that

$$\begin{aligned}
\sum_{|\mathfrak{u}|<\infty} \gamma_{\mathfrak{u}}^{\lambda} C^{|\mathfrak{u}|} &\leq \sum_{|\mathfrak{u}|<\infty} \prod_{j\in\mathfrak{u}} C\left[\sum_{\nu=1}^{\alpha}\left(\frac{2b_j}{1-\eta}\right)^{2\nu}(\nu!)^2\right]^{\lambda} \\
&\leq \exp\left[\sum_{j\geq 1} C\left[\sum_{\nu=1}^{\alpha}\left(\frac{2b_j}{1-\eta}\right)^{2\nu}(\nu!)^2\right]^{\lambda}\right] \\
&\leq \exp\left[C\sum_{\nu=1}^{\alpha}\sum_{j\geq 1}\left(\frac{2b_j}{1-\eta}\right)^{2\lambda\nu}(\nu!)^{2\lambda}\right] \\
&\leq \exp\left[C\sum_{\nu=1}^{\alpha}(\nu!)^{2\lambda}\left(\frac{2}{1-\eta}\right)^{2\lambda\nu}\sum_{j\geq 1} b_j^{2\lambda\nu}\right].
\end{aligned}$$

It remains to prove that $\sum_{j\geq 1} b_j^{2\lambda\nu} < \infty \ \forall \nu = 1, \dots, \alpha$. Since $\boldsymbol{b} \in \ell^p(\mathbb{N})$ and $b_j > 0 \ \forall j$, there exists j_0 large such that $b_j \leq 1 \ \forall j \geq j_0$. Thus, for all $\lambda \geq p/2$

$$\sum_{j\geq 1} b_j^{2\lambda\nu} \leq \sum_{j<j_0} b_j^{2\lambda\nu} + \sum_{j\geq j_0} b_j^{p} < \infty.$$

□

We are now ready to apply the results of Section 2.3 to bound the QMC error in (3.5). First of all, note that the interlaced scrambled points P^{IS} must be shifted from the box $[0,1]^s$ to $[-\frac{1}{2},\frac{1}{2}]^s$, in order to be consistent with the integration domain. This does not affect the computation in Theorem 2.7 and the result remains valid with no loss of generality.

Theorem 3.6. *Let $s \in \mathbb{N}$. Assume that (3.8) and (3.9) hold for $\kappa \in (0,1)$ and a sequence $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}}$ satisfying $b_j \in (0,1]$ and $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$. Then, there exists an interlaced scrambled polynomial lattice point set P^{IS} of cardinality N and of order $d := \left\lfloor \frac{1}{p} - \frac{1}{2} \right\rfloor + 1$ constructed with the CBC algorithm of Section 2.3, such that for all $\lambda \in [p/2, 1]$ there holds*

$$\mathbb{E}\left[|I_s(G(u_s)) - I_s(G(u_s); P^{IS})|^2\right] \le K \frac{1}{(N-1)^{\frac{1}{\lambda}}},$$

where K is a positive constant independent of s.

Proof. Let $d, \alpha \in \mathbb{N}$. By Theorem 2.7, there exists a CBC constructed P^{IS} of cardinality N satisfying

$$\begin{aligned}\mathbb{E}\left[|I_s(G(u_s)) - I_s(G(u_s); P^{IS})|^2\right] \le \\ \le \frac{1}{(N-1)^{\frac{1}{\lambda}}} \left[\sum_{\emptyset\neq \mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} C^{|\mathfrak{u}|}\right]^{\frac{1}{\lambda}} \|G(u_s)\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([-\frac{1}{2},\frac{1}{2}]^s)}\end{aligned}$$

for all $\lambda \in \left(\frac{1}{2\min(\alpha,d)+1}, 1\right]$. Recall that the condition (3.9) with $b_j \in (0,1]$ implies (3.10), so that we can apply Proposition 3.4 and 3.5. Therefore, if we define $K := \left(K_1 \cdot \max(1, K_2)^{\frac{1}{p}}\right)^2$, we obtain the desired bound for $\lambda \in [p/2, 1]$, provided that

$$\frac{1}{2\min(\alpha,d)+1} < \frac{p}{2}.$$

It remains to show the order of the QMC rule. Being $G(u_s) \in \mathcal{W}_{s,\boldsymbol{\gamma},\alpha}\left([-\frac{1}{2},\frac{1}{2}]^s\right)$ for arbitrary $\alpha \in \mathbb{N}$, the condition above is equivalent to

$$\frac{1}{2d+1} < \frac{p}{2} \iff \frac{1}{p} - \frac{1}{2} < d.$$

Since $d \in \mathbb{N}$, it is sufficient to choose $d = \left\lfloor \frac{1}{p} - \frac{1}{2} \right\rfloor + 1$ and thus the claim follows. □

If one compares these results with those in [12], where interlaced polynomial lattice point sets were used, it is clear that the main difference is the introduction of scrambling in the quadrature rule. As a result, the interlacing factor (or order) of the QMC rule can be reduced from $\left\lfloor \frac{1}{p} \right\rfloor + 1$ to $\left\lfloor \frac{1}{p} - \frac{1}{2} \right\rfloor + 1$. On the other hand, we did not bound a worst case, deterministic error but the $L^2(U)$ error because of the extra randomisation. Regarding the order of convergence, there is essentially no difference: here, the decay of the QMC error when $N \to \infty$ is of $O(N^{-\frac{1}{p}})$ independently of the dimension s, while in [12, Corollary 6.4] was instead of $O(N^{-\frac{1}{p}+\varepsilon})$ for all $\varepsilon > 0$.

3.4 High order Galerkin discretisation

To complete the error analysis, it remains to estimate the Galerkin discretisation error in (3.6). The first assumption that we introduce simplifies the construction of a FE space on the physical domain of the equation.

$$D \subset \mathbb{R}^d \text{ is a bounded polygon if d=2 or a bounded interval if d=1.} \quad (3.22)$$

It is well known that if the coefficients $a(\cdot, \boldsymbol{y})$ are sufficiently smooth and if $f \in H^m(D)$ for some $m \in \mathbb{N}_0$, then the corresponding solution of (3.1) satisfies $u \in H^{m+2}(D)$, provided that D is smooth or convex. This results are known as full regularity shift of elliptic operators. On the other hand, the solution u on non-convex polygonal domains in $\mathbb{R}^2$ can develop singularities at the corners. Therefore, we introduce a class of Sobolev spaces that allow to recover a similar regularity shift property. The idea introduced in [19] is to allow the functions in this space to have less regularity at the corners of the polygon, introducing weights that vanish on those points. To make this precise, we define the weight function

$$r_D(x) := \prod_{j \in J} |x - c_j|,$$

where $\{c_1, \dots, c_J\}$ is the (finite) set of the corners of D. We also introduce the notation $r_D^\beta(x) = \prod_{j \in J} |x - c_j|^{\beta_j}$ for a tuple $\beta = (\beta_1, \dots, \beta_J)$. Then, for some given $m, l \in \mathbb{N}_0$ with $l \leq m$, we define the spaces $H_\beta^{m,l}(D)$ as the completion of $C^\infty(\bar{D})$ under the norms

$$\|v\|_{H_\beta^{m,l}(D)} := \left(\|v\|^2_{H^{l-1}(D)} + \sum_{|\alpha|=l}^{m} \left\| |D^\alpha v| r_D^{\beta+|\alpha|-l} \right\|^2_{L^2(D)} \right)^{1/2} \quad (3.23)$$

$$\|v\|_{H_\beta^{m,0}(D)} := \left(\sum_{|\alpha|=0}^{m} \left\| |D^\alpha v| r_D^{\beta+|\alpha|} \right\|^2_{L^2(D)} \right)^{1/2} \quad (3.24)$$

with the usual multiindex notation D^α for derivatives in x. We also denote $H_\beta^{0,0}(D)$ by $L^2_\beta(D)$. These spaces are called Babuška-Kondrat'ev spaces and must not be confused with the weighted Sobolev spaces introduced in Section 1.6. We prove in the next lemmas some useful properties of those spaces. Throughout the proofs in this section, C denotes a generic constant only dependent on m, D and the inequalities in the assumptions (3.8), (3.9), (3.10) that can change within the same formula, unless differently specified.

Lemma 3.7. *If $m, m', l \in \mathbb{N}_0$ satisfy $l \leq m \leq m'$, then $H_\beta^{m',l}(D) \subseteq H_\beta^{m,l}(D)$ with continuous embedding. Moreover, if $m, l, l' \in \mathbb{N}_0$ satisfy $l \leq l' \leq m$, then $H_\beta^{m,l'}(D) \subseteq H_\beta^{m,l}(D)$ with continuous embedding.*

Proof. Clearly $\|v\|_{H_\beta^{m,l}(D)} \leq \|v\|_{H_\beta^{m',l}(D)}$ if $l \leq m \leq m'$. For the second statement, assume $l > 0$ and $l' = l + 1 \leq m$. Since $|r_D(x)| \leq \operatorname{diam}(D)^J < \infty$, for all

$v \in H^{m,l+1}_\beta(D)$ it holds,

$$\begin{aligned}
\|v\|^2_{H^{m,l}_\beta(D)} &= \|v\|^2_{H^{l-1}(D)} + \sum_{|\alpha|=l} \left\| |D^\alpha v| r_D^\beta \right\|^2_{L^2(D)} \\
&\quad + \sum_{|\alpha|=l+1}^{m} \left\| |D^\alpha v| r_D^{\beta+|\alpha|-l} \right\|^2_{L^2(D)} \\
&\leq \|v\|^2_{H^{l-1}(D)} + C \sum_{|\alpha|=l} \| |D^\alpha v| \|^2_{L^2(D)} \\
&\quad + C \sum_{|\alpha|=l+1}^{m} \left\| |D^\alpha v| r_D^{\beta+|\alpha|-(l+1)} \right\|^2_{L^2(D)} \\
&\leq C \|v\|^2_{H^{m,l+1}_\beta(D)} .
\end{aligned}$$

The case $l = 0$ is analogous. Iterating this inequality gives the claim when $l' > l + 1$. □

The next lemma rephrases [2, equation 3.2].

Lemma 3.8. *If $D \subset \mathbb{R}^2$ is bounded and $\beta_j < 1$ for all $j = 1, \dots, J$, then $L^2_\beta(D) \subseteq H^{-1}(D)$ with continuous embedding.*

Proof. Every $f \in L^2_\beta(D)$ defines a linear functional $H^1_0(D) \ni v \to \int_D fv \in \mathbb{R}$. Note that $r_D^{-\beta} \in L^{p^*}$ for some $p^* > 2$. Hence, for $p = \frac{2p^*}{p^*-2}$ Hölder inequality gives

$$\left| \int_D fv \right| \leq \left\| f r_D^\beta \right\|_{L^2(D)} \left\| v r_D^{-\beta} \right\|_{L^2(D)} \leq C \|f\|_{L^2_\beta(D)} \|v\|_{L^p(D)} .$$

By the Sobolev embedding $\|v\|_{L^p(D)} \leq C \|v\|_{H^1_0(D)}$ we conclude $\|f\|_{H^{-1}(D)} \leq C \|f\|_{L^2_\beta(D)}$. □

Lemma 3.9.

1. *Let $w \in W^{m,\infty}(D)$ and $v \in H^{m,l}_\beta(D)$. Then there is a constant C dependent on m, l, D such that,*

$$\|wv\|_{H^{m,l}_\beta(D)} \leq C \|w\|_{W^{m,\infty}(D)} \|v\|_{H^{m,l}_\beta(D)} .$$

2. *Let $w \in W^{m+1,\infty}(D)$ and $v \in H^{m+1,l+1}_\beta(D)$. Then there is a constant C dependent on m, D such that,*

$$\|\nabla w \cdot \nabla v\|_{H^{m,l}_\beta(D)} \leq C \|w\|_{W^{m+1,\infty}(D)} \|v\|_{H^{m+1,l+1}_\beta(D)} .$$

Proof. For the first item, we observe the inequality $\binom{\alpha}{\gamma} \leq 2^{|\alpha|}$. Hence,

$$\begin{aligned}
\|wv\|^2_{H^{m,l}_\beta(D)} &= \sum_{|\alpha|\leq m} \int_D (|D^\alpha(wv)| r_D^{\beta+|\alpha|-l})^2 \\
&\leq \sum_{|\alpha|\leq m} \int_D \left(\sum_{\gamma\leq\alpha} \binom{\alpha}{\gamma} |D^{\alpha-\gamma} w D^\gamma v| r_D^{\beta+|\alpha|-l} \right)^2 \\
&\leq \|w\|^2_{W^{m,\infty}(D)} \sum_{|\alpha|\leq m} \int_D \left(\sum_{\gamma\leq\alpha} \binom{\alpha}{\gamma} |D^\gamma v| r_D^{\beta+|\alpha|-l} \right)^2 \\
&\leq 2^{m+1} \|w\|^2_{W^{m,\infty}(D)} \sum_{|\alpha|\leq m} \sum_{\gamma\leq\alpha} \left\| |D^\gamma v| r_D^{\beta+|\alpha|-l} \right\|^2_{L^2(D)} \\
&\leq C \|w\|^2_{W^{m,\infty}(D)} \left(\|v\|^2_{H^{l-1}(D)} + \sum_{|\gamma|=l}^{m} \left\| |D^\gamma v| r_D^{\beta+|\gamma|-l} \right\|^2_{L^2(D)} \right)
\end{aligned}$$

where in the last step we used that $|r_D|$ is bounded and that each γ appears at most 2^m times in the double sum. For the second item, we similarly get

$$\|\nabla w \cdot \nabla v\|^2_{H^{m,l}_\beta(D)} \leq C \|w\|^2_{W^{m+1,\infty}(D)} \|\nabla v\|^2_{H^{m,l}_\beta(D)}$$

and

$$\begin{aligned}
\|\nabla v\|^2_{H^{m,l}_\beta(D)} &\leq C \left(\|v\|^2_{H^l(D)} + \sum_{|\gamma|=l+1}^{m+1} \left\| |D^\gamma v| r_D^{\beta+|\gamma|-1-l} \right\|^2_{L^2(D)} \right) \\
&\leq C \|v\|^2_{H^{m+1,l+1}_\beta(D)}
\end{aligned}$$

and the proof is complete. □

Next, we impose extra regularity for the data f and the quantity of interest G; this will later determine the order of convergence of the Galerkin solution. For $m, m' \in \mathbb{N}_0$ we require

$$f \in H^{m,0}_\beta(D), \qquad G \in H^{m',0}_\beta(D). \tag{3.25}$$

Moreover, the diffusion coefficients $a(\cdot, \boldsymbol{y})$ need to be regular in space, uniformly with respect to the parameters $\boldsymbol{y}$. In particular, we assume that

$$a(\cdot, \boldsymbol{y}) \in W^{m+1,\infty}(D) \quad \text{with} \quad \sup_{\boldsymbol{y}\in U} \|a(\cdot, \boldsymbol{y})\|_{W^{m+1,\infty}(D)} < \infty. \tag{3.26}$$

For all $h > 0$, let X_h be a family of nested dense subspaces of $H^1_0(D)$, each satisfying $M_h := \dim(X_h) = O(h^{-d})$ for $d \in \{1, 2\}$. Then, we assume that there exist $h_0 > 0$ and $C > 0$ such that for all $h < h_0$ and for all $v \in H^{m+2,2}_\beta \cap H^1_0(D)$ there holds the discretisation property

$$\inf_{v_h \in X_h} \|v - v_h\|_{H^1_0(D)} \leq C h^{m+1} \|v\|_{H^{m+2,2}_\beta(D)}. \tag{3.27}$$

This is satisfied, for example, with the explicit construction of the FE triangulation given in [6, Theorems 0.3 and 4.4], for piecewise polynomials of degree at most $m+1$ and continuous on $\bar{D}$, where h is the parameter describing the size of the elements. A similar construction is also available for 3-dimensional polyhedra (see [4]).

The regularity shift of the Laplacean stated in the following theorem was proved in [1, Theorem 2.1].

Theorem 3.10. *For a polygon $D \in \mathbb{R}^2$, denote by ω_j the interior angle of the corner c_j for $j = 1, \dots, J$. Let $m \in \mathbb{N}_0$ and $\beta_j \in [0,1)$ be such that $\beta_j > 1 - \pi/\omega_j$. Then, there is a constant C depending on m, D such that if $\Delta v \in H^{m,0}_\beta(D)$ and $v \in H^1_0(D)$, there holds that $v \in H^{m+2,2}_\beta(D)$ and*

$$\|v\|_{H^{m+2,2}_\beta(D)} \le C \|\Delta v\|_{H^{m,0}_\beta(D)} < \infty.$$

Next, we want to extend the regularity shift to the class of elliptic PDEs of our interest.

Proposition 3.11. *Define the multiindices β as in the statement of Theorem 3.10. Let $f \in H^{m,0}_\beta(D)$ and assume that (3.8), (3.10) and (3.26) hold. Then, the solution $u(\cdot, \boldsymbol{y})$ of (3.1) belongs to $H^{m+2,2}_\beta(D)$ and there exist a positive constant C independent of $\boldsymbol{y}$, such that*

$$\|\Delta u(\cdot, \boldsymbol{y})\|_{H^{m,0}_\beta(D)} \le C \|f\|_{H^{m,0}_\beta(D)} < \infty.$$

Proof. Since $a(\cdot, \boldsymbol{y}) \in W^{1,\infty}(D)$ and $\operatorname{ess\,inf} a(\cdot, \boldsymbol{y}) > 0$ there holds in $L^2(D)$

$$-\Delta u(\cdot, \boldsymbol{y}) = \frac{1}{a(\cdot, \boldsymbol{y})}(f + \nabla a(\cdot, \boldsymbol{y}) \nabla u(\cdot, \boldsymbol{y})).$$

Therefore, by Lemma 3.9, we get for $\theta_{\boldsymbol{y}} := \left\| \frac{1}{a(\cdot, \boldsymbol{y})} \right\|_{W^{m,\infty}(D)}$ that

$$\|\Delta u(\cdot, \boldsymbol{y})\|_{H^{m,0}_\beta(D)} \le C \theta_{\boldsymbol{y}} \left(\|f\|_{H^{m,0}_\beta(D)} + \|a(\cdot, \boldsymbol{y})\|_{W^{m+1,\infty}(D)} \|u(\cdot, \boldsymbol{y})\|_{H^{m+1,1}_\beta(D)} \right).$$

For all α with $|\alpha| \le m$, the multivariate Faà Di Bruno's formula as stated in [14, Corollary to Proposition 1 and 2] gives

$$D^\alpha \left(\frac{1}{a(\cdot, \boldsymbol{y})} \right) = \sum_{\mathcal{P} \in \Pi} M_{\mathcal{P}} \frac{(-1)^{|\mathcal{P}|} |\mathcal{P}|!}{a(\cdot, \boldsymbol{y})^{|\mathcal{P}|+1}} \prod_{\gamma \in \mathcal{P}} D^\gamma a(\cdot, \boldsymbol{y}),$$

where Π is the set of all partitions of the multiset (i.e. set with repetitions) $\{\underbrace{1, 1, \dots, 1}_{\alpha_1 \text{ times}}, \underbrace{2, 2, \dots, 2}_{\alpha_2 \text{ times}}\}$ and the $M_{\mathcal{P}} \in \mathbb{N}$ denote the multiplicities of the summands. In particular, $|\gamma| \le m$ so that the assumptions on $a(\cdot, \boldsymbol{y})$ imply that $D^\alpha \left(\frac{1}{a(\cdot, \boldsymbol{y})} \right) \in L^\infty(D)$ and $\sup_{\boldsymbol{y} \in U} \theta_{\boldsymbol{y}} < \infty$. It remains to show that for all $m \in \mathbb{N}_0, \boldsymbol{y} \in U$, there holds $\|u(\cdot, \boldsymbol{y})\|_{H^{m+1,1}_\beta(D)} \le C \|f\|_{H^{m,0}_\beta(D)}$. By the continuous embedding $L^2_\beta(D) \subseteq H^{-1}(D)$ we get

$$\begin{aligned} \|u(\cdot, \boldsymbol{y})\|^2_{H^{1,1}_\beta(D)} &\le \|u(\cdot, \boldsymbol{y})\|^2_{L^2(D)} + C \sum_{|\alpha|=1} \||D^\alpha u(\cdot, \boldsymbol{y})|\|^2_{L^2(D)} \\ &\le C \|u(\cdot, \boldsymbol{y})\|^2_{H^1_0(D)} \le C \|f\|^2_{H^{-1}(D)} \le C \|f\|^2_{L^2_\beta(D)}. \end{aligned}$$

This proves the case $m = 0$. For $m \geq 1$ induction gives

$$\begin{aligned}\|u(\cdot,\boldsymbol{y})\|_{H^{m+1,1}_\beta(D)} &\leq C\,\|u(\cdot,\boldsymbol{y})\|_{H^{m+1,2}_\beta(D)} \\ &\leq C\,\|\Delta u(\cdot,\boldsymbol{y})\|_{H^{m-1,0}_\beta(D)} \leq C\,\|f\|_{H^{m,0}_\beta(D)}\,,\end{aligned}$$

where we used Lemma 3.7 and Theorem 3.10. □

Recall that $B_{s,\boldsymbol{y}}$ denotes the bilinear form $B_{\boldsymbol{y}}$ from (3.7) but with truncated series for $j > s$, so that $B_{s,\boldsymbol{y}}(u_s(\cdot,\boldsymbol{y}), v) = f(v)$ for all $v \in H^1_0(D)$. The parametric Galerkin solution $u_{s,h}(\cdot,\boldsymbol{y}) \in X_h$ satisfies for all $\boldsymbol{y} \in U$,

$$B_{s,\boldsymbol{y}}(u_{s,h}(\cdot,\boldsymbol{y}), v_h) = f(v_h) \qquad \forall v_h \in X_h.$$

Since $X_h \subset H^1_0(D)$, Proposition 3.1 applies and there is a unique Galerkin solution satisfying

$$\sup_{0<h<h_0}\sup_{\boldsymbol{y}\in U}\|u_{s,h}(\cdot,\boldsymbol{y})\|_{H^1_0(D)} \leq \frac{\|f\|_{H^{-1}(D)}}{\bar{a}_{\min}(1-\bar{\kappa})}.$$

This in particular gives uniform stability of the numerical solutions with respect to both the parameters and the discretisation level. Moreover, there holds the Galerkin orthogonality property for all $\boldsymbol{y} \in U$,

$$B_{s,\boldsymbol{y}}(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}), v_h) = 0 \qquad \forall v_h \in X_h.$$

The following proposition includes a Céa estimate and the results above.

Proposition 3.12. *Let $f \in H^{m,0}_\beta(D)$ be satisfied for β as in Theorem 3.10 and assume that the conditions* (3.8), (3.10) *and* (3.26) *hold. Let $h_0 \in (0,+\infty)$ such that X_h satisfies the approximation property* (3.27) *for all $h \in (0,h_0)$. Then, if $0 < h < h_0$, there exist a positive constant C independent of h, f, s and $\boldsymbol{y}$ such that*

$$\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H^1_0(D)} \leq C\,\|f\|_{H^{m,0}_\beta(D)}\,h^{m+1}.$$

Proof. The continuity and coercivity estimates in the proof of Proposition 3.1 prove that, for all $v_h \in X_h$,

$$\begin{aligned}\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|^2_{H^1_0(D)} &\leq C B_{s,\boldsymbol{y}}\big(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}), u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\big) \\ &= C B_{s,\boldsymbol{y}}\big(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}), u_s(\cdot,\boldsymbol{y}) - v_h\big) \\ &\leq C\,\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H^1_0(D)}\|u_s(\cdot,\boldsymbol{y}) - v_h\|_{H^1_0(D)}\,.\end{aligned}$$

Thus, quasioptimality of the Galerkin approximation holds

$$\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\|_{H^1_0(D)} \leq C \inf_{v_h\in X_h}\|u_s(\cdot,\boldsymbol{y}) - v_h\|_{H^1_0(D)}\,.$$

Applying the condition (3.27), Theorem 3.10 and Proposition 3.11 we obtain the claim. □

From Proposition 3.12, it is straightforward to deduce that $\sup_{\boldsymbol{y}\in U}\big|G(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}))\big| \leq C\,\|G\|_{H^{-1}(D)}\,\|f\|_{H^{m,0}_\beta(D)}\,h^{m+1}$. However, using a Aubin-Nitsche duality argument it is possible to show faster convergence of this quantity, provided that G is more regular.

Theorem 3.13. *Let $f \in H_\beta^{m,0}(D), G \in H_\beta^{m',0}(D)$ be satisfied for some $m, m' \in \mathbb{N}_0$ and β as in Theorem 3.10. Assume that the conditions* (3.8),(3.10) *and* (3.26) *hold and let $h_0 \in (0, +\infty)$ be such that X_h satisfies the approximation property* (3.27) *for all $h \in (0, h_0)$. Then, if $0 < h < h_0$, there exist a positive constant C independent of h, f, G and $\boldsymbol{y}$ such that there holds*

$$\sup_{\boldsymbol{y}\in U} \left|G(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}))\right| \le C \left\|f\right\|_{H_\beta^{m,0}(D)} \left\|G\right\|_{H_\beta^{m',0}(D)} h^{m+m'+2}.$$

Proof. For all $\boldsymbol{y} \in U$, consider the parametric adjoint problem: given $G \in H_\beta^{m',0}(D)$, find $v_G(\cdot,\boldsymbol{y}) \in H_0^1(D)$ such that

$$B_{s,\boldsymbol{y}}(w, v_G(\cdot,\boldsymbol{y})) = G(w) \qquad \forall w \in H_0^1(D).$$

By the Lax-Milgram lemma there is a unique solution to this problem. For all $\boldsymbol{y} \in U$ and for all $v_h \in X_h$,

$$\begin{aligned}\left|G(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}))\right| &= \left|B_{s,\boldsymbol{y}}(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}), v_G(\cdot,\boldsymbol{y}))\right| \\ &= \left|B_{s,\boldsymbol{y}}(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}), v_G(\cdot,\boldsymbol{y}) - v_h)\right| \\ &\le C \left\|u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y})\right\|_{H_0^1(D)} \left\|v_G(\cdot,\boldsymbol{y}) - v_h\right\|_{H_0^1(D)}.\end{aligned}$$

The first term can be bounded by Proposition 3.12; for the second, we take the infimum over $v_h \in X_h$. By (3.27) and (3.10), we are left with

$$\left|G(u_s(\cdot,\boldsymbol{y}) - u_{s,h}(\cdot,\boldsymbol{y}))\right| \le C \left\|f\right\|_{H_\beta^{m,0}(D)} h^{m+m'+2} \left\|\Delta v_G(\cdot,\boldsymbol{y})\right\|_{H_\beta^{m,0}(D)}.$$

Since $B_{s,\boldsymbol{y}}$ is symmetric, $v_G(\cdot,\boldsymbol{y})$ solves (3.1) weakly with right hand side G. Therefore, Proposition 3.11 gives $\left\|\Delta v_G(\cdot,\boldsymbol{y})\right\|_{H_\beta^{m,0}(D)} \le C \left\|G\right\|_{H_\beta^{m',0}(D)}$ and the proof is complete. □

3.5 Combined QMC-FEM error analysis

The next theorem summarises the main results of this chapter, namely Proposition 3.2 and Theorems 3.6 and 3.13. We express the full integration error in terms of the truncation dimension s, the number N of QMC points and the dimension M_h of the FE space.

Theorem 3.14. *Let $D \subset \mathbb{R}^d$, $d \in \{1, 2\}$ be a bounded interval or a polygon. Let the smallness conditions* (3.8),(3.9) *be satisfied for a sequence $\boldsymbol{b} = (b_j)_{j\ge1} \in \ell^p(\mathbb{N})$ for some $p \in (0, 1)$ and $b_j \in (0, 1]$. Assume the regularity conditions* (3.25), (3.26) *for some β as in Theorem 3.10. Then, for sufficiently large s and M_h, an interlaced scrambled polynomial lattice rule of order $\lfloor \frac{1}{p} - \frac{1}{2} \rfloor + 1$ coupled with high order FEM realises the bound*

$$\begin{aligned}&\left\|I(G(u)) - I_s(G(u_{s,h}); P^{IS})\right\|_{L^2(U)} \\ &\qquad \le C \left(N^{-1/p} + M_h^{-(m+m'+2)/d} + \left(\sup_{j\ge s+1} b_j\right)^2 \right), \end{aligned} \tag{3.28}$$

where C is a positive constant independent of N, s, M_h and depends linearly on $\left\|f\right\|_{H_\beta^m(D)}$ and $\left\|G\right\|_{H_\beta^{m'}(D)}$.

This result gives a criteria to couple the values of N and M_h to obtain the error below some tolerance $O(\varepsilon)$. In particular, for given dimension d, summability p and $\tau := m + m' + 2$ we can choose

$$M_h^{\frac{\tau}{d}} \sim N^{\frac{1}{p}} = O(\varepsilon^{-1})$$

and the value of s accordingly. As a consequence, we must run a FEM solver with $M_h = O(\varepsilon^{-d/\tau})$ degrees of freedom $N = O(\varepsilon^{-p})$ times. Additionally, it will be shown in Section 4.1 that the CBC algorithm to compute the high order QMC rule requires $O(\alpha s N \log(N)) = O\Big((\lfloor \frac{1}{p} - \frac{1}{2} \rfloor + 1)\varepsilon^{-p}\log(\varepsilon^{-1})\Big)$ operations where the smoothness parameter α is related to p as in Theorem 3.6.

3.6 Multi-Level QMC

The error-vs-work analysis discussed previously, motivates a Multi-Level integration, which goal is to reduce the computational cost while keeping the convergence rate unchanged. This method can be understood as a progressive refinement of an initially coarse approximation. Each iteration, that is called level, involves a different choice of s, N and M_h that will be coupled in a suitable way. We denote by s_ℓ, N_ℓ and $M_\ell = M_{h_\ell}$ the respective parameters at level $\ell = 0, \dots, L$ where $L \in \mathbb{N}$ is the last level. For example, the computational cost can be controlled if N_ℓ is progressively decresed while increasing M_ℓ; this means that we evaluate less times the solution of the PDE on finer meshes. Similar arguments apply for the dimension parameters s_ℓ and we then assume that $s_\ell \geq s_{\ell-1}$. To explain the method, we define $u_\ell := u_{s_\ell, h_\ell}$ and we use the convention $u_{-1} \equiv 0$. Then we can write the telescopic sum

$$G(u_L) = \sum_{\ell=0}^{L} G(u_\ell - u_{\ell-1}).$$

Next, we apply different quadrature formulas Q_{s_ℓ, N_ℓ} to each summand. We denote by Q^L the resulting quadrature formula where each Q_{s_ℓ, N_ℓ} is chosen to be an interlaced scrambled polynomial lattice rule with N_ℓ points in s_ℓ dimensions:

$$Q^L(G(u_L)) := \sum_{\ell=0}^{L} Q_{s_\ell, N_\ell}(G(u_\ell - u_{\ell-1})). \tag{3.29}$$

Proposition 3.4 provides the conditions for boundedness of the integrand, but since we are interested in differences $G(u_\ell - u_{\ell-1})$, we want to extend some previous results to additionally obtain smallness of the integrands in $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}\left([-\frac{1}{2}, \frac{1}{2}]^s\right)$.

In Section 3.3, we used the dilated coordinates $T_{\boldsymbol{y}}(\boldsymbol{z})$ with respect to the same sequence $(b_j)_{j\geq 1}$ appearing in the sparsity assumption (3.9). As a consequence, we had in (3.17) a bound of the same form of (3.10). That smallness condition was sufficient in the case of single level QMC, while now we also need a sparsity condition similar to (3.9). Following the construction in [11, Section 4], we can dilate the coordinates with respect to the sequence $(\hat{b}_j)_{j\geq 1} := (b_j^{1-\theta})_{j\geq 1}$ for some $\theta \in [0, 1)$. If we repeat the computation in this setting, we obtain the same results with the choice of weights

$$\hat{\gamma}_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \sum_{\nu=1}^{\alpha} \left[\left(\frac{2\hat{b}_j}{1-\eta}\right)^{\nu} \nu!\right]^2, \tag{3.30}$$

provided that $\hat{\boldsymbol{b}} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, which is a stronger assumption if $\theta \in (0,1)$. On the other hand, the sparsity loss in the QMC error allows to obtain, for every $\boldsymbol{y} \in U$, that for $\bar{a}_{\boldsymbol{y}} := a(\cdot, \boldsymbol{y})$ and $\psi_{\boldsymbol{y},j} := \dfrac{\eta^{-1} - 2|y_j|}{\hat{b}_j}\psi_j$ there holds

$$\left\| \frac{\sum_{j\geq 1} |\psi_{\boldsymbol{y},j}|/b_j^\theta}{\bar{a}_{\boldsymbol{y}}} \right\|_{L^\infty(D)} \leq \frac{\kappa}{\eta} < 1. \tag{3.31}$$

In this section we denote by $\tilde{u}_{\boldsymbol{y}}$ the solution of (3.15) if we use $\hat{\boldsymbol{b}}$ as input for the dilation coordinate. Moreover, $\tilde{u}_{\boldsymbol{y},s}(\cdot, \boldsymbol{z}) = \tilde{u}_{\boldsymbol{y}}(\cdot, (z_1, \ldots, z_s, 0, 0, \ldots))$ denotes the corresponding solution with truncated diffusion coefficient and $\tilde{u}_{\boldsymbol{y},h_\ell}(\cdot, \boldsymbol{z})$ is the Galerkin solution. The new scaling permits to control the dimension truncation differences as in the following theorem, which proof is omitted here but can be found in [11, Theorem 1 and Remark 2].

Theorem 3.15. *Let the assumption* (3.8), (3.9) *be satisfied for the sequence* $\boldsymbol{b}$. *There exists a constant* $C < \infty$ *such that for every* $\boldsymbol{y} \in U$ *and for every* $s \in \mathbb{N}$ *and every* $\theta \in [0,1]$

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} \frac{1}{(\boldsymbol{\nu}!)^2} \left\| \partial_{\boldsymbol{z}}^{\boldsymbol{\nu}} \big(\tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) - \tilde{u}_{\boldsymbol{y},s}(\cdot, \boldsymbol{z}) \big) \big|_{\boldsymbol{z}=\boldsymbol{0}} \right\|_{H_0^1(D)}^2 \leq C \, \|f\|_{H^{-1}(D)}^2 \sup_{j>s} b_j^{2\theta}.$$

Moreover, the same estimate holds if we replace $\tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})$ *and* $\tilde{u}_{\boldsymbol{y},s}(\cdot, \boldsymbol{z})$ *by the respective Galerkin solutions* $\tilde{u}_{\boldsymbol{y},h}(\cdot, \boldsymbol{z})$ *and* $\tilde{u}_{\boldsymbol{y},s,h}(\cdot, \boldsymbol{z})$, *with constant* C *independent of* h.

Next, in order to bound Galerkin error differences, we apply the Leibniz rule to the equation $-\operatorname{div}(a \nabla u) = f$: for sufficiently smooth a, u and f there holds

$$-a \Delta D^\alpha u - \sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} D^\gamma a \Delta D^{\alpha-\gamma} u - \sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \nabla D^\gamma a \cdot \nabla D^{\alpha-\gamma} u = D^\alpha f. \tag{3.32}$$

We also need certain regularity of f and $\tilde{a}_{\boldsymbol{y}}$, that allows to apply the above formula for $a = \tilde{a}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})$ and $u = \tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})$, where equality is meant in $L^2_{\beta+|\alpha|}(D)$. In fact, it is sufficient to have $f \in H^{m,0}_\beta(D)$ and $\tilde{a}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) \in W^{m+1,\infty}(D)$. This can be seen as in [3, Lemma 4.3], with the only difference that the polygonal domain D requires Babuška-Kondrat'ev spaces in our setting. In terms of the initial diffusion coefficient $a(\cdot, \boldsymbol{y})$, it is sufficient to assume that all the ψ_j belong to $W^{m+1,\infty}(D)$ and that, for all α with $|\alpha| \leq m+1$,

$$\sup_{\boldsymbol{y} \in U} \|a(\cdot, \boldsymbol{y})\|_{W^{m+1,\infty}(D)} < \infty \quad \text{and} \quad \sum_{j \geq 1} \frac{|D^\alpha \psi_j|}{\hat{b}_j} \in L^\infty(D). \tag{3.33}$$

In particular, if we define

$$K_\alpha := \Bigg\| \sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} \left(|D^\gamma \bar{a}_{\boldsymbol{y}}| + \sum_{j \geq 1} |D^\gamma \psi_{\boldsymbol{y},j}| \right) + \sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \left(|\nabla D^\gamma \bar{a}_{\boldsymbol{y}}| + \sum_{j \geq 1} |\nabla D^\gamma \psi_{\boldsymbol{y},j}| \right) \Bigg\|_{L^\infty(D)},$$

we also get $K_\alpha < \infty$ for all α with $|\alpha| \leq m$. We are now ready to prove a bound on higher order derivatives of the Taylor coefficients (cp. Lemma 3.3).

Proposition 3.16. *Let $m \in \mathbb{N}_0, f \in H^{m,0}_\beta$. Assume that the diffusion coefficient of the PDE* (3.1) *satisfies* (3.33) *and that β is defined as in Theorem 3.10. Moreover, assume that conditions* (3.8), (3.9) *hold with $0 < b_j \leq 1$ for all $j \geq 1$. Then, for all $\alpha : |\alpha| \leq m$, there is a constant $C > 0$ independent of f and $\boldsymbol{y}$ such that*

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \frac{1}{(\boldsymbol{\nu}!)^2} \left\| \Delta \left(\partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})\big|_{\boldsymbol{z}=\boldsymbol{0}} \right) \right\|^2_{H^{m,0}_\beta(D)} \leq C \left\| f \right\|^2_{H^{m,0}_\beta(D)} .$$

Proof. Throughout the proof, C denotes a generic constant independent of f and $\boldsymbol{y}$ that can change within the same formula. Define $t_{\boldsymbol{y},\boldsymbol{\nu}} = \frac{1}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})\big|_{\boldsymbol{z}=\boldsymbol{0}}$. Then, It is sufficient to prove that for all α with $|\alpha| \leq m$, there holds

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \left\| \Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{L^2_{\beta+|\alpha|}(D)} \leq C \left\| f \right\|^2_{H^{m,0}_\beta(D)} . \tag{3.34}$$

For $|\alpha| = 0$, this follows from [11, Proposition 2]. Assuming that (3.34) holds for all γ with $|\gamma| < |\alpha|$, we will prove it for α, arguing similarly as in [3, Theorem 4.2]. Let $\phi_\alpha := r_D^{\beta+|\alpha|}$ and define for all $\boldsymbol{\nu} \in \mathcal{F}$ the quantities

$$c^\alpha_{\boldsymbol{\nu}} := \int_D \bar{a}_{\boldsymbol{y}} |\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi^2_\alpha \qquad \delta^\alpha_{\boldsymbol{\nu}} := \int_D |\nabla D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi^2_\alpha .$$

Being $\bar{a}_{\boldsymbol{y}} > 0$ a.e. in D, it is sufficient to show that $\sum_{\boldsymbol{\nu}\in\mathcal{F}} c^\alpha_{\boldsymbol{\nu}} \leq C \left\| f \right\|^2_{H^{m,0}_\beta(D)}$. The pointwise multiplication by ϕ_α is an isometric isomorphism from $L^2_{\beta+|\alpha|}(D)$ to $L^2(D)$. Therefore, (3.32) can be written as follows: for all $v \in L^2(D)$, there holds

$$\begin{aligned} \int_D (D^\alpha f) v \phi_\alpha = &- \int_D \left(\tilde{a}_{\boldsymbol{y}} \Delta D^\alpha \tilde{u}_{\boldsymbol{y}} \right) v \phi_\alpha - \int_D \left(\sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} D^\gamma \tilde{a}_{\boldsymbol{y}} \Delta D^{\alpha-\gamma} \tilde{u}_{\boldsymbol{y}} \right) v \phi_\alpha \\ &- \int_D \left(\sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \nabla D^\gamma \tilde{a}_{\boldsymbol{y}} \cdot \nabla D^{\alpha-\gamma} \tilde{u}_{\boldsymbol{y}} \right) v \phi_\alpha . \end{aligned}$$

We take $\frac{1}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}}_{\boldsymbol{z}}$ of this relation for some $\boldsymbol{\nu} \in \mathcal{F} \backslash \{\boldsymbol{0}\}$ and we evaluate at $\boldsymbol{z} = \boldsymbol{0}$. As a result, we have a recursive formula for the (derivatives of) Taylor coefficients,

given by

$$\begin{aligned}\int_D (\bar{a}_{\boldsymbol{y}} \Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}) v\phi_\alpha = &- \int_D \left(\sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} \psi_{\boldsymbol{y},j} \Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j} \right) v\phi_\alpha \\ &- \int_D \left(\sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} D^\gamma \bar{a}_{\boldsymbol{y}} \Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}} \right) v\phi_\alpha \\ &- \int_D \left(\sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \nabla D^\gamma \bar{a}_{\boldsymbol{y}} \cdot \nabla D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}} \right) v\phi_\alpha \\ &- \int_D \left(\sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} D^\gamma \psi_{\boldsymbol{y},j} \Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j} \right) v\phi_\alpha \\ &- \int_D \left(\sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} \nabla D^\gamma \psi_{\boldsymbol{y},j} \cdot \nabla D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j} \right) v\phi_\alpha.\end{aligned}$$

Next, Proposition 3.11 implies that $\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}} \in L^2_{\beta+|\alpha|}(D)$ so we can choose the test function $v = (\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}})\phi_\alpha$. If we apply Young's inequality, we then obtain that for all $\varepsilon > 0$, there holds

$$\begin{aligned}c_{\boldsymbol{\nu}}^\alpha \leq\ & \varepsilon \int_D \sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} |D^\gamma \bar{a}_{\boldsymbol{y}}| \Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 \\ &+ \frac{1}{4\varepsilon} \int_D \sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} |D^\gamma \bar{a}_{\boldsymbol{y}}| |\Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 \\ &+ \varepsilon \int_D \sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} |\nabla D^\gamma \bar{a}_{\boldsymbol{y}}| \Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 \\ &+ \frac{1}{4\varepsilon} \int_D \sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} |\nabla D^\gamma \bar{a}_{\boldsymbol{y}}| |\nabla D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 \\ &+ \frac{1}{2} \int_D \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} |\psi_{\boldsymbol{y},j}| |\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 + \frac{1}{2} \int_D \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} |\psi_{\boldsymbol{y},j}| |\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \phi_\alpha^2 \\ &+ \varepsilon \int_D \sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} |D^\gamma \psi_{\boldsymbol{y},j}| |\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 \\ &+ \frac{1}{4\varepsilon} \int_D \sum_{0 \neq \gamma \leq \alpha} \binom{\alpha}{\gamma} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} |D^\gamma \psi_{\boldsymbol{y},j}| |\Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \phi_\alpha^2 \\ &+ \varepsilon \int_D \sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} |\nabla D^\gamma \psi_{\boldsymbol{y},j}| |\Delta D^\alpha t_{\boldsymbol{y},\boldsymbol{\nu}}|^2 \phi_\alpha^2 \\ &+ \frac{1}{4\varepsilon} \int_D \sum_{\gamma \leq \alpha} \binom{\alpha}{\gamma} \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} |\nabla D^\gamma \psi_{\boldsymbol{y},j}| |\nabla D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}|^2 \phi_\alpha^2 \\ &=: T_1 + T_2 + \ldots + T_{10}.\end{aligned}$$

First, we apply (3.17) to T_5 and T_6, obtaining

$$\sum_{|\boldsymbol{\nu}|=k} (T_5 + T_6) \leq \frac{\kappa}{2\eta} \sum_{|\boldsymbol{\nu}|=k} c_{\boldsymbol{\nu}}^{\alpha} + \frac{\kappa}{2\eta} \sum_{|\boldsymbol{\nu}|=k-1} c_{\boldsymbol{\nu}}^{\alpha}. \tag{3.35}$$

Defining $K'_\alpha = \frac{K_\alpha}{\bar{a}_{\boldsymbol{y},\min}}$ (see (3.16)), we also have

$$\sum_{|\boldsymbol{\nu}|=k} (T_1 + T_3 + T_7 + T_9) \leq \varepsilon K'_\alpha \sum_{|\boldsymbol{\nu}|=k} c_{\boldsymbol{\nu}}^{\alpha}. \tag{3.36}$$

Analogously, applying that $r_D^{\alpha-\gamma}$ is bounded for all $\gamma \leq \alpha$, for all $\varepsilon > 0$ there is a finite constant C_ε depending also on K_α such that

$$\begin{aligned}\sum_{|\boldsymbol{\nu}|=k} (T_2 + T_4 + T_8 + T_{10}) \leq C_\varepsilon \sum_{0\neq\gamma\leq\alpha} \sum_{|\boldsymbol{\nu}|=k-1}^{k} c_{\boldsymbol{\nu}}^{\alpha-\gamma} \\ + C_\varepsilon \sum_{\gamma\leq\alpha} \sum_{|\boldsymbol{\nu}|=k-1}^{k} \delta_{\boldsymbol{\nu}}^{\alpha-\gamma}.\end{aligned} \tag{3.37}$$

Combining (3.35), (3.36) and (3.37) we get

$$\begin{aligned}\left(1 - \varepsilon K'_\alpha - \frac{\kappa}{2\eta}\right) \sum_{|\boldsymbol{\nu}|=k} c_{\boldsymbol{\nu}}^{\alpha} \leq \frac{\kappa}{2\eta} \sum_{|\boldsymbol{\nu}|=k-1} c_{\boldsymbol{\nu}}^{\alpha} + C_\varepsilon \sum_{0\neq\gamma\leq\alpha} \sum_{|\boldsymbol{\nu}|=k-1}^{k} c_{\boldsymbol{\nu}}^{\alpha-\gamma} \\ + C_\varepsilon \sum_{\gamma\leq\alpha} \sum_{|\boldsymbol{\nu}|=k-1}^{k} \delta_{\boldsymbol{\nu}}^{\alpha-\gamma}.\end{aligned}$$

We choose $\varepsilon > 0$ satisfying $\frac{\kappa}{2\eta} < 1 - \varepsilon K'_\alpha - \frac{\kappa}{2\eta}$, that is $\varepsilon < \frac{\eta - \kappa}{K'_\alpha \eta}$; thus, summing over $k \geq 1$ yields,

$$\sum_{\mathbf{0}\neq\boldsymbol{\nu}\in\mathcal{F}} c_{\boldsymbol{\nu}}^{\alpha} \leq C \left(c_{\mathbf{0}}^{\alpha} + \sum_{0\neq\gamma\leq\alpha} \sum_{\boldsymbol{\nu}\in\mathcal{F}} c_{\boldsymbol{\nu}}^{\alpha-\gamma} + \sum_{\gamma\leq\alpha} \sum_{\boldsymbol{\nu}\in\mathcal{F}} \delta_{\boldsymbol{\nu}}^{\alpha-\gamma} \right).$$

From $\bar{a}_{\boldsymbol{y}} \in L^\infty(D)$ we obtain the inequalities

$$\begin{aligned} c_{\mathbf{0}}^{\alpha} &\leq C \left\| u(\cdot, \boldsymbol{y}) \right\|^2_{H^{m+2,2}_\beta(D)} \leq C \left\| f \right\|^2_{H^{m,0}_\beta(D)}, \\ c_{\boldsymbol{\nu}}^{\alpha-\gamma} &\leq C \left\| \Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{L^2_{\beta+|\alpha-\gamma|}(D)}, \end{aligned}$$

where we also used Theorem 3.10 and Proposition 3.11. Finally, again Theorem 3.10 gives

$$\begin{aligned}\sum_{\gamma\leq\alpha} \delta_{\boldsymbol{\nu}}^{\alpha-\gamma} &\leq C \left\| t_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{H^{|\alpha|+1,1}_\beta(D)} \\ &\leq C \left\| \Delta t_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{H^{|\alpha|-1,0}_\beta(D)} \\ &= C \sum_{0\neq\gamma\leq\alpha} \left\| \Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{L^2_{\beta+|\alpha-\gamma|}(D)}.\end{aligned}$$

Therefore, the claim follows by the inductive hypothesis that, for all $0 \neq \gamma \leq \alpha$, there exists a constant $C > 0$ such that

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} \left\| \Delta D^{\alpha-\gamma} t_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{L^2_{\beta+|\alpha-\gamma|}(D)} \leq C \left\| f \right\|^2_{H^{m,0}_\beta(D)} .$$

□

Remark As in [11, Remark 3], we have that for any finite s, the truncated solution $\tilde{u}_{\boldsymbol{y},s}(\cdot, \boldsymbol{z})$ satisfies $\partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y},s}(\cdot, \mathbf{0}) = \partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y}}(\cdot, \mathbf{0})$ if $\operatorname{supp}(\boldsymbol{\nu}) \subseteq \{1 : s\}$ and $\partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y},s}(\cdot, \mathbf{0}) = 0$ else. Therefore, Proposition 3.16 also gives an upper bound for the truncated solutions because it only consists of more terms.

Once we have summability of higher derivatives of the Taylor coefficients, we can generalise [11, Proposition 3] and [11, Theorem 2], to have an upper bound on Galerkin error differences.

Proposition 3.17. *Let the assumption in* (3.8), (3.9), (3.33) *be satisfied for* $0 < \kappa < \eta < 1$ *and let* β *be as in Theorem 3.10. Then there exists a constant* $C > 0$ *such that for every* $\boldsymbol{y} \in U$, $f \in H^{m,0}_\beta(D)$ *and for every* $\ell = 0, \dots, L$ *there holds*

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} \frac{1}{(\boldsymbol{\nu}!)^2} \left\| \partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \left(\tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) - \tilde{u}_{\boldsymbol{y},h_\ell}(\cdot, \boldsymbol{z}) \right) \big|_{\boldsymbol{z}=\mathbf{0}} \right\|^2_{H^1_0(D)} \leq C M_\ell^{-2(m+1)/d} \left\| f \right\|^2_{H^{m,0}_\beta(D)} .$$

Moreover the same estimate holds if we replace $\tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z})$ *and* $\tilde{u}_{\boldsymbol{y},h_\ell}(\cdot, \boldsymbol{z})$ *by the truncated solutions* $\tilde{u}_{\boldsymbol{y},s_\ell}(\cdot, \boldsymbol{z})$ *and* $\tilde{u}_{\boldsymbol{y},h_\ell,s_\ell}(\cdot, \boldsymbol{z})$, *respectively, with constant* C *independent of* s_ℓ.

Proof. We show the first claim following the steps of [11, Proposition 3], while the second follows along the lines of the remark above. For any $\boldsymbol{y} \in U$ and $\boldsymbol{\nu} \in \mathcal{F}$ we define the Taylor coefficient as

$$t_{\boldsymbol{y},\boldsymbol{\nu}} := \frac{1}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y}}(\cdot, \boldsymbol{z}) \big|_{\boldsymbol{z}=\mathbf{0}} \qquad \text{and} \qquad t^\ell_{\boldsymbol{y},\boldsymbol{\nu}} := \frac{1}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} \tilde{u}_{\boldsymbol{y},h_\ell}(\cdot, \boldsymbol{z}) \big|_{\boldsymbol{z}=\mathbf{0}}.$$

The recursive formula of Taylor coefficients (see proof of Lemma 3.3) implies that $\forall v \in X_\ell$

$$\int_D \bar{a}_{\boldsymbol{y}} \nabla (t_{\boldsymbol{y},\boldsymbol{\nu}} - t^\ell_{\boldsymbol{y},\boldsymbol{\nu}}) \cdot \nabla v = - \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} \int_D \psi_{\boldsymbol{y},j} \nabla (t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j} - t^\ell_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j}) \cdot \nabla v,$$

where $X_\ell := X_{h_\ell}$ is a subspace of $H^1_0(D)$ that satisfies the approximation property (3.27). For any $\boldsymbol{y} \in U$ we define the dilated Galerkin projection $\mathcal{P}_{\boldsymbol{y},\ell} : H^1_0 \to X_\ell$ via the relation

$$\int_D \bar{a}_{\boldsymbol{y}} \nabla (w - \mathcal{P}_{\boldsymbol{y},\ell} w) \cdot \nabla v = 0 \qquad \forall v \in X_\ell.$$

If we test the recursion formula with $v = \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^\ell_{\boldsymbol{y},\boldsymbol{\nu}}) \in X_\ell$, Cauchy-Schwarz

inequality and (3.17) yield, for all $k \geq 1$,

$$\begin{aligned}
&\sum_{|\boldsymbol{\nu}|=k} \int_D \bar{a}_{\boldsymbol{y}} |\nabla \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}})|^2 \\
&\quad \leq \sum_{|\boldsymbol{\nu}|=k} \int_D \bar{a}_{\boldsymbol{y}} \nabla(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}}) \cdot \nabla \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}}) \\
&\quad \leq \frac{1}{2} \sum_{|\boldsymbol{\nu}|=k} \int_D \sum_{j \in \mathrm{supp}(\boldsymbol{\nu})} |\psi_{\boldsymbol{y},j}| \left(|\nabla(t_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}-\boldsymbol{e}_j})|^2 + |\nabla \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}})|^2 \right) \\
&\quad \leq \frac{\kappa}{2\eta} \sum_{|\boldsymbol{\nu}|=k-1} \int_D \bar{a}_{\boldsymbol{y}} |\nabla(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}})|^2 + \frac{\kappa}{2\eta} \sum_{|\boldsymbol{\nu}|=k} \int_D \bar{a}_{\boldsymbol{y}} |\nabla \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}})|^2.
\end{aligned}$$

Define the energy norm $\|u\|^2_{\bar{a}_{\boldsymbol{y}}} := \int_D \bar{a}_{\boldsymbol{y}} |\nabla u|^2$ for all $u \in H^1_0(D)$, hence we can rewrite the above estimate as

$$\sum_{|\boldsymbol{\nu}|=k} \left\| \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}}) \right\|^2_{\bar{a}_{\boldsymbol{y}}} \leq \frac{\kappa}{2\eta - \kappa} \sum_{|\boldsymbol{\nu}|=k-1} \left\| t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{\bar{a}_{\boldsymbol{y}}}. \tag{3.38}$$

Let $\mathcal{I}$ denote the identity operator in $H^1_0(D)$. By triangular inequality and Young's inequality, there holds

$$\begin{aligned}
\sum_{|\boldsymbol{\nu}|=k} \left\| t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{\bar{a}_{\boldsymbol{y}}} &\leq \sum_{|\boldsymbol{\nu}|=k} \left(\left\| \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}}) \right\|_{\bar{a}_{\boldsymbol{y}}} + \|(\mathcal{I} - \mathcal{P}_{\boldsymbol{y},\ell}) t_{\boldsymbol{y},\boldsymbol{\nu}}\|_{\bar{a}_{\boldsymbol{y}}} \right)^2 \\
&\leq (1+\varepsilon) \sum_{|\boldsymbol{\nu}|=k} \left\| \mathcal{P}_{\boldsymbol{y},\ell}(t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}}) \right\|^2_{\bar{a}_{\boldsymbol{y}}} \\
&\quad + \left(1 + \frac{1}{\varepsilon}\right) \sum_{|\boldsymbol{\nu}|=k} \|(\mathcal{I} - \mathcal{P}_{\boldsymbol{y},\ell}) t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}}.
\end{aligned}$$

Thus, summing over $k \geq 1$, (3.38) allows to deduce

$$\begin{aligned}
\sum_{\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}} \left\| t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{\bar{a}_{\boldsymbol{y}}} &\leq \frac{(1+\varepsilon)\kappa}{2\eta - \kappa} \sum_{\boldsymbol{\nu} \in \mathcal{F}} \left\| t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{\bar{a}_{\boldsymbol{y}}} \\
&\quad + \left(1 + \frac{1}{\varepsilon}\right) \sum_{\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}} \|(\mathcal{I} - \mathcal{P}_{\boldsymbol{y},\ell}) t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}}.
\end{aligned}$$

Since $\kappa < \eta < 1$ there is an $\varepsilon > 0$ small with $\dfrac{(1+\varepsilon)\kappa}{2\eta - \kappa} < 1$; for such choice there is a constant $C < \infty$ satisfying

$$\sum_{\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}} \left\| t_{\boldsymbol{y},\boldsymbol{\nu}} - t^{\ell}_{\boldsymbol{y},\boldsymbol{\nu}} \right\|^2_{\bar{a}_{\boldsymbol{y}}} \leq C \left(\left\| t_{\boldsymbol{y},\mathbf{0}} - t^{\ell}_{\boldsymbol{y},\mathbf{0}} \right\|^2_{\bar{a}_{\boldsymbol{y}}} + \sum_{\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}} \|(\mathcal{I} - \mathcal{P}_{\boldsymbol{y},\ell}) t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{\bar{a}_{\boldsymbol{y}}} \right).$$

Recall that the energy norm and the $H^1_0(D)$ norm are equivalent. Therefore, it remains to bound $\left\| t_{\boldsymbol{y},\mathbf{0}} - t^{\ell}_{\boldsymbol{y},\mathbf{0}} \right\|^2_{H^1_0(D)}$ and $\sum_{\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}} \|(\mathcal{I} - \mathcal{P}_{\boldsymbol{y},\ell}) t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{H^1_0(D)}$. For the first term we apply the steps of Proposition 3.12 (with $B_{\boldsymbol{y}}$ instead of $B_{s,\boldsymbol{y}}$).

For the second, a similar reasoning and an application of Propostion 3.16 allow to conclude

$$\sum_{\mathbf{0}\neq\boldsymbol{\nu}\in\mathcal{F}} \|(\mathcal{I}-\mathcal{P}_{\boldsymbol{y},\ell})t_{\boldsymbol{y},\boldsymbol{\nu}}\|^2_{H^1_0(D)} \leq C h_\ell^{2(m+1)} \|f\|^2_{H^{m,0}_\beta(D)}\,.$$

Since $M_\ell = O(h_\ell^{-d})$ (see 48), the proof is complete. □

Theorem 3.18. *Let the assumption in* (3.8), (3.9), (3.33) *be satisfied and let* β *be as in Theorem 3.10. Then there exists a constant* $C>0$ *such that for every* $\boldsymbol{y}\in U$, $f\in H^{m,0}_\beta(D)$, $G\in H^{m',0}_\beta(D)$ *and for every* $\ell=0,\dots,L$ *there holds*

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \frac{1}{(\boldsymbol{\nu}+\mathbf{1})!\boldsymbol{\nu}!} \left|\partial^{\boldsymbol{\nu}}_{\boldsymbol{z}} G\big(\tilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z}) - \tilde{u}_{\boldsymbol{y},h_\ell}(\cdot,\boldsymbol{z})\big)\big|_{\boldsymbol{z}=\mathbf{0}}\right|^2 \leq C M_\ell^{-2(m+m'+2)/d} \|f\|^2_{H^{m,0}_\beta(D)} \|G\|^2_{H^{m',0}_\beta(D)}\,.$$

Moreover the same estimate holds if we replace $\tilde{u}_{\boldsymbol{y}}(\cdot,\boldsymbol{z})$ *and* $\tilde{u}_{\boldsymbol{y},h_\ell}(\cdot,\boldsymbol{z})$ *by the truncated solutions* $\tilde{u}_{\boldsymbol{y},s_\ell}(\cdot,\boldsymbol{z})$ *and* $\tilde{u}_{\boldsymbol{y},h_\ell,s_\ell}(\cdot,\boldsymbol{z})$ *respectively.*

Proof. The proof follows by the Aubin-Nitsche duality argument along the same lines of [11, Theorem 2] and using Proposition 3.17. □

We are now ready to present the main result of the Thesis. By coupling high order FEM with high order QMC integration, we construct a Multi-Level QMC quadrature rule using interlaced scrambled lattice points. As a result, the error decay of the single level QMC is preserved, while the computational cost is reduced consistently. First, we summarise many of the results obtained so far in one statement, to control the error in terms of the number of integration points, the dimension of the FE space and the dimension truncation of the integral.

Theorem 3.19. *Let* $L,\alpha\in\mathbb{N}, p,\theta\in(0,1)$ *satisfy that* $\alpha\geq\lfloor\frac{1}{p}-\frac{1}{2}\rfloor+1$ *and* $\theta<1-p$. *Let* $\boldsymbol{\gamma},\hat{\boldsymbol{\gamma}}$ *be defined as in* (3.20) *and* (3.30) *respectively. Assume that* (3.8) *and* (3.9) *are satisfied for a sequence* $\boldsymbol{b}\in\ell^p(\mathbb{N})$ *with* $b_j\in(0,1]$ *and assume that* (3.33) *is satisfied for* $\hat{\boldsymbol{b}} := ((\alpha+1)b_j^{1-\theta})_{j\geq1}$. *Let* $f\in H^{m,0}_\beta(D), G\in H^{m',0}_\beta(D)$ *and* $\sup_{\boldsymbol{y}\in U}\|a(\cdot,\boldsymbol{y})\|_{W^{m+1,\infty}(D)}<\infty$, *where* β *was defined in the statement of Theorem 3.10. Then, there exists a Multi-Level interlaced scrambled polynomial lattice rule* Q^L *of order* $\lfloor\frac{1}{p}-\frac{1}{2}\rfloor+1$ *with* $s_\ell\geq s_{\ell-1}, M_\ell\geq M_{\ell-1}$ *and* $N_\ell\leq N_{\ell-1}$ *for all* $\ell=1,\dots,L$, *satisfying*

$$\begin{aligned}\big\|I(G(u))-Q^L(G(u_L))\big\|_{L^2(U)} \leq\ & C_1\left(\sup_{j>s_L} b_j^2 + M_L^{-\frac{m+m'+2}{d}} + N_0^{-1/p}\right)\\ & + C_1\sum_{\ell=1}^{L} N_\ell^{-\frac{1-\theta}{p}}\left(M_{\ell-1}^{-\frac{m+m'+2}{d}} + \delta_{s_\ell,s_{\ell-1}}\sup_{j>s_{\ell-1}} b_j^\theta\right),\end{aligned}$$

where $C_1 := C\|G\|_{H^{m',0}_\beta(D)}\|f\|_{H^{m,0}_\beta(D)}$, *for some finite constant* C *independent of* $s_\ell, N_\ell, M_\ell, f$ *and* G.

Proof. In this proof we omit the domain of the weighted spaces $\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}(\cdot)$, which we assume fixed equal to $[-\frac{1}{2},\frac{1}{2}]^s$. Since for all $\ell=1,\dots,L$, $\otimes_{j=s_{\ell-1}}^{s_\ell}[-\frac{1}{2},\frac{1}{2}]$ has Lebesgue measure 1 and there holds $u_\ell(\cdot,\boldsymbol{y}) = u_{h_\ell}(\cdot,(y_1,\dots,y_{s_\ell},0,0,\dots))$, an

application of Fubini's theorem implies that $I_{s_\ell}(G(u_{\ell-1})) = I_{s_{\ell-1}}(G(u_{\ell-1}))$; hence

$$I(G(u)) - Q^L(G(u_L)) = I(G(u)) - I_{s_L}(G(u_L)) + \sum_{\ell=0}^{L} \Big[I_{s_\ell}(G(u_\ell - u_{\ell-1})) - Q_{s_\ell,N_\ell}(G(u_\ell - u_{\ell-1})) \Big].$$

Note that $\boldsymbol{b} \in \ell^p(\mathbb{N})$ implies that $\hat{\boldsymbol{\gamma}} \in \ell^{\frac{p}{1-\theta}}(\mathbb{N})$, with $\frac{p}{1-\theta} \in (0,1)$ by assumption. Triangular inequality and (3.4), together with Theorem 2.7 (applied with $d = \alpha$) give

$$\begin{aligned}
\big\|I(G(u)) - Q^L(G(u_L))\big\|_{L^2(U)} &\le |I(G(u)) - I_{s_L}(G(u_L))| \\
&\quad + \sum_{\ell=0}^{L} \|I_{s_\ell}(G(u_\ell - u_{\ell-1})) - Q_{s_\ell,N_\ell}(G(u_\ell - u_{\ell-1}))\|_{L^2(U)} \\
&\le |I(G(u)) - I_{s_L}(G(u_{s_L}))| + \sup_{\boldsymbol{y}\in U} |(G(u_{s_L} - u_L))| \\
&\quad + CN_0^{-1/p} \|G(u_0)\|_{\mathcal{W}_{s_0,\boldsymbol{\gamma},\alpha}} + \sum_{\ell=0}^{L} CN_\ell^{-(1-\theta)/p} \|G(u_\ell - u_{\ell-1})\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}}.
\end{aligned}$$

We note that, by Proposition 3.5, the constant C is independent of ℓ. Since b_j decay to zero, we can choose $\tilde{s}$ the smallest integer such that for all $s_L =: s \ge \tilde{s}$, (3.12) holds. If instead $s_L < \tilde{s}$, we can apply (3.11) to obtain similarly

$$\begin{aligned}
|I(G(u)) - I_{s_L}(G(u_{s_L}))| &\le C \|G\|_{H^{m'}_\beta(D)} \|f\|_{H^m_\beta(D)} \left(\sup_{j>s_L} b_j \cdot \max_{j=s_L,\dots,\tilde{s}} b_j \right) \\
&= C \|G\|_{H^{m'}_\beta(D)} \|f\|_{H^m_\beta(D)} \left(\sup_{j>s_L} b_j^2 \right).
\end{aligned}$$

Next, we apply Theorem 3.13 to bound

$$\sup_{\boldsymbol{y}\in U} |(G(u_{s_L} - u_L))| \le C \|G\|_{H^{m'}_\beta(D)} \|f\|_{H^m_\beta(D)} M_L^{-\frac{\tau}{d}},$$

where we define $\tau := m + m' + 2$. Moreover, Lemma 3.3 holds also for the Taylor coefficients of the FEM solution u_{h_0}, because only the variational formulation is used in its proof. Therefore, Proposition 3.4 can be applied when $\ell = 0$ to obtain

$$\|G(u_{s_0,h_0})\|_{\mathcal{W}_{s_0,\boldsymbol{\gamma},\alpha}} \le C \|G\|_{H^{-1}(D)} \|f\|_{H^{-1}(D)}.$$

Following the steps of [11, Theorem 3], we split the Galerkin differences from the dimension truncation in the weighted space norm:

$$\begin{aligned}
&\|G(u_\ell - u_{\ell-1})\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}} \\
&\quad \le \big\|G(u_{s_\ell,h_\ell} - u_{s_\ell,h_{\ell-1}})\big\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}} + \big\|G(u_{s_\ell,h_{\ell-1}} - u_{s_{\ell-1},h_{\ell-1}})\big\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}}.
\end{aligned}$$

Both terms can be treated as in the proof of Proposition 3.4. In particular,

Theorem 3.13 implies

$$\begin{aligned}\left\|G(u_{s_\ell,h_\ell}-u_{s_\ell,h_{\ell-1}})\right\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}} &\le \|G(u_{s_\ell}-u_{s_\ell,h_\ell})\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}} \\ &\quad + \left\|G(u_{s_\ell}-u_{s_\ell,h_{\ell-1}})\right\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}} \\ &\le C\,\|f\|_{H^{m,0}_\beta(D)}\,\|G\|_{H^{m',0}_\beta(D)}\left(M_\ell^{-\frac{\tau}{d}}+M_{\ell-1}^{-\frac{\tau}{d}}\right) \\ &\le C\,\|f\|_{H^{m,0}_\beta(D)}\,\|G\|_{H^{m',0}_\beta(D)}\,M_{\ell-1}^{-\frac{\tau}{d}},\end{aligned}$$

where the coefficient $(\alpha+1)$ in the definition of $\hat{\boldsymbol{b}}$ compensates for the factor $\frac{1}{(\boldsymbol{\nu}+\mathbf{1})!}$ in Theorem 3.13. Similar arguments, but applying instead Theorem 3.15, yield

$$\left\|G(u_{s_\ell,h_{\ell-1}}-u_{s_{\ell-1},h_{\ell-1}})\right\|_{\mathcal{W}_{s_\ell,\hat{\boldsymbol{\gamma}},\alpha}} \le C\delta_{s_\ell,s_{\ell-1}}\,\|G\|_{H^{-1}(D)}\,\|f\|_{H^{-1}(D)}\sup_{j>s_{\ell-1}} b_j^\theta$$

where $\delta_{i,j}$ denotes the Kronecker's delta. The proof is hence complete. □

3.7 Error vs. work analysis of Multi-Level QMC

Similarly as in Section 3.5, we discuss how to match the values of $(N_\ell)_{\ell=0,\dots,L}$, $(M_\ell)_{\ell=0,\dots,L}$ and $(s_\ell)_{\ell=0,\dots,L}$ in the Multi-Level setting, for fixed L. We want to control the total work necessary given a tolerance $O(\varepsilon)$ for the error, extending to high order the analysis in [12, Section 8] and [11, Section 6]. First, we assume that the fluctuation functions $(\psi_j)_{j\ge1}$ define a multiresolution analysis (MRA) in $L^2(D)$, $D\subset\mathbb{R}^d$ and verify the smallness and sparsity condition of the fluctuations in this case. In particular, we take a function $\psi\in W^{m+1,\infty(D)}$ with $\|\psi\|_{L^{\infty(D)}}=1$ and compact support, such that all the ψ_j can be obtained by scaling and translation of ψ: for some $\sigma,\rho>0$ to be determined later, for all $l\in\mathbb{N}_0$ and suitable $x_0,\dots,x_{\bar{k}_l}\in\mathbb{R}^d$, we define

$$\psi_{l,k}(x) := \sigma 2^{-\rho l}\psi(2^l(x-x_k)).$$

We also ask that, for each resolution level $l\in\mathbb{N}_0$, there holds $\bar{k}_l=O(2^{dl})<\infty$. Hence, we have a well-defined bijective enumeration $(l,k)\mapsto j\in\mathbb{N}$ given by the lexicographic order and we can identify $\psi_j:=\psi_{l,k}$. As a consequence, choosing $\forall\ell=0,\dots,L$, $s_\ell=2^{\bar{l}+1}-1$ for some $\bar{l}$, we are including all and only the fluctuations $\psi_{l,k}$ up to resolution level $\bar{l}$. In order to control the overlap of the fluctuations, we also assume that the points x_k satisfy that there exists a finite $K\in\mathbb{N}$ such that, for all $x\in D$ and $l\in\mathbb{N}_0$, there holds $|\{k:\psi_{l,k}(x)=0\}|\le K$. Therefore, we can enforce the condition (3.10) as follows:

$$\left\|\frac{\sum_{l\ge0}\sum_k|\psi_{l,k}|}{2\bar{a}}\right\|_{L^\infty(D)} \le \frac{\sigma K}{2\bar{a}_{\min}}\sum_{l\ge0}2^{-\rho l}=\frac{\sigma K}{2\bar{a}_{\min}}\frac{2^\rho}{2^\rho-1}=\bar{\kappa},$$

if we choose $\sigma:=\dfrac{2(2^\rho-1)\bar{a}_{\min}\bar{\kappa}}{2^\rho K}$. Next, we wish to find a sequence $(b_j)_{j\ge1}\in\ell^p(\mathbb{N})$, for $p\in(0,1)$, such that $b_j\in(0,1]$ and (3.9) holds. Let $\eta,c>0$ be parameters to be determined later and define, for all $l\in\mathbb{N}_0$ and for all $j\in\{2^l,2^l+1,\dots,2^{l+1}-1\}$

$$b_j := (1+c2^{\eta l})^{-1}<1.$$

Then we can enforce (3.9) as follows:

$$\left\| \frac{\sum_{l\geq 0}\sum_k |\psi_{l,k}|(1+c2^{\eta l})}{2\bar{a}} \right\|_{L^\infty(D)} \leq \bar{\kappa} + \frac{c}{2\bar{a}_{\min}} \left\| \sum_{l\geq 0} 2^{\eta l} \sum_k |\psi_{l,k}| \right\|_{L^\infty(D)}$$
$$\leq \bar{\kappa} + \frac{cK\sigma}{2\bar{a}_{\min}} \frac{1}{1-2^{\eta-\rho}} = \kappa,$$

where we impose the constraint $\rho > \eta$ for the convergence of the geometric series and we define $c := \dfrac{2(\kappa-\bar{\kappa})(1-2^{\eta-\rho})\bar{a}_{\min}}{K\sigma}$. Note that, sparsity of the b_j is implied by $\eta > d/p$, because

$$\sum_{j\geq 1} b_j^p \sim \sum_{l\geq 0} 2^{(d-p\eta)l} < \infty.$$

Finally, it remains to verify (3.33). Since $D^\alpha \psi_j = \sigma 2^{(|\alpha|-\rho)l} D^\alpha \psi$ and $\psi \in W^{m+1,\infty}(D)$, we obtain that there exists a constant $C > 0$ such that, for all α with $0 \leq |\alpha| \leq m+1$, there holds

$$\left\| \sum_{j\geq 1} \frac{|D^\alpha \psi_j|}{(\alpha+1)b_j^{1-\theta}} \right\|_{L^\infty(D)} \leq CK\sigma \sum_{l\geq 0} 2^{(m+1-\rho+\eta(1-\theta))l}.$$

The right hand side is finite if and only if $m+1-\rho+\eta(1-\theta) < 0$, but since we require $1-\theta > p$ (cp. Theorem 3.19), we need to choose $\theta \in \left(\frac{1+m-\rho}{\eta}+1, 1-p\right)$. This is possible if we pick $\eta > d/p$ and $\rho > \max(m+1+\eta p, \eta)$.

When we evaluate the work below, we assume that the QMC lattices of N_ℓ points in s_ℓ dimensions are already available for all levels. Given a QMC point $\boldsymbol{y} \in [-\frac{1}{2}, \frac{1}{2}]^{s_\ell}$ and $x \in D$, we have that

$$|\{j : j \leq s_\ell : \psi_j(x) \neq 0\}| \leq K \log(s_\ell).$$

Thus, in order to assemble the stiffness matrix for one QMC point, we need $O(M_\ell \log(s_\ell))$ operations. Moreover, we assume that the solution of the linear FE system can be done in $O(M_\ell)$ operations using sparse matrices, even for high order FEM (i.e. the order only affects the constant hidden in the $O(\cdot)$ notation). This is repeated N_ℓ times for each level, so that the total work is asymptotically

$$\text{work} = O\left(\sum_{\ell=0}^{L} N_\ell M_\ell \log(s_\ell)\right).$$

We are now ready to choose the parameters N_ℓ, s_ℓ and M_ℓ. The FE mesh that we need for the approximation property (3.27) is not quasi-uniform, but it can be also progressively refined at each level: if $d = 1$, we split each interval in 2 parts (not necessarily with the same lenght) and, if $d = 2$ we split each triangle in 4 parts as shown in [6, Section 5]. Thus, it is meaningful to assume $M_\ell \sim 2^{d\ell}$, for $\ell = 0, \ldots, L$. Since $b_j \sim j^{-\eta/d}$, the estimate in Theorem 3.19 suggests to choose

$$s_L \sim 2^{d\lceil L\tau/(2\eta)\rceil},$$
$$s_\ell \sim \min(2^{d\lceil \ell\tau/(\theta\eta)\rceil}, s_L) \qquad \forall \ell = 0, \ldots, L-1 \quad ,$$

where $\tau := m + m' + 2$. With these choices, the error takes the form

$$\text{error} = O\left(M_L^{-\tau/d} + N_0^{-\frac{1}{p}} + \sum_{\ell=1}^{L} M_{\ell-1}^{-\tau/d} N_\ell^{-\frac{1-\theta}{p}}\right).$$

Note that we allow different QMC order for the first level (that is $1/p$), and for the rest of levels (that is $(1-\theta)/p$). Adapting the steps in [11, 22], we minimise error vs work by finding stationary points of the Lagrangian $g(\lambda)$ with respect to the variables N_ℓ, where

$$g(\lambda) := M_L^{-\tau/d} + N_0^{-\frac{1}{p}} + \sum_{\ell=1}^{L} M_{\ell-1}^{-\tau/d} N_\ell^{-\frac{1-\theta}{p}} + \lambda \sum_{\ell=0}^{L} N_\ell M_\ell \log(s_\ell).$$

In particular, $\frac{\partial g(\lambda)}{\partial N_0} = 0$ induces $\lambda = \frac{N_0^{-1/p-1}}{pM_0 \log(s_0)}$. On the other hand, $\frac{\partial g(\lambda)}{\partial N_\ell} = 0$ implies that natural choices for N_ℓ, $\ell \geq 1$ are

$$N_\ell = \left\lceil N_0^{\frac{1+p}{1-\theta+p}} \left(\frac{(1-\theta) M_{\ell-1}^{-\tau/d} M_0 \log(s_0)}{M_\ell \log(s_\ell)}\right)^{\frac{p}{1-\theta+p}} \right\rceil.$$

Since $M_\ell \sim M_{\ell-1}$, for all $\ell \geq 1$ and M_0, s_0 are constants, we get

$$\text{error} = O\left(M_L^{-\tau/d} + N_0^{-\frac{1}{p}} + N_0^{-\frac{1}{p}+\frac{\theta}{1-\theta+p}} \sum_{\ell=1}^{L} E_\ell\right), \tag{3.39}$$

$$\text{work} = O\left(N_0 + N_0^{\frac{1+p}{1-\theta+p}} \sum_{\ell=1}^{L} E_\ell\right), \tag{3.40}$$

where $E_\ell := \left(\log(s_\ell) M_\ell^{1-\frac{\tau p}{d(1-\theta)}}\right)^{\frac{1-\theta}{1-\theta+p}}$. It remains to determine N_0. To this end, we note that, for all $0 \neq r_1 \in \mathbb{R}$ and $r_2 > 0$, there holds

$$\sum_{\ell=0}^{L} 2^{r_1 \ell} \ell^{r_2} \leq \frac{2^{r_1(L+1)} - 1}{2^{r_1} - 1} L^{r_2}.$$

Thus, since $\log(s_\ell) = O(\ell)$, we can deduce

$$\sum_{\ell=1}^{L} E_\ell \sim \sum_{\ell=0}^{L} 2^{\ell \frac{d(1-\theta)-\tau p}{1-\theta+p}} \ell^{\frac{1-\theta}{1-\theta+p}} = \begin{cases} O(1) & \text{if } d(1-\theta) < \tau p \\ O\left(L^{1+\frac{1-\theta}{1-\theta+p}}\right) & \text{if } d(1-\theta) = \tau p \\ O\left(2^{L\frac{d(1-\theta)-\tau p}{1-\theta+p}} L^{\frac{1-\theta}{1-\theta+p}}\right) & \text{if } d(1-\theta) > \tau p. \end{cases}$$

This implies that the last term of the error in (3.39) is always larger than $N_0^{-\frac{1}{p}}$, independently of the value of θ admissible. Analogously, the work for first level

N_0 is always dominated by the other terms. Therefore,

$$\begin{aligned}\text{error} &= O\left(M_L^{-\tau/d} + N_0^{-\frac{1}{p}+\frac{\theta}{1-\theta+p}} \sum_{\ell=1}^{L} E_\ell\right) \\ &= \begin{cases} O\left(2^{-\tau L} + N_0^{-\frac{1}{p}+\frac{\theta}{1-\theta+p}}\right) & \text{if } d(1-\theta) < \tau p \\ O\left(2^{-\tau L} + N_0^{-\frac{1}{p}+\frac{\theta}{1-\theta+p}} L^{1+\frac{1-\theta}{1-\theta+p}}\right) & \text{if } d(1-\theta) = \tau p \\ O\left(2^{-\tau L} + N_0^{-\frac{1}{p}+\frac{\theta}{1-\theta+p}} 2^{L\frac{d(1-\theta)-\tau p}{1-\theta+p}} L^{\frac{1-\theta}{1-\theta+p}}\right) & \text{if } d(1-\theta) > \tau p. \end{cases}\end{aligned}$$

The choices of N_0, leading to the overall error of $O(2^{-\tau L})$, are determined by

$$N_0 = \begin{cases} \left\lceil 2^{\frac{\tau L p(1-\theta+p)}{(1+p)(1-\theta)}} \right\rceil & \text{if } d(1-\theta) < \tau p \\ \left\lceil (2^{\tau L} L)^{\frac{p(1-\theta+p)}{(1+p)(1-\theta)}} L^{\frac{p}{1+p}} \right\rceil & \text{if } d(1-\theta) = \tau p \\ \left\lceil 2^{\frac{Lp(\tau+d)}{1+p}} L^{\frac{p}{1+p}} \right\rceil & \text{if } d(1-\theta) > \tau p. \end{cases}$$

By (3.40), the work necessary is

$$\text{work} = \begin{cases} O\left(2^{\frac{\tau L p}{(1-\theta)}}\right) & \text{if } d(1-\theta) < \tau p \\ O\left(2^{\frac{\tau L p}{(1-\theta)}} L^{2+\frac{p}{(1-\theta)}}\right) & \text{if } d(1-\theta) = \tau p \\ O\left(2^{Ld} L\right) & \text{if } d(1-\theta) > \tau p. \end{cases}$$

In conclusion for a prescribed accuracy $\varepsilon \sim 2^{-\tau L}$, we constructed a Multi-Level QMC FEM algorithm that realises error $= O(\varepsilon)$ and requires work

$$\text{work} = \begin{cases} O\left(\varepsilon^{-\frac{p}{(1-\theta)}}\right) & \text{if } d(1-\theta) < \tau p \\ O\left(\varepsilon^{-\frac{p}{(1-\theta)}} \log(\varepsilon^{-1})^{2+\frac{p}{(1-\theta)}}\right) & \text{if } d(1-\theta) = \tau p \\ O\left(\varepsilon^{-d/\tau} \log(\varepsilon^{-1})\right) & \text{if } d(1-\theta) > \tau p. \end{cases}$$

We note that, employing higher order FEM, the condition $d(1-\theta) < \tau p$ becomes less restrictive and that the corresponding work is (asymptotically) lower than in the other two cases.

Chapter 4

Numerical Experiments

4.1 Fast CBC construction for product weights

In this section we focus on some implementation aspects of the CBC algorithm that was presented in Section 2.3. Here we will always assume product weights, that means $\gamma_{\mathfrak{v}} := \prod_{j\in\mathfrak{v}} \gamma_j$ for a positive sequence $(\gamma_j)_{j\geq 1}$. For the first step of the algorithm, we assume that we can check irreducibility of polynomials in $p \in \mathbb{Z}_b[x]$ of arbitrary degree m – while this is an algebraic problem of independent interest, MATLAB provides default choices for p up to $m = 16$ and we will not need higher values in our application.

What requires the most work is then the evaluation arg min in the last step, for all $\tau = 1, \ldots, ds$. Since in each iteration we need to compute $b^m - 1$ values of $B_{\alpha,d,\boldsymbol{\gamma}}$, we prefer to represent this quantity in terms of the digits of the polynomial lattice points, rather than by equation (2.7). We show the following equivalent expression for $B_{\alpha,d,\boldsymbol{\gamma}}$, as proved in [13, Lemma 4] and [5, Lemma 7.4].

Lemma 4.1. *Let $\alpha, d \in \mathbb{N}$ be such that $\bar{\alpha} := 2\min(\alpha, d)$. Let $\phi_{\alpha,d} : [0,1) \to \mathbb{R}$ satisfy that $\phi_{\alpha,d}(0) = \dfrac{(b-1)^2}{b^\alpha(b^{\bar{\alpha}}-1)}$ and*

$$\phi_{\alpha,d}(z) = \frac{(b-1)(b-1-b^{\bar{\alpha}\lfloor \log_b z\rfloor}(b^{\bar{\alpha}+1}-1))}{b^\alpha(b^{\bar{\alpha}}-1)} \qquad \forall z \in (0,1)$$

Given a polynomial lattice $P(\boldsymbol{q}, p) = \{(z_{n,1}, \ldots, z_{n,ds}) \in [0,1)^{ds} : 0 \leq n < b^m\}$, there holds

$$B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p) = \frac{1}{b^m} \sum_{n=0}^{b^m-1} \sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:s\}} \gamma_{\mathfrak{v}} D^{|\mathfrak{v}|} \prod_{j\in\mathfrak{v}} \left[-1 + \prod_{k=1}^{d} (1 + \phi_{\alpha,d}(z_{n,(j-1)d+k}))\right]$$

In particular, if the weights are in product form $\gamma_{\mathfrak{v}} = \prod_{j\in\mathfrak{v}} \gamma_j$, then

$$B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p) = -1 + \frac{1}{b^m} \sum_{n=0}^{b^m-1} \prod_{j=1}^{s} \left[1 - \gamma_j D + \gamma_j D \prod_{k=1}^{d} (1 + \phi_{\alpha,d}(z_{n,(j-1)d+k}))\right]$$

Proof. By *P5* of Proposition 1.10, we obtain that equation (2.7) can be rewritten as

$$\begin{aligned} B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q},p) &= \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \sum_{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u},\mathbf{0}) \frac{1}{b^m} \sum_{n=0}^{b^m-1} \mathrm{wal}_{(\boldsymbol{k}_\mathfrak{u},\mathbf{0})}(\boldsymbol{z}_n) \\ &= \frac{1}{b^m} \sum_{n=0}^{b^m-1} \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \sum_{\boldsymbol{k}_\mathfrak{u}\in\mathbb{N}^{|\mathfrak{u}|}} r_{\alpha,d}(\boldsymbol{k}_\mathfrak{u},\mathbf{0}) \,\mathrm{wal}_{(\boldsymbol{k}_\mathfrak{u},\mathbf{0})}(\boldsymbol{z}_n) \\ &= \frac{1}{b^m} \sum_{n=0}^{b^m-1} \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}} \gamma_{\mathfrak{v}(\mathfrak{u})} D^{|\mathfrak{v}(\mathfrak{u})|} \prod_{j\in\mathfrak{u}} \sum_{k_j\in\mathbb{N}} r_{\alpha,d}(k_j)\,\mathrm{wal}_{k_j}(z_{n,j}). \end{aligned}$$

For $z=0$, there holds

$$\begin{aligned} \sum_{k\in\mathbb{N}} r_{\alpha,d}(k)\,\mathrm{wal}_k(0) &= \frac{b-1}{b^{\alpha-1}} \sum_{l=1}^{\infty} (b-1)\frac{b^{l-1}}{b^{(\bar{\alpha}+1)l}} \\ &= \frac{(b-1)^2}{b^\alpha} \sum_{l=1}^{\infty} \frac{1}{b^{\bar{\alpha}l}} = \phi_{\alpha,d}(0). \end{aligned}$$

On the other hand, for all $z := \sum_{i=1}^{\infty} \zeta_i b^{-i} \in (0,1)$ we get

$$\sum_{k\in\mathbb{N}} r_{\alpha,d}(k)\,\mathrm{wal}_k(z) = \frac{b-1}{b^{\alpha-1}} \sum_{l=1}^{\infty} \frac{1}{b^{(\bar{\alpha}+1)l}} \sum_{k=b^{l-1}}^{b^l-1} \mathrm{wal}_k(z). \tag{4.1}$$

For $l=1$, the inner sum is

$$\sum_{k=1}^{b-1} \mathrm{wal}_k(z) = \sum_{\kappa_0=1}^{b-1} \omega_b^{\kappa_0\zeta_1} = \begin{cases} -1 & \text{if } \zeta_1 \neq 0 \\ b-1 & \text{if } \zeta_1 = 0 \end{cases}.$$

Now fix $l \in \mathbb{N}\backslash\{1\}$ and let $k = \sum_{i=0}^{l-1} \kappa_i b^i$. Observe that $b^{l-1} \le k < b^l$ implies $\kappa_{l-1} \neq 0$, so that

$$\begin{aligned} \sum_{k=b^{l-1}}^{b^l-1} \mathrm{wal}_k(z) &= \sum_{\kappa_{l-1}=1}^{b-1} \omega_b^{\kappa_{l-1}\zeta_l} \cdot \prod_{i=0}^{l-2} \sum_{\kappa_i=0}^{b-1} \omega_b^{\kappa_i\zeta_{i+1}} \\ &= \begin{cases} 0 & \text{if } \zeta_i \neq 0 \text{ for some } i=1,\dots,l-1 \\ b^{l-1} \sum_{\kappa_{l-1}=1}^{b-1} \omega_b^{\kappa_{l-1}\zeta_l} & \text{else} \end{cases} \\ &= \begin{cases} 0 & \text{if } \zeta_i \neq 0, \text{ for some } i=1,\dots,l-1 \\ -b^{l-1} & \text{if } \zeta_l \neq 0, \text{ and } \zeta_i = 0 \ \forall i=1,\dots,l-1 \\ b^{l-1}(b-1) & \text{if } \zeta_i = 0, \ \forall i=1,\dots,l. \end{cases} \end{aligned}$$

We note that $\lfloor \log_b(z) \rfloor = -i_0$, where ζ_{i_0} is the first non-zero digit of z, that is $\zeta_i = 0$ for all $i < i_0$ and $\zeta_{i_0} \neq 0$. In particular, the summands in (4.1) for $l \ge i_0 + 1$ vanish and we have

$$\sum_{k\in\mathbb{N}} r_{\alpha,d}(k)\,\mathrm{wal}_k(z) = \frac{b-1}{b^{\alpha-1}} \left(\frac{-b^{i_0-1}}{b^{(\bar{\alpha}+1)i_0}} + \sum_{l=1}^{i_0-1} \frac{b^{l-1}(b-1)}{b^{(\bar{\alpha}+1)l}} \right),$$

where the sum on the right hand side vanishes by convention in the case $i_0 = 1$. Therefore,

$$\begin{aligned}\sum_{k\in\mathbb{N}} r_{\alpha,d}(k)\,\mathrm{wal}_k(z) &= -\frac{b-1}{b^{\alpha+\bar{\alpha}i_0}} + \frac{(b-1)^2}{b^\alpha}\sum_{l=1}^{i_0-1}\frac{1}{b^{\bar{\alpha}l}}\\ &= -\frac{(b-1)b^{-\bar{\alpha}i_0}}{b^\alpha} + \frac{(b-1)^2(1-b^{-\bar{\alpha}i_0+\bar{\alpha}})}{b^\alpha(b^{\bar{\alpha}}-1)}\\ &= \frac{(b-1)(b-1-b^{-\bar{\alpha}i_0}(b^{\bar{\alpha}+1}-1))}{b^\alpha(b^{\bar{\alpha}}-1)}\\ &= \phi_{\alpha,d}(z).\end{aligned}$$

Using this equality we thus obtain

$$\begin{aligned}B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q},p) &= \frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\emptyset\neq\mathfrak{u}\subseteq\{1:ds\}}\gamma_{\mathfrak{v}(\mathfrak{u})}D^{|\mathfrak{v}(\mathfrak{u})|}\prod_{j\in\mathfrak{u}}\phi_{\alpha,d}(z_{n,j})\\ &= \frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:s\}}\gamma_{\mathfrak{v}}D^{|\mathfrak{v}|}\sum_{\substack{\mathfrak{u}\subseteq\{1:ds\}\\ \mathfrak{v}=\mathfrak{v}(\mathfrak{u})}}\left[\prod_{j\in\mathfrak{u}}\phi_{\alpha,d}(z_{n,j})\right]\\ &= \frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\emptyset\neq\mathfrak{v}\subseteq\{1:s\}}\gamma_{\mathfrak{v}}D^{|\mathfrak{v}|}\sum_{\substack{\mathfrak{u}\subseteq\{1:ds\}\\ \mathfrak{v}=\mathfrak{v}(\mathfrak{u})}}\prod_{j\in\mathfrak{v}(\mathfrak{u})}\prod_{\substack{k=1\\ (j-1)d+k\in\mathfrak{u}}}^{d}\phi_{\alpha,d}(z_{n,(j-1)d+k})\end{aligned}$$

Define $\mathfrak{s}_j := \{(j-1)d+1 : jd\}$ and $\mathfrak{u}_j := \mathfrak{s}_j \cap \mathfrak{u}$ for all $j \in \{1:s\}$. For any fixed $\mathfrak{v}$, there holds $\mathfrak{v} = \mathfrak{v}(\mathfrak{u})$ if and only if $\mathfrak{u}_j \neq \emptyset$ for all $j \in \mathfrak{v}$. Hence,

$$\begin{aligned}B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q},p) &= \frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\substack{\mathfrak{v}\subseteq\{1:s\}\\ \mathfrak{v}\neq\emptyset}}\gamma_{\mathfrak{v}}D^{|\mathfrak{v}|}\prod_{j\in\mathfrak{v}}\left[\sum_{\substack{\mathfrak{u}_j\subseteq\mathfrak{s}_j\\ \mathfrak{u}_j\neq\emptyset}}\prod_{\substack{k=1\\ (j-1)d+k\in\mathfrak{u}_j}}^{d}\phi_{\alpha,d}(z_{n,(j-1)d+k})\right]\\ &= \frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\substack{\mathfrak{v}\subseteq\{1:s\}\\ \mathfrak{v}\neq\emptyset}}\gamma_{\mathfrak{v}}D^{|\mathfrak{v}|}\prod_{j\in\mathfrak{v}}\left[-1+\prod_{k=1}^{d}\left(1+\phi_{\alpha,d}(z_{n,(j-1)d+k})\right)\right]\end{aligned}$$

and the first claim holds. For product weights, we can further simplify the above as follows:

$$\begin{aligned}B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q},p) &= \frac{1}{b^m}\sum_{n=0}^{b^m-1}\sum_{\substack{\mathfrak{v}\subseteq\{1:s\}\\ \mathfrak{v}\neq\emptyset}}\prod_{j\in\mathfrak{v}}\left[-\gamma_j D+\gamma_j D\prod_{k=1}^{d}\left(1+\phi_{\alpha,d}(z_{n,(j-1)d+k})\right)\right]\\ &= -1+\frac{1}{b^m}\sum_{n=0}^{b^m-1}\prod_{j=1}^{s}\left[1-\gamma_j D+\gamma_j D\prod_{k=1}^{d}\left(1+\phi_{\alpha,d}(z_{n,(j-1)d+k})\right)\right].\end{aligned}$$

The proof is now complete. □

We now describe the fast CBC algorithm, as presented in [13]. We can extend the formula shown above to all $\tau = 1, \ldots, ds-1$, where for $1 \le d_0, d_1 \le d$ we

write $\tau = (j_0 - 1)d + d_0$ and $\tau + 1 = (j_1 - 1)d + d_1$. This yields

$$B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}_\tau, p) = -1 + \frac{1}{b^m} \sum_{n=0}^{b^m-1} \left[1 - \gamma_{j_0} D + \gamma_{j_0} D \prod_{k=1}^{d_0} \left(1 + \phi_{\alpha,d}(z_{n,(j_0-1)d+k}) \right) \right] \times \prod_{j=1}^{j_0-1} \left[1 - \gamma_j D + \gamma_j D \prod_{k=1}^{d} \left(1 + \phi_{\alpha,d}(z_{n,(j-1)d+k}) \right) \right].$$

We assume that we have already found the first τ generating polynomials and we are searching the component $q_{\tau+1}(x)$. We define the quantities, only dependent on the first τ components,

$$P_{n,\tau} := \prod_{j=1}^{j_1-1} \left[1 - \gamma_j D + \gamma_j D \prod_{k=1}^{d} \left(1 + \phi_{\alpha,d}(z_{n,(j-1)d+k}) \right) \right]$$
$$Q_{n,\tau} := \prod_{k=1}^{d_1-1} \left(1 + \phi_{\alpha,d}(z_{n,(j_1-1)d+k}) \right)$$

and recall the definition of the points

$$z_{n,(j-1)d+k} = v_m \left(\frac{n(x) q_{(j-1)d+k}(x)}{p(x)} \right).$$

Therefore, using that $(j_1, d_1) = (j_0 + 1, 1)$ if $d_0 = d$ or otherwise $(j_1, d_1) = (j_0, d_0 + 1)$, we can deduce

$$B_{\alpha,d,\boldsymbol{\gamma}}((\boldsymbol{q}_\tau, q), p) = -1 + \frac{1}{b^m} \sum_{n=0}^{b^m-1} P_{n,\tau} \left[1 - \gamma_{j_1} D + \gamma_{j_1} D Q_{n,\tau} \left(1 + \phi_{\alpha,d} \left(v_m \left(\frac{n(x) q(x)}{p(x)} \right) \right) \right) \right].$$

Next, we note that the only term that is influenced by q is

$$\sum_{n=1}^{b^m-1} P_{n,\tau} Q_{n,\tau} \ \phi_{\alpha,d} \left(v_m \left(\frac{n(x) q(x)}{p(x)} \right) \right)$$

and only this needs to be minimised. Since v_m truncates the integer part, we can assume $n(x)q(x)$ as an element in $\mathbb{Z}_b[x]/(p)$; being p irreducible, $\mathbb{Z}_b[x]/(p)$ is a field whose multiplicative group is cyclic, hence there is a $g \in \mathbb{Z}_b[x]/(p)$ such that

$$(\mathbb{Z}_b[x]/(p)) \backslash \{0\} = \{g^0 = g^{b^m-1} = 1, g^1, \ldots, g^{b^m-2}\}.$$

Thus, defining $a_n := P_{n,\tau} Q_{n,\tau}$, we have that there is a $1 \leq z < b^m$ such that, up to a reordering of the sum over n,

$$\sum_{n=1}^{b^m-1} P_{n,\tau} Q_{n,\tau} \ \phi_{\alpha,d} \left(v_m \left(\frac{n(x) q(x)}{p(x)} \right) \right) = \sum_{n=1}^{b^m-1} a_n \phi_{\alpha,d} \left(v_m \left(\frac{g^{z-n}(x)}{p(x)} \right) \right).$$

This is a discrete linear convolution, which result $\boldsymbol{c} = (c_z)_{z=1}^{b^m-1}$ can be computed using FFT by the convolution theorem in $O(mb^m)$ time. Then we select the z_0

which achieves the minimum c_{z_0}. Finally, we update $P_{n,\tau}$ and $Q_{n,\tau}$ according to

$$\begin{cases} P_{n,\tau+1} = P_{n,\tau}\left[1-\gamma_{j_1}D+\gamma_{j_1}DQ_{n,\tau}\left(1+v_m\left(\frac{g^{z_0-n}(x)}{p(x)}\right)\right)\right] & \text{if } d_1 = d \\ P_{n,\tau+1} = P_{n,\tau} & \text{else} \end{cases}$$

and

$$\begin{cases} Q_{n,\tau+1} = 1 & \text{if } d_1 = d \\ Q_{n,\tau+1} = Q_{n,\tau}\left(1+v_m\left(\frac{g^{z_0-n}(x)}{p(x)}\right)\right) & \text{else} \end{cases}$$

These operations can be done in $O(b^m)$ time, thus without affecting the asymptotic complexity. The last step of the CBC algorithm has to be performed ds times, so that the overall computational time is $O(dsmb^m)$. Moreover, we need to store the vectors $P_{n,\tau}$ and $Q_{n,\tau}$ only at the step τ, so that we need $O(b^m)$ memory. In the implementation we consider only the case $b = 2$ so that digit operations are bit operations.

4.2 Implementation of a scrambling algorithm

In order to implement random linear scrambled nets as in Definition 1.5, we repeat the construction in [24, Section 6.2.3] that is valid for general digital nets. As a final step, we operate digit interlacing, thus obtaining an interlaced scrambled polynomial lattice point set. Here, we assume that the CBC algorithm has already been performed so that we have a vector of generating polynomials. Any digital net of b^m points in $[0,1)^{ds}$, including a polynomial lattice, can be defined in terms of generating matrices $C_1, \ldots, C_{ds} \in {\mathbb{Z}_b}^{m\times m}$. This is done as follows: let $\bar{n} \in {\mathbb{Z}_b}^m$ be the vector of digits of some $n \in \{0, \ldots, b^m - 1\}$, then the relation

$$\bar{x}_{n,j} := C_j\bar{n} \qquad \mod b$$

defines the vector $\bar{x}_{n,j}$, which corresponds to the m digits of the j-th component of the n-th point of the digital net. In particular, each point $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,ds})$ is determined by $x_{n,j} := (\bar{x}_{n,j})_1 b^{-1} + \ldots + (\bar{x}_{n,j})_m b^{-m}$.

Algorithm - Interlaced Scrambled Lattice generator. Since we set $b = 2$, we can summarise the algorithm in the following steps, where all arithmetic operations have to be performed in the field $\mathbb{Z}_2$.

1. Build the generating matrices $C_1, \ldots, C_{ds} \in \mathbb{Z}_2^{m\times m}$ of the polynomial lattice rule in ds dimensions;

2. multiply each matrix C_j, $j \in \{1 : ds\}$ by some random lower triangular $R_j \in \mathbb{Z}_2^{m\times m}$, with non-zero diagonal entries;

3. generate random vectors $\bar{g}_j \in \mathbb{Z}_2^m$ for $j \in \{1 : ds\}$;

4. for all $n \in \mathbb{N}$ in the range $0 \le n < b^m$, store the digits of n in a column vector $\bar{n}$ and compute $\bar{r}_{n,j} := R_jC_j\bar{n} + \bar{g}_j \in \mathbb{Z}_2^m$. The result contains the digits of the j-th component of the n-th point of the scrambled lattice;

5. interlacing consists in reshaping the 3-dimensional array $(\bar{r}_{n,j})_{n,j}$.

Finally, we briefly mention how to compute the generating matrices C_j in the first part of the algorithm; for further details we refer to [10, Section 10.1]. We assume that we are given the generating polynomials $\boldsymbol{q} \in (\mathbb{Z}_2[x])^{ds}$ of degree strictly smaller than m and the modulus $p(x) \in \mathbb{Z}_2[x]$ of degree m. Let $q_j(x) = q_1^{(j)} x^{m-1} + \ldots + q_{m-1}^{(j)} x + q_m^{(j)}$ and $p(x) = x^m + p_1 x^{m-1} + \ldots + p_{m-1} x + p_m$ be polynomials. Therefore, observe that

$$\frac{q_j(x)}{p(x)} = \sum_{l=1}^{\infty} u_l^{(j)} x^{-l}$$

where, for $l \leq m$ the coefficients $u_l^{(j)}$ are determined by the triangular linear system

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ p_1 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & & 0 \\ p_{m-1} & p_{m-2} & \cdots & p_1 & 1 \end{pmatrix} \begin{pmatrix} u_1^{(j)} \\ u_2^{(j)} \\ \vdots \\ \vdots \\ u_m^{(j)} \end{pmatrix} = \begin{pmatrix} q_1^{(j)} \\ q_2^{(j)} \\ \vdots \\ \vdots \\ q_m^{(j)} \end{pmatrix}$$

and for $l > m$ one can use the recursion $u_l^{(j)} + u_{l-1}^{(j)} p_1 + u_{l-2}^{(j)} p_2 + \ldots + u_{l-m}^{(j)} p_m = 0$. The generating matrix C_j then is the $m \times m$ Hankel matrix corresponding to the sequence $(u_l^{(j)})_{l \in \mathbb{N}}$ (see [10, Remark 10.2]).

4.3 Numerical results

In the first numerical experiment, we compare the worst case error in variance $WCE^2 = B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p)$ against the number of QMC points, given various choices of weights. We only consider product weights and the basis $b = 2$, that is $N = 2^m$ is the number of QMC points. We give as input $(\alpha, d) \in \{(1,1), (2,2)\}$, so that we expect a decay of approximately $O(N^{-3})$ and $O(N^{-5})$ respectively, provided that the weights have the necessary summability.

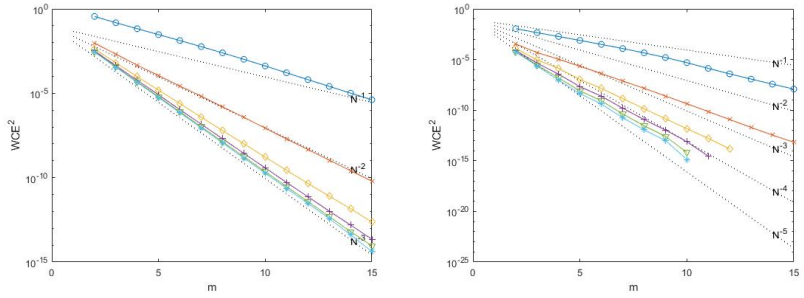


Figure 4.1: Values of $B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p)$ against m, for $s = 1000$, $(\alpha, d) = (1,1)$ (left) and $(2,2)$ (right) and choices of weights $\gamma_j = D^{-1} j^{-n}$ with $n \in \{1,2,3,4,5,6\}$ marked respectively by *circle,cross,diamond,plus,down triangle* and *asterisk*.

In Figure 4.1, it is clear that smaller weights imply faster decay of the worst case error, in both examples. As predicted by Corollary 2.8, the best convergence

rate is attained when $n = 3$ in the left picture. In the right picture, we similarly have that $n = 5$ gives already a good behaviour, but it is difficult to observe the best decay within double precision. On the other hand, extra summability of the weights does not improve the rate beyond $O(N^{-(2\min(\alpha,d)+1)})$. Next, we repeat the same experiment without the scaling of the weights D^{-1}, where we recall that $D = 4^{(\max(d-\alpha,0)}2^{(2d-1)\alpha}$ (see Proposition 2.1).

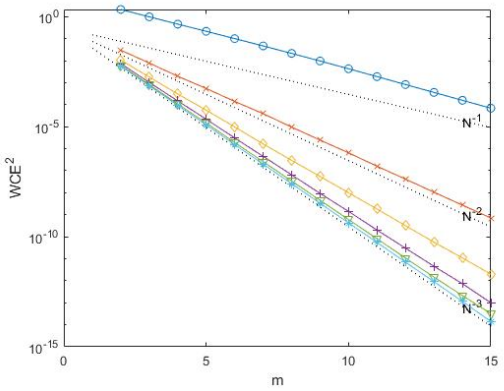


Figure 4.2: Values of $B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p)$ against m, for $s = 1000$, $(\alpha, d) = (1, 1)$ and choices of weights $\gamma_j = j^{-n}$ with $n \in \{1, 2, 3, 4, 5, 6\}$ marked respectively by *circle, cross, diamond, plus, down triangle* and *asterisk*.

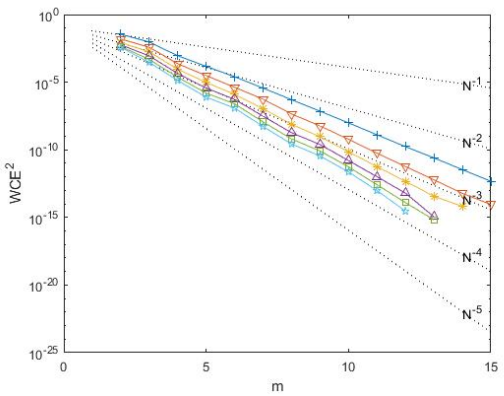


Figure 4.3: Values of $B_{\alpha,d,\boldsymbol{\gamma}}(\boldsymbol{q}, p)$ against m, for $s = 1000$, $(\alpha, d) = (2, 2)$ and choices of weights $\gamma_j = j^{-n}$ with $n \in \{4, 5, 6, 7, 8, 9\}$ marked respectively by *plus, down triangle, asterisk, triangle, square* and *star*.

Unlike with scaling, the worst case error converges slower that what the theory suggests. This is in particular evident in Figure 4.3 as we observe a loss of 2 orders of convergence. Note that, in Figure 4.2, the scaling is compensated by the extra summability of the weights, i.e. when $n \geq 4$. A similar effect can be observed in Figure 4.3 for $n \geq 7$, even if the desired order of convergence is never reached in our tests.

We also test the QMC rule to compute the multivariate integral of $g_\eta(\boldsymbol{y}) := \exp(-\sum_{j=1}^{s} j^{-\eta} y_j) \in \mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ for some product weights $\left(\prod_{j\in\mathfrak{u}} \gamma_j\right)_{\mathfrak{u}\subseteq\{1:s\}}$ to be determined. We use as reference value the exact integral over $[0,1]^s$, that is

$$I_s(g_\eta) = \int_{[0,1]^s} g_\eta = \prod_{j=1}^{s} \int_0^1 \exp(-j^{-\eta} y_j) \mathrm{d}y_j = \prod_{j=1}^{s} j^\eta (1 - \exp(-j^{-\eta})).$$

We have that the parameter $\eta \in \mathbb{N}$ describes the summability of the corresponding weights; in fact, for all $\boldsymbol{\nu} = (\nu_1, \dots, \nu_s) \in \mathbb{N}^s$, we have that $\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} g_\eta(\boldsymbol{y}) = g_\eta(\boldsymbol{y}) \prod_{j \in \mathrm{supp}(\boldsymbol{\nu})} j^{-\eta\nu_j}$. As a consequence,

$$\begin{aligned}\|g_\eta\|_{\mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)} &\leq \sup_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \sum_{\boldsymbol{\nu}\in\{1:\alpha\}^{|\mathfrak{u}|}} \prod_{j\in\mathfrak{u}} j^{-2\eta\nu_j} \sup_{\boldsymbol{y}\in[0,1]^s} g_\eta^2(\boldsymbol{y}) \\ &= \sup_{\mathfrak{u}\subseteq\{1:s\}} \prod_{j\in\mathfrak{u}} \frac{1}{\gamma_j} \sum_{\nu=1}^{\alpha} j^{-2\eta\nu}\end{aligned}$$

and $g_\eta \in \mathcal{W}_{s,\boldsymbol{\gamma},\alpha}([0,1]^s)$ uniformly in $s, \alpha \in \mathbb{N}$, provided that $\gamma_j \geq c \sum_{\nu=1}^{\alpha} j^{-2\eta\nu}$, for some $c > 0$. We can then choose $\gamma_j = D^{-1} j^{-2\eta}$. According to Corollary 2.8, we can obtain L^2 error decay with rate arbitrarily close to $d + \frac{1}{2}$ with the choice of interlacing factor $d = \max\{1, \lfloor \eta - \frac{1}{2} \rfloor\}$.

One can estimate the L^2 error at least in two ways: using the unbiased estimator of the variance or comparing with the exact integral; in [8], the first was used. However, if we scramble a 1-D digital net using Matoušek scrambling and have no interlacing, (see Definition 1.5) we get only a permutation of the QMC points and the unbiased estimator is deterministic. Therefore, we compare the two methods only in higher dimension. For an interlaced scrambled polynomial lattice rule $Q_{s,N}$ of N points in s dimensions with realisations $(q_i)_{i=1}^r$ and empirical mean $\bar{q} = \frac{1}{r}\sum_{i=1}^r q_i$, define the random variables

$$\mathrm{err}_1 := \mathrm{err}_1(r, N) = \sqrt{\frac{1}{r-1}\sum_{i=1}^{r}(\bar{q} - q_i)^2}$$

$$\mathrm{err}_2 := \mathrm{err}_2(r, N) = \sqrt{\frac{1}{r}\sum_{i=1}^{r}\left(I_s(g_\eta) - q_i\right)^2}$$

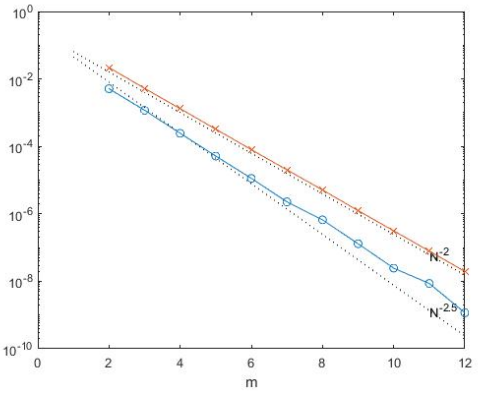


Figure 4.4: displays the values of $\mathrm{err}_1, \mathrm{err}_2$, marked by *circle* and *cross* respectively, against the number of QMC nodes $N = 2^m$. Here, $r = 300$, $(\alpha, d) = (2, 2)$, $s = 5$, $\eta = 4$ and the weights are $\gamma_j = D^{-1} j^{-2\eta}$.

In Figure 4.4, the rate $O(N^{-5/2+\varepsilon})$ expected by the theory is not observed, but we observe instead $O(N^{-2})$, especially when the exact integral is used as comparison. The unbiased estimator of the variance gives slightly better convergence (for sufficiently large N) resembling the results in [8]. Even if the

experiment is repeated, the results are similar and this disparity seems not to be caused by a low number r of independent repetitions.

Before testing the QMC-FEM algorithm, we want to observe the QMC error decay for PDEs. We only consider the one dimensional case and we take, for example, the domain $D = (0, 2)$. We define the hat function $\psi(x) := \max(1 - |x|, 0)$ and the fluctuations by $\psi_j := C_j \psi(\frac{x - x_j}{c_j})$. The values of C_j, c_j and x_j can be set as follows: pick an $\hat{l} \in \mathbb{N}$ and define equispaced nodes

$$x_0 = 0 < x_1 < \ldots < x_{2^{\hat{l}}} = 2 \quad \text{with } x_{j+1} = x_j + h \text{ for fixed } h.$$

For all $j = 1, \ldots, 2^{\hat{l}} - 1$ there exist unique $l, k \in \mathbb{N}$ with $l \leq \hat{l}$ and $k \leq 2^{l-1}$ such that $j = 2^{\hat{l}-l}(2k - 1)$. Then we can scale horizontally the functions according to $c_j := 2^{-l+1}$ to obtain a hierarchical basis functions with $\hat{l}$ levels. Finally, in order to ensure the summability of the fluctuations we can scale their $L^\infty(D)$ norms setting $C_j := a_{\min}/2^{l-1}$, obtaining

$$\psi_j(x) = \frac{a_{\min}}{2^{l-1}} \psi\left(2^{l-1}x - 2k + 1\right). \tag{4.2}$$

Since these functions do not overlap if they belong to the same level, the sparsity condition is also satisfied with the choice of $b_j := 2^{-(l-1)} l^{-\frac{1}{p-\varepsilon}}$ for some $\varepsilon > 0$. This is restrictive on the weights but it can be relaxed using a different basis for the fluctuations, as in [12]. In all the following experiments, the number of fluctuations – and therefore the dimension of the parameter domain – is fixed at $s = 2^{\hat{l}} - 1$ and no dimension truncation is considered.

For the experiment in Figure 4.5, we chose the input $\bar{a}(x) \equiv 1$, $f(x) \equiv 1$ on the domain $D = (0, 2)$, with mixed Dirichlet/Neumann boundary conditions $u(0) = 0$, $u'(2) = 0$. Then, with the fluctuations defined in (4.2), we have the diffusion coefficient

$$a(x, \boldsymbol{y}) = 1 + \sum_{j=1}^{2^{\hat{l}}-1} y_j \psi_j(x).$$

As an example, we check the quantity of interest $G(u) = u(2, \boldsymbol{y})$, which integral can be evaluated with the formula

$$\begin{aligned} \int_{[-\frac{1}{2}, \frac{1}{2}]^s} G(u(x, \boldsymbol{y}))\mathrm{d}\boldsymbol{y} &= \mathbb{E}_{\boldsymbol{y}}\left[\int_0^2 \frac{1}{a(x, \boldsymbol{y})} \int_x^2 f(t)\mathrm{d}t\mathrm{d}x\right] \\ &\approx \frac{1}{\hat{N}} \sum_{i=0}^{\hat{N}-1} \int_0^2 \frac{2 - x}{a(x, \boldsymbol{y}_i)}\mathrm{d}x =: I_{\hat{N}}(G(u)). \end{aligned}$$

The reference value $I_{\hat{N}}(G(u))$ is approximated with standard trapezoidal integration with mesh size δ on $(0, 2)$; this is repeated $\hat{N}$ times with QMC points $\boldsymbol{y}_i$, and we denote the result by $I_{\delta,\hat{N}}(G(u)))$. As QMC quadrature rule we use $Q_{s,N}(F) = I_s(F; P^{IS})$ where N is the cardinality of the interlaced scrambled polynomial lattice P^{IS} and define the random variable

$$\mathrm{err} := \mathrm{err}(h, N) = |Q_{s,N}(u_h(2)) - I_{\delta,\hat{N}}(G(u))|.$$

Since we fixed M_h, in Figure 4.5 we observe the second order decay only for small values of N: after some threshold, the overall error is essentially the Galerkin error and remains constant.

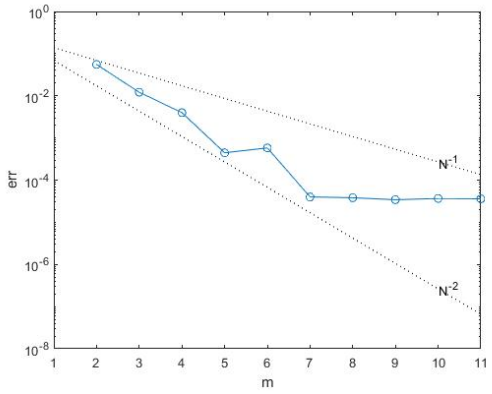


Figure 4.5: displays the values of err against the number of QMC nodes $N = 2^m$. Here, $(\alpha, d) = (2, 2)$, $\hat{l} = 2$, that is $s = 3$, $M_h = 65$ and the weights are $\gamma_j = D^{-1} j^{-4}$. For the reference value, $\hat{N} = 2^{13}$ and $\delta = 2^{-10}$.

In the following experiment (Figures 4.6 and 4.7) we analyse the convergence of the single level QMC-FEM as in Section 3.5. We consider the nominal operator $\bar{a}(x) = 1.5 + \cos(\pi x)$, $f \equiv 1$ and mixed Dirichlet/Neumann boundary conditions $u(0) = 0$, $u'(2) = 0$. We use again the same quantity of interest $G(u) = u(2, \boldsymbol{y})$ and we compute similarly the reference value. Moreover, since we use a piecewise linear Finite Elements solver,so that $\tau = 1$ in Section 3.5: to obtain first order convergence we therefore need to set $M_h \sim N$. For second order we need analogously that $M_h \sim N^2$ provided that $p \leq 1/2$ and so on for higher order.

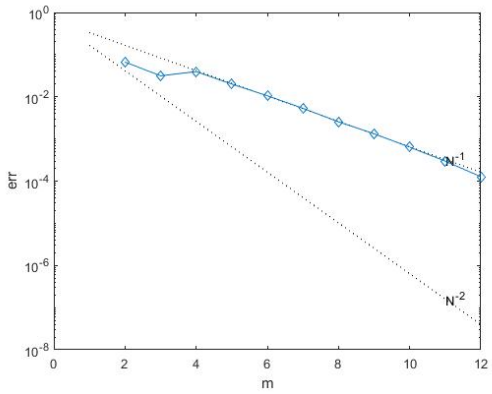


Figure 4.6: displays the values of err against the number of QMC nodes $N = 2^m$. Here, $(\alpha, d) = (1, 1)$, $\hat{l} = 4$, i.e $s = 15$ $M_h = N$ and the weights are uniform: $\gamma_j = D^{-1}$. For the reference value, $\hat{N} = 2^{14}$ and $\delta = 2^{-10}$.

In Figure 4.6, the convergence rate of $O(N^{-1})$ is clear, despite the uniform weights. This suggests that the decay condition on the weights imposed by the theory (cp. equation (3.20) and the definition b_j above) is not sharp, as full rate can already be observed in this case. Since we only had $s = 15$, in order to make this observation more explicit, we repeat the experiment with decaying weights and larger dimension.

In Figure 4.7, we have $s = 63$, so that the decaying weights are considerably smaller than the uniform ones. However, we observe that there is no difference in the decay rate of the errors.

Finally, we conclude with a second order example, with two different choices of QMC-FEM couplings. The order of err $= O(N^{-2})$ can be observed in Figure

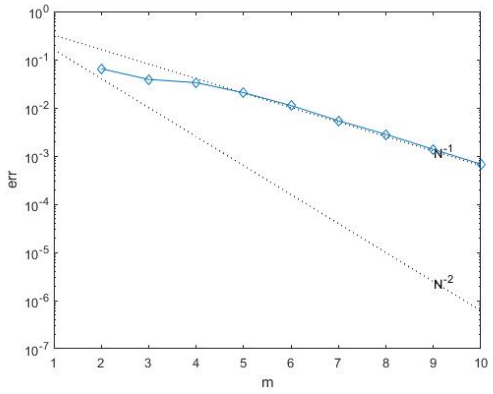


Figure 4.7: values of err against the number of QMC nodes $N = 2^m$. Here, $(\alpha, d) = (1, 1)$, $\hat{l} = 6$, i.e. $s = 63$ $M_h = N$ and the weights are $\gamma_j = D^{-1} j^{-4}$. For the reference value, $\hat{N} = 2^{14}$ and $\delta = 2^{-10}$.

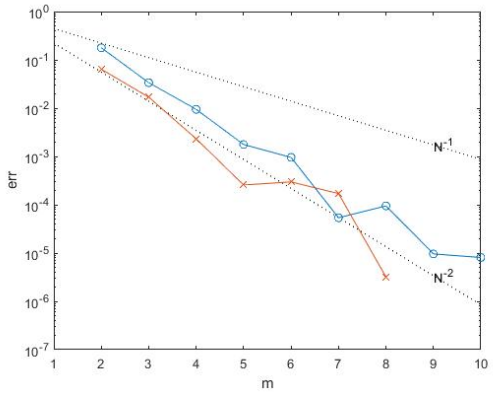


Figure 4.8: values of err against the number of QMC nodes $N = 2^m$. Here, $(\alpha, d) = (2, 2)$, $\hat{l} = 4$, i.e. $s = 15$ and the weights are $\gamma_j = D^{-1}$. For the reference value, $\hat{N} = 2^{14}$ and $\delta = 2^{-10}$. $M_h = N, N^2$ are marked by *circle* and *cross* respectively.

4.8. Note that when N increases, the choice $M_h = N$ is less stable and the convergence deteriorates. On the other hand, the computational cost increases too quickly if $M_h = N^2$, so that the constraint on the first order FEM plays a crucial role.

Appendix A

MATLAB Codes

```
-------------------------------------------------------------------
%   Finds first m coefficients of q(x)/p(x) in Z_2
%   (see Dick, Pillichshammer - Digital nets and sequences
%       definition 10.1, remark 10.2 and proposition 10.4)
%
%   assumes p(x) irreducible and monic of degree m, q(x) of
%   degree < m
%   INPUT:  p       coefficients of p(x) with increasing
%                   power up to x^{m-1}, vector of size m.
%           q       coefficients of q(x), matrix of size [m x s_max]
%   OUTPUT: u       Laurent coefficients, matrix of size [m x s_max]

function u = vmcoeff(p,q)

m = size(q,1);
s_max = size(q,2);

reshape(p, 1, length(p));
u = zeros(m,s_max);
q=double(q);
p=double(p);
for i= 1:m
    u(i,:) = mod(q(m-i+1,:) - p(m-i+2:m)* u(1:i-1,:),2);
end
end


-------------------------------------------------------------------

% Interlaced scrambled polynomial lattice rule over Z_2/f with
% deg(f) = m.
% Performs the CBC construction of the generating polynomials
% and computes the worst case error squared.
% See Goda, Dick - Construction of interlaced scrambled
% polynomial lattice rules of arbitrary high order p.1272-1273
%
%   INPUT:  m       degree of the irreducible polynomial,scalar
%           d       interlacing factor, scalar
%           alpha   smoothness of function, scalar
%           s_max   number of dimensions, scalar
%           gamma   parameters for weighting the dimensions,
%                   vector [s_max x 1]
%           prune   use pruning to avoid repetition of components,
%                   logical
```

```
%
%   OUTPUT: C       Generating matrices of the polynomial lattice,
%                   C(:,:,i) are matrices of size [m x m]
%                   for i=1,...,s_max*d
%           WCE2    worst case error, quality criterion for the
%                   variance of the estimator


% (C) 2007, <dirk.nuyens@cs.kuleuven.ac.be>
% File modified by:
% Takashi Goda <goda@iba.t.u-tokyo.ac.jp>. Feb. 2013
% Marcello Longo. Mar. 2019

function [C,WCE2] = polyLatticeCBC(m, d, alpha, s_max, gamma,prune)

C = zeros(m,m,d*s_max);

z = zeros(s_max*d, 1);

N = pow2(m);
a2 = 2*min(alpha, d);
phi = @(x) (1-pow2(a2*floor(log2(x)))*(pow2(a2+1)-1))...
        /pow2(alpha)/(pow2(a2)-1);

g = gf(2, m); % generator g(x) = 2 = (10)_2 = x
perm = gf(zeros(N-1, 1), m);
perm(1) = 1;
for j=2:N-1
    % perm(j) is g^(j-1) for all j=1,...,N-1
    perm(j) = perm(j-1)*g;
end
q = perm.x;
q = (de2bi(q,m))';  %extract binary from decimal
p = de2bi(primpoly(m));
p = p(1:end-1);         % p is a monic polynomial

psi = phi(pow2(-(1:m))*vmcoeff(p,q)); %psi = \phi(v_m(q/p))
psi = psi';

fft_psi = fft(psi);

gamma = gamma*power(4,max(d-alpha,0))*pow2((2*d-1)*alpha);

P = ones(N-1, 1);
Q = ones(N-1, 1);

for s = 1:s_max
    for k = 1:d
        j = (s - 1) * d + k;
        a = P.*Q;
        E2 = real(ifft(fft_psi .* fft(a))); % convolution theorem

        if prune
            notvalid = [1;z(1:j-1)]; % previous components

            % if exist available generating vectors
            if ~isempty(setdiff(1:N-1,notvalid))

                E2(notvalid) = NaN;  % min among valid components

            end
            %else no pruning
```

```
        end
        [min_E2, z(j)] = min(E2);    % CBC argmin
        if j == 1
            z(j) = 1;
        end
        for n=1:N-1      %Update Q
            genPow = mod(z(j)-n,N-1);
            Q(n) = (1 + psi(genPow+1)) * Q(n);
        end

        fprintf('j=%4d, z=%6d\n', j, z(j));
        C(:,:,j) = genMatrix(p,q(:,z(j)));
    end
    P = (1 - gamma(s) + gamma(s) * Q) .* P; % Update P
    Q = ones(N-1,1);
end
P0_ds = prod(1-gamma + gamma * power(1+phi(0),d)); %error for n==0

WCE2 = -1 + mean([P; P0_ds ]);
end
```

```
%--------------------------------------------------------------------
%   Computes one generating matrix of a polynomial lattice
%   point set P(q,p) over Z_2.
%   (see Dick, Pillichshammer - Digital nets and sequences
%       definition 10.1, remark 10.2 and proposition 10.4)
%
%   Assumes p(x) irreducible and monic of degree m, q(x) of
%   degree < m
%   INPUT : p       coefficients of p(x) with increasing power
%                   up to x^{m-1}, column vector of size m.
%           q       coefficients of q(x), column vector of size m

%   OUTPUT: C       Generating matrix of the polynomial lattice,
%                   matrix of size [m x m]

%   To generate the i-th component of the n-th point,
%   do { C(:,:,i)*x } where x contains the m digits of n<2^m
%   in increasing order

function C = genMatrix(p,q)

m = size(q,1);
u = zeros(2*m-1,1);
C = zeros(m,m);

u(1:m) = vmcoeff(p,q);

for j = 1:(m-1)
    u(j+m,:) = mod(- p*u(j:(j+m-1),:),2);
end
C = hankel(u(1:m),u(m:end));

end
```

```
%--------------------------------------------------------------------
%   Implements the affine Matousek scrambling of digital nets
%   over the field Z_2, see
%   Matousek - On L2 discrepancy for anchored boxes, pag. 540
%
%   INPUT:  C       Generating matrices of the polynomial lattice,
```

```
%                   C(:,:,i) are matrices of size [m x m]
%                   for i=1,...,s_max
%   OUTPUT: G       Generating matrices of the scrambled points,
%                   same class and dimension as C
%           B       bias (affine scrambling), B(:,i) are
%                   vectors of size m for i=1,...,s_max
%
%   To generate the i-th component of the n-th point,
%   do { G(:,:,i)*x XOR B(:,i) } where x contains the m digits of
%   n<2^m in increasing order

function [G,B] = affineNestedScrambling(C)
m=size(C,1);
s_max=size(C,3);
G=zeros(m,m,s_max);

rng('shuffle');

for i=1:s_max
    %random, nonsingular LT matrix in Z_2
    R = eye(m,m) + tril(randi([0 1],m,m),-1);

    G(:,:,i) = mod(R * C(:,:,i),2);
end

B = randi([0 1],m,s_max);     %random vectors in Z_2

end


-------------------------------------------------------------------
% The function interlace2 computes the digit interlacing function
% in [0,1)^{d*s}
% Works for any base b
% INPUTS :  x       matrix of size [m x d*s_max], with entries in
%                                       Z_b: each column represents
%                   the digits of the i-th component of x
%           d       interlacing factor,scalar
%
% OUTPUT :  y       one interlaced point, [d*m x s_max] matrix
%                   with entries in Z_b: each column represents
%                   the digits of the i-th component of y

function y = interlace2(x,d)
% if x has wrong size this discards last components
s_max = floor(size(x,2)/d);

m = size(x,1);
y=zeros(d*m,s_max);

for s= 1:s_max
    digitMatrix = (x(:,(s-1)*d + 1 : s*d))';
    y(:,s) = reshape(digitMatrix,d*m,1);  %rearrange digits
end
end


-------------------------------------------------------------------
%   interlScrambLatticeGenerator generates an interlaced scrambled
%   polynomial lattice point set over Z_2 for a function with
%   bounded generalized weighted Hardy and Krause variation
%   INPUT:  m       2^m = number of points, scalar, must be <= 16
%           d       interlacing factor, scalar
%           alpha   smoothness factor, scalar
```

```
%            s_max    number of variables, scalar
%            gamma    weights, column vector of size s_max
%            prune    use pruning to avoid repeated components,
%                     logical
%
%   OUTPUT: y         QMC nodes, matrix of size [2^m x s_max]
%                     (each row is a point)
%

function y = interlScrambLatticeGenerator(m, d, alpha, s_max, gamma,prune)

N = pow2(m);
y = zeros(N, s_max);
x = zeros(m, d*s_max);
C = polyLatticeCBC(m, d, alpha, s_max, gamma,prune);

[G,B] = affineNestedScrambling(C);
disp("Scrambled generating matrices ready");

for n = 1: N
    bits = (de2bi(n-1,m))'; % array of digits of n
    for s = 1:d*s_max

        % Matousek scrambling
        x(:,s)=mod(G(:,:,s)*bits+B(:,s),2);
    end
    %interlacing
    y(n,:) = pow2(-(1:d*m))*interlace2(x,d);
end
end

-------------------------------------------------------------------
% Generates the hierarchical basis (without pyramid scheme)
% for the affine parametric diffusion coefficient of
% the PDE -(a *u')'=f on the interval D
% and Dirichlet BC.
%
%   INPUT:  intL        left extremum of the interval D, scalar
%           intR        right extremum of the interval D, scalar
%           ell_max     maximum level of hierarchical basis,
%                       scalar
%           a_bar       nominal operator, function handle
%
%   OUTPUT: psi         function handle of the basis ordered by
%                       level, output is column of size 2^(ell_max)-1

function psi = hierarchical_assemble(intL,intR, a_bar,ell_max)

x = linspace(intL,intR,pow2(ell_max)+1);     % equispaced nodes
x([1, end]) = [];                            % with no extrema

a_min = min(a_bar(x));
assert(a_min > 0, "Nominal ellipticity failed");

hat_handle = @(x) max(1-abs(x),0);  %mother of hat basis

s=0;

psi = "@(x) [";
for ell=1:ell_max
    % scale basis to ensure summability of fluctuations
    powell = pow2(ell-1);
    for k=1:powell
```

```
        % x(j) is the peak of current hierarchical basis function
        j = pow2(ell_max-ell)*(2*k-1);

        % concatenate string of new fluctuation
        psi = psi + "a_min/"+ num2str(powell) +"*hat_handle((x-"...
            + num2str(x(j))+")*"+ num2str(pow2(ell)) + ...
            "/(intR-intL)),";
        s=s+1;

    end
end
psi = psi + "]";
psi=eval(psi);  % evaluates string to produce a function handle

% plot (linspace(intL,intR),psi(linspace(intL,intR)'));
end
```

The following code assumes the availability of a FEM solver for second order elliptic PDEs: the functions `load_vec` and `stiff` compute respectively the load vector and the stiffness matrix starting from anonymous functions. Note that, in principle, we could implement a more efficient pyramidal scheme to evaluate all the fluctuations. However, this is only possible if we use a FEM solver that requires point values of the diffusion coefficient as input instead of one function handle.

```
% -------------------------------------------------------------------
%  Computes FEM solutions of affine parametric PDE for each
%  istance of QMC points
%   INPUT:  intL        left extremum of the interval D, scalar
%           intR        right extremum of the interval D, scalar
%           ell_max     maximum level of hierarchical basis, scalar
%           a_bar       nominal operator, function handle
%           f_handle    RHS of the PDE, function handle
%           y           QMC points, matrix of size [2^m x s_max]
%                       (each row is a point in [0,1]^s_max )
%           M           number of FE intervals + 1, scalar
%
%   OUTPUT: u           values of functions for all QMC points,
%                       matrix of size [M x 2^m] (each column is a
%                       FEM solution)

function u = QMCFEM(intL,intR,a_bar,f_handle,ell_max,y,M)

psi = hierarchical_assemble(intL,intR,a_bar,ell_max);

x = linspace(intL,intR,M)';    %FEM mesh

N = size(y,1);
u = zeros(M,N);

F = load_vec(x,f_handle);

%set for homogeneous Dirichlet BC
%dof = 2:M-1;

%set for Mixed BC : Dirichlet at intL and Neumann at intR
dof = 2:M;

y=y-0.5;     %shift the QMC points to [-1/2,1/2]
```

```
for i=1:N
    %diffusion coefficient
    a_handle = @(x) a_bar(x) + psi(x)*(y(i,:)');

    %stiffness matrix
    A = stiff(x,a_handle,@(x) 0,@(x) 0);

    u(dof,i) = (A(dof,dof) \ F(dof));
end
end
```

Bibliography

[1] I. Babuška and B. Q. Guo. Regularity of the solution of elliptic problems with piecewise analytic data. I. Boundary value problems for linear elliptic equation of second order. *SIAM J. Math. Anal.*, 19(1):172–203, 1988.

[2] I. Babuška, R. B. Kellogg, and J. Pitkäranta. Direct and inverse error estimates for finite elements with mesh refinements. *Numer. Math.*, 33(4):447–471, 1979.

[3] Markus Bachmayr, Albert Cohen, Dinh Dũng, and Christoph Schwab. Fully discrete approximation of parametric and stochastic elliptic PDEs. *SIAM J. Numer. Anal.*, 55(5):2151–2186, 2017.

[4] Constantin Bacuta, Victor Nistor, and Ludmil T. Zikatanov. Improving the rate of convergence of high-order finite elements on polyhedra. II. Mesh refinements and interpolation. *Numer. Funct. Anal. Optim.*, 28(7-8):775–824, 2007.

[5] Jan Frederik Baldeaux. *Higher order nets and sequences*. PhD thesis, The University of New South Wales, 2010.

[6] Constantin Băcuţă, Victor Nistor, and Ludmil T. Zikatanov. Improving the rate of convergence of 'high order finite elements' on polygons and domains with cusps. *Numer. Math.*, 100(2):165–184, 2005.

[7] Kai Lai Chung. *A course in probability theory*. Academic Press, Inc., San Diego, CA, third edition, 2001.

[8] Josef Dick. Higher order scrambled digital nets achieve the optimal rate of the root mean square error for smooth integrands. *Ann. Statist.*, 39(3):1372–1398, 2011.

[9] Josef Dick, Frances Y. Kuo, Quoc T. Le Gia, Dirk Nuyens, and Christoph Schwab. Higher order QMC Petrov-Galerkin discretization for affine parametric operator equations with random field inputs. *SIAM J. Numer. Anal.*, 52(6):2676–2702, 2014.

[10] Josef Dick and Friedrich Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi–Monte Carlo Integration*. Cambridge University Press, 2010.

[11] Robert N. Gantner, Lukas Herrmann, and Christoph Schwab. Multilevel QMC with product weights for affine-parametric, elliptic PDEs. In *Contemporary computational mathematics—a celebration of the 80th birthday of Ian Sloan. Vol. 1, 2*, pages 373–405. Springer, Cham, 2018.

[12] Robert N. Gantner, Lukas Herrmann, and Christoph Schwab. Quasi-Monte Carlo integration for affine-parametric, elliptic PDEs: local supports and product weights. *SIAM J. Numer. Anal.*, 56(1):111–135, 2018.

[13] Takashi Goda and Josef Dick. Construction of interlaced scrambled polynomial lattice rules of arbitrary high order. *Found. Comput. Math.*, 15(5):1245–1278, 2015.

[14] Michael Hardy. Combinatorics of partial derivatives. *Electron. J. Combin.*, 13(1):Research Paper 1, 13, 2006.

[15] L. Herrmann and Ch. Schwab. Multilevel qmc uncertainty quantification for advection-reaction-diffusion. Technical Report 2019-06, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2019.

[16] Lukas Herrmann and Christoph Schwab. Multilevel quasi-monte carlo integration with product weights for elliptic pdes with lognormal coefficients. Technical Report 2017-19 (revised), Seminar for Applied Mathematics, ETH Zürich, 2017.

[17] Lukas Herrmann and Christoph Schwab. Qmc integration for lognormal-parametric, elliptic pdes: local supports and product weights. *Numer. Math.*, 141(1):63–102, 2019.

[18] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948.

[19] V. A. Kondratiev. Boundary value problems for elliptic equations in domains with conical or angular points. *Trudy Moskov. Mat. Obšč.*, 16:209–292, 1967.

[20] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Math. Comp.*, 79(270):953–966, 2010.

[21] Frances Y. Kuo, Christoph Schwab, and Ian H. Sloan. Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(6):3351–3374, 2012.

[22] Frances Y. Kuo, Christoph Schwab, and Ian H. Sloan. Multi-level quasi-monte carlo finite element methods for a class of elliptic pdes with random coefficients. *Found. Comput. Math.*, 15(2):411–449, April 2015.

[23] Gerhard Larcher and Claudia Traunfellner. On the numerical integration of Walsh series by number-theoretic methods. *Math. Comp.*, 63(207):277–291, 1994.

[24] Christiane Lemieux. *Monte Carlo and quasi-Monte Carlo sampling.* Springer Series in Statistics. Springer, New York, 2009.

[25] Jiří Matoušek. On the L_2-discrepancy for anchored boxes. *J. Complexity*, 14(4):527–556, 1998.

[26] Harald Niederreiter. Low-discrepancy point sets obtained by digital constructions over finite fields. *Czechoslovak Math. J.*, 42(117)(1):143–166, 1992.

[27] Erich Novak. *Deterministic and stochastic error bounds in numerical analysis*, volume 1349 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988.

[28] Art B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In *Monte Carlo and quasi-Monte Carlo methods in scientific computing (Las Vegas, NV, 1994)*, volume 106 of *Lect. Notes Stat.*, pages 299–317. Springer, New York, 1995.

[29] Shu Tezuka and Henri Faure. I-binomial scrambling of digital nets and sequences. *J. Complexity*, 19(6):744–757, 2003. Information-Based Complexity Workshop (Minneapolis, MN, 2002).

[30] J. L. Walsh. A Closed Set of Normal Orthogonal Functions. *Amer. J. Math.*, 45(1):5–24, 1923.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

HIGHER ORDER QMC INTEGRATION WITH SCRAMBLING FOR ELLIPTIC PDES WITH RANDOM COEFFICIENTS

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
LONGO	MARCELLO

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 23/04/2019	Marcello Longo

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.