NUMERICS OF HIGH-DIMENSIONAL PROBLEMS

MICHAEL FEISCHL

Contents

Date: May 24, 2024.

1. High dimensional quadrature

This section is largely based on [3]. Integration in high-dimensions is very important for applications in finance (computing expectations), physics (Schrödinger equation), and other fields. We consider integration on the unit cube $[0,1]^s$, for some $s \in \mathbb{N}$. Hence, we would like to approximate

$$I_s(f) := \int_{[0,1]^s} f(\boldsymbol{x})\, d\boldsymbol{x} := \int_0^1 \cdots \int_0^1 f(x_1, \ldots, x_s)\, dx_1 \cdots dx_s$$

in the case where $s \in \mathbb{N}$ is large, possibly in the hundreds or thousands.

1.1. **The problem with tensor-quadrature.** In classical numerical analysis, one would approximate $\int_0^1 f(x)\, dx \approx \sum_{i=1}^n w_i f(z_i)$ for some (Gauss)-quadrature points $z_1, \ldots, z_n$. To obtain a quadrature rule for the multi-dimensional integral, the standard way forward is to use tensor quadrature, i.e.,

$$I_s(f) \approx T_n(f) := \sum_{i_1=1}^n \cdots \sum_{i_s=1}^n w_{i_1} w_{i_2} \cdots w_{i_s} f(z_{i_1}, \ldots, z_{i_s}).$$

Under some assumptions on the regularity of f, we can prove that the error satisfies

$$|I_s(f) - T_n(f)| \lesssim n^{-r} \|f\|_{C^r([0,1]^s)}.$$

However, in terms of number of evaluations N of f, the rate of convergence looks less impressive, i.e.,

$$|I_s(f) - T_n(f)| \lesssim N^{-r/s} \|f\|_{C^r([0,1]^s)},$$

since we have $N = n^s$ tensor points $\boldsymbol{z_i} := (z_{i_1}, \ldots, z_{i_s})$ for all $\boldsymbol{i} \in \{1, \ldots, n\}^s$. The convergence rate depends on the dimension of the problem.

1.2. **Monte Carlo quadrature.** The classical Monte Carlo (MC) method is an equal-weight quadrature rule of the form

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

where $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}$ are i.i.d. uniform random samples from $[0,1]^s$.

Theorem 1. *For all $f \in L^2([0,1]^s)$, the root-mean squared error satisfies*

$$\sqrt{\mathbb{E}(|I_s(f) - Q_n(f)|^2)} = \sqrt{\mathrm{Var}(f)}\, n^{-1/2},$$

where $\mathrm{Var}(f) = I_s(f^2) - I_s(f)^2$. Here, the expectation is taken with respect to the uniform random samples $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}$.

Proof. We have

$$\mathbb{E}(|I_s(f) - Q_n(f)|^2) = \mathbb{E}(I_s(f)^2) - 2\mathbb{E}(Q_n(f) I_s(f)) + \mathbb{E}(Q_n(f)^2).$$

Obviously, we have $\mathbb{E}(I_s(f)^2) = I_s(f)^2$ and $\mathbb{E}(Q_n(f) I_s(f)) = \mathbb{E}(Q_n(f)) I_s(f)$. Moreover, there holds

$$\mathbb{E}(Q_n(f)) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}(f(\boldsymbol{x}_i)) = \frac{1}{n} \sum_{i=0}^{n-1} \int_{[0,1]^s} f(\boldsymbol{x})\, d\boldsymbol{x} = I_s(f)$$

as well as

$$\begin{aligned}\mathbb{E}(Q_n(f)^2) &= \sum_{i,j=0}^{n-1} \mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)) \\ &= \frac{1}{n^2}\Big(\sum_{i=0}^{n-1} \mathbb{E}(f(\boldsymbol{x}_i)^2) + \sum_{i\neq j} \mathbb{E}(f(\boldsymbol{x}_i))\mathbb{E}(f(\boldsymbol{x}_j))\Big) \\ &= \frac{1}{n^2}\Big(\sum_{i=0}^{n-1} I_s(f^2) + \sum_{i\neq j} I_s(f)^2\Big) = \frac{1}{n} I_s(f^2) + \frac{n-1}{n} I_s(f)^2.\end{aligned}$$

Altogether, this shows

$$\mathbb{E}(|I_s(f) - Q_n(f)|^2) = \frac{I_s(f^2) - I_s(f)^2}{n}$$

and hence we conclude the proof. □

We note that the convergence rate of the error is independent of the dimension $s \in \mathbb{N}$ and (almost) independent of the smoothness of the integrand f. For $f \in C^2([0,1]^s)$, the tenor trapezoidal rule achieves a convergence rate of $N^{-2/s}$. This means that for $s > 4$, the MC method is actually faster than the tensor quadrature.

1.3. **Quasi-Monte Carlo quadrature.** The MC method is the best method for low regularity integrands. In fact, Bakhvalov (1959) proved that the rate $\mathcal{O}(n^{-1/2})$ can not be improved for $f \in L^2([0,1]^s)$. On the other hand, the MC method will not benefit of more regular integrands. Hence, we want to discuss quasi-Monte Carlo methods (QMC).

QMC methods are equal weight quadrature rules of the form

$$Q_n(f) := \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{z}_i)$$

for some (deterministically chosen) points $\boldsymbol{z}_0, \dots, \boldsymbol{z}_{n-1} \in [0,1]^s$. In the following, we will develop an error analysis for those QMC rules. The main difference to the classical error analysis of quadrature problems is, that we do not bootstrap the interpolation problem, i.e., the fact that quadrature is not harder than interpolation is used to develop Gaussian quadrature rules for example. However, it turns out that in higher dimensions, quadrature is much easier than interpolation.

Definition 2 (reproducing kernel Hilbert space)**.** *Let $\mathcal{X}$ be a Hilbert space of functions $f\colon O \to \mathbb{R}$ for an arbitrary set O equipped with the vector space structure of pointwise addition and scalar multiplication. If the functionals $L_{\boldsymbol{x}}\colon \mathcal{X} \to \mathbb{R}$, $L_{\boldsymbol{x}}(f) := f(\boldsymbol{x})$ are bounded for all $\boldsymbol{x} \in O$, then $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS).*

Lemma 3. *If $\mathcal{X}$ is a RKHS, then there exists a unique function $K\colon O \times O \to \mathbb{R}$ (the kernel) such that*

(i) *$K(\cdot, \boldsymbol{x}) \in \mathcal{X}$ for all $\boldsymbol{x} \in O$,*
(ii) *$f(\boldsymbol{x}) = \langle f\,,\, K(\cdot, \boldsymbol{x})\rangle_{\mathcal{X}}$ for all $f \in \mathcal{X}$ and all $\boldsymbol{x} \in O$ (the reproducing property).*
(iii) *$K(\cdot,\cdot)$ is symmetric, i.e., $K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in O$.*

(iv) $K(\cdot,\cdot)$ *is positive semidefinite, i.e., for all* $n \in \mathbb{N}$, $\mu_1, \dots, \mu_n \in \mathbb{R}$, *and* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in O$, *there holds*

$$\sum_{i,j=1}^{n} \mu_i \mu_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0.$$

(v) *The set* $\operatorname{span}\{K(\cdot, \boldsymbol{x}) \colon \boldsymbol{x} \in O\}$ *is dense in* $\mathcal{X}$.

Moreover, if a function $K \colon O \times O \to \mathbb{R}$ *satisfies (iii)–(iv), there exists a unique RKHS* $\mathcal{X}$ *with kernel* K.

Proof. Since L_x is bounded by definition, we find a Riesz representer $K_{\boldsymbol{x}} \in \mathcal{X}$ such that $L_{\boldsymbol{x}}(\cdot) = \langle K_{\boldsymbol{x}}\,,\,\cdot\rangle_{\mathcal{X}}$. We may define $K(\boldsymbol{x},\boldsymbol{y}) := \langle K_{\boldsymbol{x}}\,,\,K_{\boldsymbol{y}}\rangle_{\mathcal{X}}$ and see (iii) immediately. Moreover, we have $K(\cdot,\boldsymbol{y}) = (\boldsymbol{x} \mapsto K(\boldsymbol{x},\boldsymbol{y})) = (\boldsymbol{x} \mapsto \langle K_{\boldsymbol{x}}\,,\,K_{\boldsymbol{y}}\rangle_{\mathcal{X}}) = (\boldsymbol{x} \mapsto K_{\boldsymbol{y}}(\boldsymbol{x}))$ and hence $K(\cdot,\boldsymbol{y}) \in \mathcal{X}$, i.e. (i). The same argument also shows $\langle f\,,\,K(\cdot,\boldsymbol{x})\rangle_{\mathcal{X}} = \langle f\,,\,K_{\boldsymbol{x}}\rangle_{\mathcal{X}} = f(\boldsymbol{x})$, i.e. (ii). Positive semi-definiteness (iv) follows from

$$\sum_{i,j=1}^{n} \mu_i \mu_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \sum_{i,j=1}^{n} \mu_i \mu_j \langle K_{\boldsymbol{x}_i}\,,\,K_{\boldsymbol{x}_j}\rangle_{\mathcal{X}} = \langle \sum_{i=1}^{n} \mu_i K_{\boldsymbol{x}_i}\,,\,\sum_{i=1}^{n} \mu_i K_{\boldsymbol{x}_i}\rangle_{\mathcal{X}} \geq 0,$$

since $\langle\cdot\,,\,\cdot\rangle_{\mathcal{X}}$ is positive definite. The uniqueness of the kernel follows from (ii) and the fact that Riesz representers are unique. To see (v), we consider a function $f \in \mathcal{X}$ with $f \perp \operatorname{span}\{K(\cdot,\boldsymbol{x})\colon \boldsymbol{x} \in O\}$, i.e., $\langle f\,,\,K(\cdot,\boldsymbol{x})\rangle_{\mathcal{X}} = 0$ for all $\boldsymbol{x} \in O$. This implies $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in O$ and hence $f = 0$. The fact that (iii)–(iv) imply the existence of a RKHS is a consequence of the Moore–Aronszajn theorem [1]. □

In the following, we will consider the Kernel

$$K_s(\boldsymbol{x},\boldsymbol{y}) := \prod_{j=1}^{s} (2 - \max\{x_j, y_j\})$$

inducing the RKHS $\mathcal{X}_s$ and show that the corresponding inner product is given by

$$\langle f\,,\,g\rangle_{\mathcal{X}_s} := \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} \int_{[0,1]^{|\mathfrak{u}|}} \partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\partial_{\boldsymbol{x}_\mathfrak{u}} g(\boldsymbol{x}_\mathfrak{u};1)\,d\boldsymbol{x}_\mathfrak{u},$$

where $(\boldsymbol{x}_M;1)$ is a shorthand for the vector $\widetilde{\boldsymbol{x}} \in \mathbb{R}^s$ defined by $\widetilde{x}_j = x_j$ for $j \in \mathfrak{u}$ and $\widetilde{x}_j = 1$ for $j \notin \mathfrak{u}$. Moreover $\mathfrak{u} = \emptyset$ is interpreted as the point evaluation $\int_{[0,1]^{|\emptyset|}} \partial_{\boldsymbol{x}_\emptyset} f(\boldsymbol{x}_\emptyset;1)\partial_{\boldsymbol{x}_\emptyset} g(\boldsymbol{x}_\emptyset;1)\,d\boldsymbol{x}_\emptyset := f(\mathbf{1})g(\mathbf{1})$ with $\mathbf{1} := (1,\dots,1) \in [0,1]^s$.

Lemma 4. $\mathcal{X}_s$ *is a RKHS with kernel* $K_s(\cdot,\cdot)$ *that contains all functions* $f\colon [0,1]^s \to \mathbb{R}$ *such that* $\langle f\,,\,f\rangle_{\mathcal{X}_s} < \infty$.

Proof that $\mathcal{X}_s$ *is a RKHS.* First, we check that $\langle\cdot\,,\,\cdot\rangle_{\mathcal{X}_s}$ induces a Hilbert space. Lets define $\mathcal{X}_s$ as the set of all functions $f\colon [0,1]^s \to \mathbb{R}$ such that $\langle f\,,\,f\rangle_{\mathcal{X}_s} < \infty$. On $\mathcal{X}_s$, we immediately see that $\langle\cdot\,,\,\cdot\rangle_{\mathcal{X}_s}$ is bilinear, symmetric, and positive semi-definite. To see definiteness, we first show by induction on s for $f \in \mathcal{X}_s$ and $\boldsymbol{x} \in [0,1]^s$ that

$$f(\boldsymbol{x}) = \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} (-1)^{|\mathfrak{u}|} \int_{\prod_{j\in\mathfrak{u}}[x_j,1]} \partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\,d\boldsymbol{x}_\mathfrak{u}. \tag{1}$$

For $s = 1$, we have $f(x) = f(1) - \int_x^1 \partial_x f(y)\,dy$ and hence confirm (1). Assume (1) holds for $s-1$. Then, we have

$$\begin{aligned} f(\boldsymbol{x}) &= f(x_1,\dots,x_{s-1},1) - \int_{x_s}^1 \partial_{x_s} f(x_1,\dots,x_{s-1},y)\,dy \\ &= \sum_{\mathfrak{u}\subseteq\{1,\dots,s-1\}} \Big((-1)^{|\mathfrak{u}|} \int_{\prod_{j\in\mathfrak{u}}[x_j,1]} \partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\,d\boldsymbol{x}_\mathfrak{u} - (-1)^{|\mathfrak{u}|} \int_{x_s}^1 \int_{\prod_{j\in\mathfrak{u}}[x_j,1]} \partial_{\boldsymbol{x}_\mathfrak{u}} f((\boldsymbol{x}_\mathfrak{u};1),y)\,d\boldsymbol{x}_\mathfrak{u}\,dy\Big) \\ &= \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} (-1)^{|\mathfrak{u}|} \int_{\prod_{j\in\mathfrak{u}}[x_j,1]} \partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\,d\boldsymbol{x}_\mathfrak{u} \end{aligned}$$

which confirms (1) for s. With (1) at hand, we immediately see that $\langle f\,,\,f\rangle_{\mathcal{X}_s} = 0$ implies $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}\in[0,1]^s$ and hence $f = 0$. This shows definiteness of $\langle\cdot\,,\,\cdot\rangle_{\mathcal{X}_s}$ and hence confirms that $\mathcal{X}_s$ is a Hilbert space. Moreover, (1) shows that $f \mapsto f(\boldsymbol{x})$ is a bounded by the norm $\langle f\,,\,f\rangle_{\mathcal{X}_s}^{1/2}$ and hence $\mathcal{X}_s$ is a RKHS by Definition 2. □

Proof that $K_s(\cdot,\cdot)$ is the kernel of $\mathcal{X}_s$. First, we compute

$$\partial_{\boldsymbol{x}_\mathfrak{u}} K_s(\boldsymbol{x},\boldsymbol{y}) = \partial_{\boldsymbol{x}_\mathfrak{u}} \prod_{j=1}^s (2 - \max\{x_j,y_j\}) = \prod_{j\in\mathfrak{u}} -\mathbf{1}_{\{x_j\ge\}}(y_j) = (-1)^{|\mathfrak{u}|} \prod_{j\in\mathfrak{u}} \mathbf{1}_{\{x_j\ge\}}(y_j),$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function of the set $\{\cdot\}$. This shows immediately that $K_s(\cdot,\boldsymbol{y}) \in \mathcal{X}_s$ for all $\boldsymbol{y}\in[0,1]^s$. It remains to check that $\langle f\,,\,K(\cdot,\boldsymbol{y})\rangle_{\mathcal{X}_s} = f(\boldsymbol{y})$, i.e.,

$$\begin{aligned} f(\boldsymbol{y}) &= \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} (-1)^{|\mathfrak{u}|} \int_{[0,1]^{|\mathfrak{u}|}} \partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1) \prod_{j\in\mathfrak{u}} \mathbf{1}_{\{x_j\ge\}}(y_j)\,d\boldsymbol{x}_\mathfrak{u} \\ &= \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} (-1)^{|\mathfrak{u}|} \int_{\prod_{j\in\mathfrak{u}}[y_j,1]} \partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\,d\boldsymbol{x}_\mathfrak{u}. \end{aligned} \tag{2}$$

This, however, is exactly the identity (1). Hence, $K_s(\cdot,\cdot)$ is the unique kernel of $\mathcal{X}_s$. □

1.4. **The worst case error.** Given a point set $P\subset[0,1]^s$ with $|P| = n$, the worst-case error of the QMC rule $Q_n(f)$ defined by the set P is defined by

$$e_n(P,\mathcal{X}) := \sup_{\|f\|_\mathcal{X}\le 1} |I_s(f) - Q_n(f)|.$$

By linearity of the quadrature, we obtain immediately

$$|I_s(f) - Q_n(f)| \le e_n(P,\mathcal{X})\|f\|_\mathcal{X}.$$

In a RKHS, the worst case error is actually easy to compute.

Theorem 5. *Let $K\colon [0,1]^s\times[0,1]^s\to\mathbb{R}$ be a reproducing kernel that satisfies*

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

Then, there holds for the induced RKHS $\mathcal{X}$ that

$$\begin{aligned} e_n^2(P,\mathcal{X}) = \int_{[0,1]^s}\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y} - \frac{2}{n}\sum_{\boldsymbol{z}\in P}\int_{[0,1]^s} K(\boldsymbol{z},\boldsymbol{y})\,d\boldsymbol{y} \\ + \frac{1}{n^2}\sum_{\boldsymbol{z}\in P}\sum_{\boldsymbol{z}'\in P} K(\boldsymbol{z},\boldsymbol{z}'), \end{aligned}$$

where for the case $n = 0$, only the first term remains.

Proof. The reproducing kernel property shows for $f \in \mathcal{X}$

$$\frac{1}{n}\sum_{\boldsymbol{z}\in P} f(\boldsymbol{z}) = \frac{1}{n}\sum_{\boldsymbol{z}\in P}\langle f\,,\, K(\cdot,\boldsymbol{z})\rangle_{\mathcal{X}} = \langle f\,,\, \frac{1}{n}\sum_{\boldsymbol{z}\in P} K(\cdot,\boldsymbol{z})\rangle_{\mathcal{X}}.$$

Next, we show that $f \mapsto I_s(f)$ is a bounded linear operation on $\mathcal{X}$. To that end, observe

$$|I_s(f)| = \Big|\int_{[0,1]^s} \langle K(\cdot,\boldsymbol{x})\,,\, f\rangle_{\mathcal{X}}\, d\boldsymbol{x}\Big| \le \|f\|_{\mathcal{X}} \int_{[0,1]^s} \|K(\cdot,\boldsymbol{x})\|_{\mathcal{X}}\, d\boldsymbol{x}.$$

Since $\|K(\cdot,\boldsymbol{x})\|_{\mathcal{X}}^2 = \langle K(\cdot,\boldsymbol{x})\,,\, K(\cdot,\boldsymbol{x})\rangle_{\mathcal{X}} = K(\boldsymbol{x},\boldsymbol{x})$, the assumption in the statement of the theorem shows $\int_{[0,1]^s} \|K(\cdot,\boldsymbol{x})\|_{\mathcal{X}}\, d\boldsymbol{x} < \infty$ and hence proves boundedness of $f \mapsto I_s(f)$. Thus, we find a representer $R \in \mathcal{X}$ with $I_s(f) = \langle R\,,\, f\rangle_{\mathcal{X}}$ for all $f \in \mathcal{X}$. This representer can be characterized easily with the reproducing property of K, i.e.

$$R(\boldsymbol{y}) = \langle R\,,\, K(\cdot,\boldsymbol{y})\rangle_{\mathcal{X}} = I_s(K(\cdot,\boldsymbol{y})) = \int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x}.$$

Hence, we obtain

$$I_s(f) = \int_{[0,1]^s} f(\boldsymbol{x})\, dx = \langle f\,,\, \int_{[0,1]^s} K(\boldsymbol{x},\cdot)\, d\boldsymbol{x}\rangle_{\mathcal{X}}.$$

Subtraction of the identities for $I_s(f)$ and $Q_n(f)$ reveals

$$I_s(f) - Q_n(f) = \langle f\,,\, \int_{[0,1]^s} K(\cdot,\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{n}\sum_{\boldsymbol{z}\in P} K(\cdot,\boldsymbol{z})\rangle_{\mathcal{X}} = \langle f\,,\, \zeta\rangle_{\mathcal{X}},$$

where $\zeta(\boldsymbol{y}) := \int_{[0,1]^s} K(\boldsymbol{y},\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{n}\sum_{\boldsymbol{z}\in P} K(\boldsymbol{y},\boldsymbol{z})$ is the error representer. Hence, $e_n(P,\mathcal{X}) = \|\zeta\|_{\mathcal{X}}$ by definition. Computing the norm

$$\begin{aligned}\|\zeta\|_{\mathcal{X}}^2 = &\int_{[0,1]^s}\int_{[0,1]^s} \langle K(\cdot,\boldsymbol{x})\,,\, K(\cdot,\boldsymbol{x}')\rangle_{\mathcal{X}}\, d\boldsymbol{x}\, d\boldsymbol{x}' \\ &- \frac{2}{n}\sum_{\boldsymbol{z}\in P}\int_{[0,1]^s} \langle K(\cdot,\boldsymbol{x})\,,\, K(\cdot,\boldsymbol{z})\rangle_{\mathcal{X}}\, d\boldsymbol{x} \\ &+ \frac{1}{n^2}\sum_{\boldsymbol{z},\boldsymbol{z}'\in P} \langle K(\cdot,\boldsymbol{z}')\,,\, K(\cdot,\boldsymbol{z})\rangle_{\mathcal{X}}\, d\boldsymbol{x}\end{aligned}$$

concludes the proof.

□

Using the formula from the previous Theorem, a lengthy but straighforward computation shows that the kernel $K_s(\cdot,\cdot)$ from above admits the worst case error

$$e_n^2(P,\mathcal{X}_s) = \Big(\frac{4}{3}\Big)^s - \frac{2}{n}\sum_{\boldsymbol{z}\in P}\prod_{j=1}^{s}\Big(\frac{3-z_j^2}{2}\Big) + \frac{1}{n^2}\sum_{\boldsymbol{z},\boldsymbol{z}'\in P}\prod_{j=1}^{s}(2-\max\{z_j,z_j'\}).$$

1.5. **Geometric discrepancy.** In the proof of Theorem 5, we derived the estimate

$$|I_s(f) - Q_n(f)| = |\langle f\,,\, \zeta\rangle_{\mathcal{X}}|, \tag{3}$$

where for $\mathcal{X} = \mathcal{X}_s$, the error representer ζ takes the form

$$\begin{aligned}\zeta(\boldsymbol{y}) &= \int_{[0,1]^s} K(\boldsymbol{y},\boldsymbol{x})\,d\boldsymbol{x} - \frac{1}{n}\sum_{\boldsymbol{z}\in P} K(\boldsymbol{y},\boldsymbol{z}) \\ &= \prod_{j=1}^{s}\Big(\frac{3-y_j^2}{2}\Big) - \frac{1}{n}\sum_{\boldsymbol{z}\in P}\prod_{j=1}^{s}(2-\max\{y_j,z_j\}),\end{aligned}$$

where P is the QMC point-set and $\boldsymbol{y}\in[0,1]^s$. We may compute the first mixed partial derivatives for any subset $\mathfrak{u}\subseteq\{1,\dots,s\}$

$$\partial_{\boldsymbol{x}_\mathfrak{u}}\zeta(\boldsymbol{x}_\mathfrak{u};1) = (-1)^{|\mathfrak{u}|}\Big(\prod_{j\in\mathfrak{u}} x_j - \frac{1}{n}\sum_{\boldsymbol{z}\in P}\prod_{j\in\mathfrak{u}}\mathbf{1}_{\{\geq z_j\}}(x_j)\Big).$$

With the local discrepancy function Δ_P in s-dimensions

$$\Delta_P(\boldsymbol{x}) := \frac{1}{n}\sum_{\boldsymbol{z}\in P}\prod_{j=1}^{s}\mathbf{1}_{\{\geq z_j\}}(x_j) - \prod_{j=1}^{s} x_j,$$

we may write

$$\partial_{\boldsymbol{x}_\mathfrak{u}}\zeta(\boldsymbol{x}_\mathfrak{u};1) = (-1)^{|\mathfrak{u}|+1}\Delta_P(\boldsymbol{x}_\mathfrak{u};1).$$

Since we already know the scalar product $\langle\cdot\,,\,\cdot\rangle_{\mathcal{X}_s}$, we obtain from (3)

$$|I_s(f)-Q_n(f)| = \Big|\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}(-1)^{|\mathfrak{u}|}\int_{[0,1]^{|\mathfrak{u}|}}\partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\Delta_P(\boldsymbol{x}_\mathfrak{u};1)\,d\boldsymbol{x}_\mathfrak{u}\Big|.$$

Thus, by application of Hölder's inequality, we obtain the following theorem:

Theorem 6 (Koksma-Hlawka inequality)**.** *There holds*

$$\begin{aligned}|I_s(f)-Q_n(f)| \leq \Big(\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}\Big(\int_{[0,1]^{|\mathfrak{u}|}}\Big|\partial_{\boldsymbol{x}_\mathfrak{u}} f(\boldsymbol{x}_\mathfrak{u};1)\Big|^p\,d\boldsymbol{x}_\mathfrak{u}\Big)^{p'/p}\Big)^{1/p'} \\ \cdot\Big(\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}\Big(\int_{[0,1]^{|\mathfrak{u}|}}\Big|\Delta_P(\boldsymbol{x}_\mathfrak{u};1)\Big|^q\,d\boldsymbol{x}_\mathfrak{u}\Big)^{q'/q}\Big)^{1/q'},\end{aligned}$$

where $1<p,p',q,q'<\infty$ and $1/p+1/q=1=1/p'+1/q'$. The estimate also holds for one or more of p,p',q,q' equal to infinity with the obvious modifications to the norms.

Proof. We already derived an estimate of the form

$$|I_s(f)-Q_n(f)| \leq \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}\langle a_\mathfrak{u}\,,\,b_\mathfrak{u}\rangle_{L^2([0,1]^s)}$$

for particular functions $a_\mathfrak{u}$ and $b_\mathfrak{u}$. Two applications of Hölder's inequalities show

$$|I_s(f)-Q_n(f)| \leq \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}\|a_\mathfrak{u}\|_{L^p}\|b_\mathfrak{u}\|_{L^q} \leq \Big(\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}\|a_\mathfrak{u}\|_{L^p}^{p'}\Big)^{1/p'}\Big(\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}}\|b_\mathfrak{u}\|_{L^q}^{q'}\Big)^{1/q'}.$$

This concludes the proof. □

The previous theorem shows that the integration error is controlled by a factor which depends only on the set P. This discrepancy term has a geometric interpretation in the sense that

$$\sum_{\boldsymbol{z}\in P}\prod_{j=1}^{s}\mathbf{1}_{\{\geq z_j\}}(x_j)$$

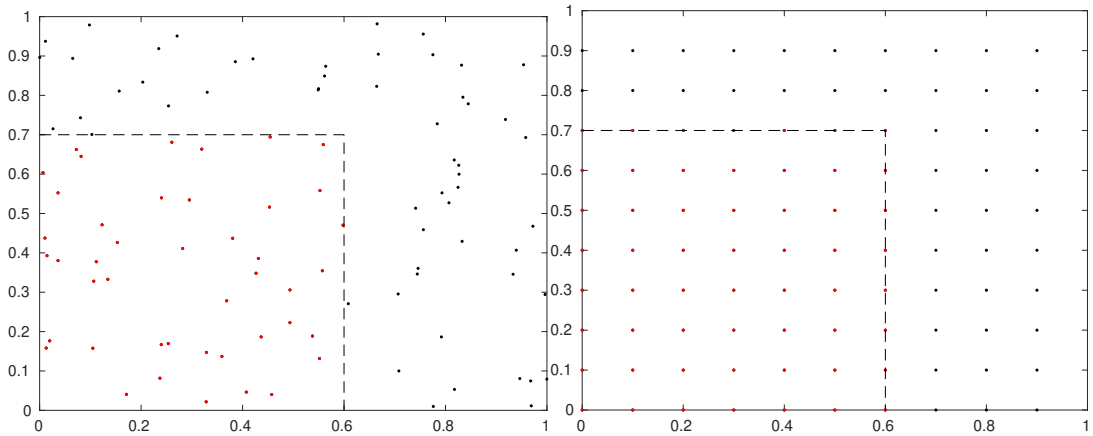

FIGURE 1. On the left hand side we see a sample of uniformly distributed points P with $n = 100$. There are 44 points (red) in the box $[0, 0.6] \times [0, 0.7]$ (a share of 0.44) and the actual volume of the box is 0.42. Hence, we obtain $\Delta_P(0.6, 0.7) = 0.44 - 0.42 = 0.02$. On the right-hand side, we see a tensor point-set P with $n = 100$. There are 56 points in the box, which results in a discrepancy of $\Delta_P(0.6, 0.7) = 0.14$.

gives the number of points in P which are in the axis parallel box $[\mathbf{0}, \boldsymbol{x}]$ and $\prod_{j=1}^{s} x_j$ is just the volume of that box. Hence, $\Delta_P(\boldsymbol{x})$ compares the actual volume of the box $[\mathbf{0}, \boldsymbol{x}]$ with the share of points that are located in this box (see Figure 1 for an example). It is rather obvious, that tensor points are not well-suited to minimize the discrepancy.

1.6. **Stratification.** Obviously, random points will not always be well distributed in the unit cube and thus fail to minimize the discrepancy. One way to improve this situation is to divide the unit cube into smaller disjoint parts $[0, 1]^s = \bigcup_{\ell=1}^{L} D_\ell$ (so called *strata*) and run a MC quadrature in each of those parts. The stratified MC quadrature rule reads

$$Q_{\text{strat}}(f) := \sum_{\ell=1}^{L} \frac{|D_\ell|}{n_\ell} \sum_{i=0}^{n_\ell} f(\boldsymbol{x}_{i,\ell}),$$

where the $\boldsymbol{x}_{i,\ell}$ are chosen independently and uniformly in D_ℓ. Given a budget of function evaluations $N \in \mathbb{N}$, it makes sense to choose $n_\ell := \lceil N|D_\ell| \rceil$ proportionally to the size of D_ℓ. However, other choices (e.g., proportionally to the variance of f within D_ℓ) are usefull as well.

We immediately see that Q_{strat} is unbiased, i.e.,

$$\mathbb{E}(Q_{\text{strat}}(f)) = \sum_{\ell=1}^{L} \frac{|D_\ell|}{n_\ell} \sum_{i=0}^{n_\ell} \mathbb{E}(f(\boldsymbol{x}_{i,\ell})) = \sum_{\ell=1}^{L} |D_\ell| \frac{1}{|D_\ell|} \int_{D_\ell} f \, dx = I_s(f).$$

To compute the quadrature error, we proceed as in the standard MC case, i.e.,

$$\mathbb{E}|Q_{\text{strat}}(f) - I_s(f)|^2 = \sum_{\ell,\ell'=1}^{L} \mathbb{E}\Big(\frac{|D_\ell|}{n_\ell} \sum_{i=0}^{n_\ell} f(\boldsymbol{x}_{i,\ell}) - \int_{D_\ell} f \, dx \Big) \Big(\frac{|D_{\ell'}|}{n_{\ell'}} \sum_{i=0}^{n_{\ell'}} f(\boldsymbol{x}_{i,\ell'}) - \int_{D_{\ell'}} f \, dx \Big).$$

Independence of the sample points shows

$$\begin{aligned}
&\mathbb{E}|Q_{\mathrm{strat}}(f) - I_s(f)|^2 \\
&\quad = \sum_{\ell\neq\ell'=1}^{L} \mathbb{E}\Big(\frac{|D_\ell|}{n_\ell}\sum_{i=0}^{n_\ell} f(\boldsymbol{x}_{i,\ell}) - \int_{D_\ell} f\,dx\Big)\mathbb{E}\Big(\frac{|D_{\ell'}|}{n_{\ell'}}\sum_{i=0}^{n_{\ell'}} f(\boldsymbol{x}_{i,\ell'}) - \int_{D_{\ell'}} f\,dx\Big). \\
&\qquad + \sum_{\ell=1}^{L} \mathbb{E}\Big(\frac{|D_\ell|}{n_\ell}\sum_{i=0}^{n_\ell} f(\boldsymbol{x}_{i,\ell}) - \int_{D_\ell} f\,dx\Big)^2.
\end{aligned}$$

The terms in the second sum correspond to the standard MC error which we already computed in Theorem 1. The terms in the first double sum disappear because of the unbiasedness of the MC quadrature in each strata D_ℓ. Thus, we end up with

$$\mathbb{E}|Q_{\mathrm{strat}}(f) - I_s(f)|^2 = \sum_{\ell=1}^{L} |D_\ell|^2 \frac{\sigma_\ell^2}{n_\ell} \leq \frac{1}{N}\sum_{\ell=1}^{L} |D_\ell|\sigma_\ell^2,$$

where $\sigma_\ell^2 := |D_\ell|^{-1}\int_{D_\ell} f^2\,dx - \big(|D_\ell|^{-1}\int_{D_\ell} f\,dx\big)^2$. Obviously, this can be an improvement over plain MC as $\sigma_\ell = 0$ for functions f which are constant in each of the D_ℓ.

With $\mu_\ell := |D_\ell|^{-1}\int_{D_\ell} f\,dx$, we also see that the error estimate is never worse than for plain MC. To that end, compute

$$\begin{aligned}
\mathrm{Var}(f)^2 &= \int_{[0,1]^s} f^2\,dx - \Big(\int_{[0,1]^s} f\,dx\Big)^2 \\
&= \sum_{\ell=1}^{L} |D_\ell|\frac{1}{|D_\ell|}\int_{D_\ell} f^2\,dx - \sum_{\ell,\ell'=1}^{L} |D_\ell||D_{\ell'}|\mu_\ell\mu_{\ell'} \\
&= \sum_{\ell=1}^{L} |D_\ell|\sigma_\ell^2 + \sum_{\ell=1}^{L} |D_\ell|\mu_\ell^2 - \sum_{\ell,\ell'=1}^{L} |D_\ell||D_{\ell'}|\mu_\ell\mu_{\ell'}.
\end{aligned}$$

There holds with Hölder

$$\sum_{\ell,\ell'=1}^{L} |D_\ell||D_{\ell'}|\mu_\ell\mu_{\ell'} \leq \Big(\sum_{\ell,\ell'=1}^{L} |D_\ell||D_{\ell'}|\mu_\ell^2\Big)^{1/2}\Big(\sum_{\ell,\ell'=1}^{L} |D_\ell||D_{\ell'}|\mu_{\ell'}^2\Big)^{1/2} = \sum_{\ell=1}^{L} |D_\ell|\mu_\ell^2.$$

This, together with the above shows $\mathrm{Var}(f)^2 \geq \sum_{\ell=1}^{L} |D_\ell|\sigma_\ell^2$. Hence, the stratified MC error estimate is never worse than the plain MC error estimate from Theorem 1. However, the rate of $N^{-1/2}$ is still the best we can hope for.

1.7. **Digital nets.** In the following, we will discuss an important class of QMC point-sets P_m, which perform better than the MC methods.

Definition 7 (Digital (t,m,s)-net)**.** *A dyadic interval $I_{\boldsymbol{m},\boldsymbol{a}} \subseteq [0,1]^s$ in s-dimensions is defined by multi-indices $\boldsymbol{a},\boldsymbol{m} \in \mathbb{N}_0^s$ with $0 \leq a_j \leq 2^{m_j-1}$ for all $j = 1,\dots,s$ and*

$$I_{\boldsymbol{m},\boldsymbol{a}} := \prod_{j=1}^{s} [a_j 2^{-m_j}, (a_j+1)2^{-m_j}).$$

Note that $|I_{\boldsymbol{m},\boldsymbol{a}}| = 2^{-|\boldsymbol{m}|}$ and for fixed $\boldsymbol{m}$, $\bigcup_{0\leq\boldsymbol{a}\leq 2^{\boldsymbol{m}-1}} I_{\boldsymbol{m},\boldsymbol{a}} = [0,1)^s$ form a partition of the unit cube. For $t,m \in \mathbb{N}_0$ with $t \leq m$, a point set $P \subseteq [0,1)^s$ with $\#P = 2^m$ is called a digital (t,m,s)-net if

$$\#(P \cap I_{\boldsymbol{m},\boldsymbol{a}}) = 2^t$$

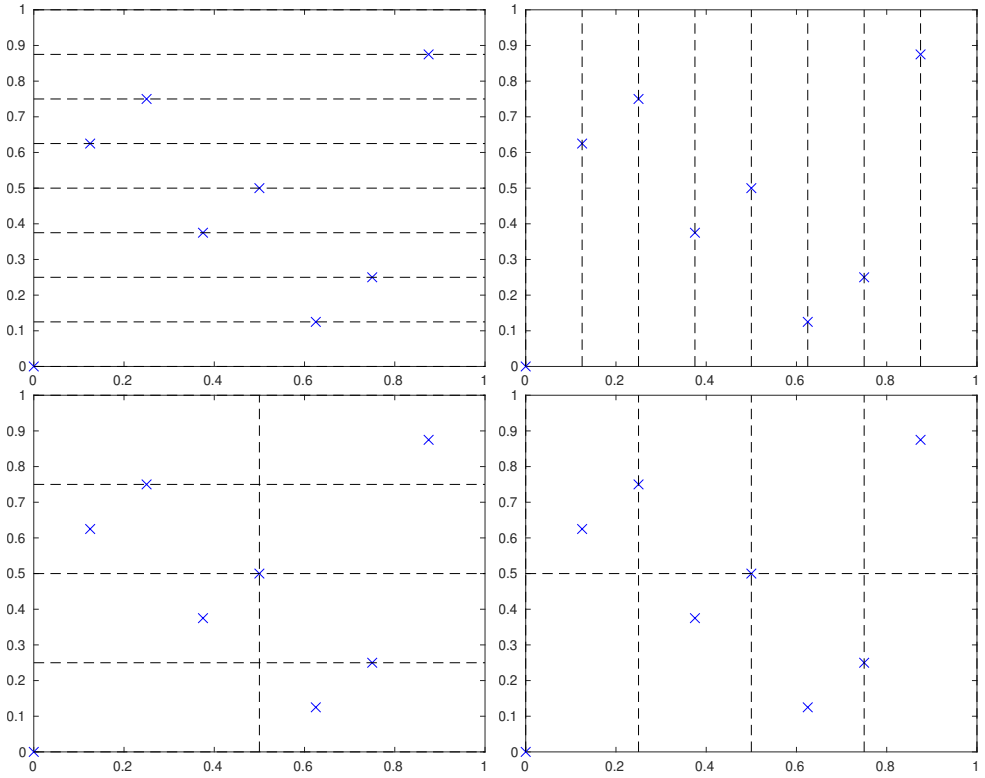

FIGURE 2. The geometric properties of a digital $(0, 3, 2)$-net. The dyadic intervals are indicated with dashed lines. Each of them contains exactly one point of the digital net.

for all $\boldsymbol{m} \in \mathbb{N}_0^s$ *with* $|\boldsymbol{m}| = m - t$. *This condition suggests that a* (t, m, s)*-net is well distributed in the unit cube and the distribution properties are better if* t *is small since the dyadic intervals can be chosen finer (see Figure 2 for an illustration). Note that digital nets can be defined in other bases than the binary base* 2.

One way to generate a digital (t, m, s)-net is to define matrices $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_s \in \{0, 1\}^{m \times m}$ such that

$$\begin{pmatrix} (\boldsymbol{C}_1)_{1:m_1,:} \\ \vdots \\ (\boldsymbol{C}_s)_{1:m_s,:} \end{pmatrix}$$

has full rank for all $\boldsymbol{m} = (m_1, \ldots, m_s) \in \mathbb{N}_0^s$ with $|\boldsymbol{m}| = m_1 + \ldots + m_s = m - t$ and some $0 \leq t \leq m$. This is also sometimes called the (t, m, s)-net property for matrices. An example for $s = 2$ and $t = 0$ is given by:

$$\boldsymbol{C}_1 := \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & \ldots & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{C}_2 := \begin{pmatrix} 0 & 0 & \ldots & 1 \\ 0 & \ldots & 1 & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \ldots & 0 \end{pmatrix}.$$

We generate a point-set $P = \{\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{2^m-1}\}$ as follows:
For $\ell = 0, \ldots, 2^m - 1$ do:

(1) Write ℓ in its base-2 representation, i.e.,

$$\ell = \ell_0 + 2\ell_1 + 2^2\ell_2 + \ldots + 2^{m-1}\ell_{m-1} \tag{4}$$

for coefficients $\ell_i \in \{0,1\}$.

(2) Compute for $j = 1, 2, \ldots, s$ modulo 2

$$\boldsymbol{y}_j := \boldsymbol{C}_j \boldsymbol{\ell},$$

where $\boldsymbol{\ell} = (\ell_0, \ell_1, \ldots, \ell_{m-1})$ and $\boldsymbol{y}_j \in \{0,1\}^m$.

(3) Set

$$\boldsymbol{z}_\ell := (z_{\ell,1}, \ldots, z_{\ell,s}) \quad \text{with} \quad z_{\ell,j} := y_{j,0}2^{-1} + y_{j,1}2^{-2} + \ldots y_{j,m-1}2^{-m}.$$

The resulting point-set P_m is called a digital (t,m,s)-net. It has some rather nice geometric properties.

Lemma 8. *Given $\boldsymbol{m} \in \mathbb{N}_0^s$ with $|\boldsymbol{m}| = m - t \in \mathbb{N}$ and $\boldsymbol{a}$ with $0 \le a_j \le 2^{m_j} - 1$, define the interval $I_{\boldsymbol{m},\boldsymbol{a}} := \prod_{j=1}^s [a_j 2^{-m_j}, (a_j+1)2^{-m_j}) \subseteq [0,1)^s$. Then, there holds $|P_m \cap I_{\boldsymbol{m},\boldsymbol{a}}| = 2^t$ for all choices of $\boldsymbol{m}$ and $\boldsymbol{a}$. Thus, P_m is a digital (t,m,s)-net according to Definition 7.*

Proof. We recall that every point $z \in P_m$ is of the form

$$\boldsymbol{z}_\ell := (z_{\ell,1}, \ldots, z_{\ell,s}) \quad \text{with} \quad z_{\ell,j} := y_{j,0}2^{-1} + y_{j,1}2^{-2} + \ldots y_{j,m-1}2^{-m}.$$

We write $a_j = a_{j,m_j} + a_{j,m_j-1}2 + \ldots + a_{j,1}2^{m_j-1}$ and observe that $z_{\ell,j} \in [a_j 2^{-m_j}, (a_j+1)2^{-m_j})$ implies that

$$\begin{aligned} a_{j,m_j}2^{-m_j} + a_{j,m_j-2}2^{-m_j+1} + \ldots + a_{j,1}2^{-1} &\le y_{j,1}2^{-1} + y_{j,2}2^{-2} + \ldots y_{j,m-1}2^{-m+1} \\ &< (a_{j,m_j}+1)2^{-m_j} + a_{j,m_j-1}2^{-m_j+1} + \ldots + a_{j,1}2^{-1}. \end{aligned}$$

By comparing the digits, it is obvious that $a_{j,i} = y_{j,i}$ for all $i = 1, \ldots, m_j$ in order to satisfy the above estimate. Since $|\boldsymbol{m}| = m - t$, the restriction $\boldsymbol{z}_\ell \in I_{\boldsymbol{m},\boldsymbol{a}}$ determines $m - t$ digits in total. In other words, the number of points in $I_{\boldsymbol{m},\boldsymbol{a}}$ is equivalent to the number of solutions $\boldsymbol{\ell} \in \{0,1\}^m$ of

$$\begin{pmatrix} a_{1,1} \\ \vdots \\ a_{1,m_1} \\ a_{2,1} \\ \vdots \\ a_{2,m_2} \\ \vdots \\ a_{s,1} \\ \vdots \\ a_{s,m_2} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{C}_1)_{1:m_1,:} \\ (\boldsymbol{C}_2)_{1:m_2,:} \\ \vdots \\ (\boldsymbol{C}_2)_{1:m_s,:} \end{pmatrix} \boldsymbol{\ell} \mod 2.$$

Since, by definition of the $\boldsymbol{C}_j$, the combined matrix above in $\mathbb{R}^{(m-t)\times m}$ has full rank, the solution space is a t-dimensional subspace of $\{0,1\}^m$ (note that we carried out all calculations in the field $\mathbb{Z}_2$). Such a subspace has precisely 2^t elements. This can be seen by noticing that a t-dimensional subspace is spanned by t different basis functions $b_1, \ldots, b_t$. Each element of the subspace writes as $\sum_{j=1}^t \gamma_j b_j$ with $\gamma_j \in \mathbb{Z}_2$. This leaves 2^t possible choices of coefficients $\gamma_1, \ldots, \gamma_t$. Hence, we conclude the proof. □

Lemma 9. *For a digital net as constructed above, there holds $\|\Delta_P\|_{L^\infty([0,1]^2)} \lesssim m^{s-1}2^{-m+t}$ for all $m \in \mathbb{N}$.*

Proof. We give the proof for $s = 2$, but the idea carries over to the general case. Let $\boldsymbol{x} \in [0,1)^2$ define the box $B_{\boldsymbol{x}} := [0, x_1) \times [0, x_2)$ and, without loss of generality, assume that $x_1 \geq x_2$. We approximate $B_{\boldsymbol{x}}$ with a union of the intervals $I_{\boldsymbol{m},\boldsymbol{a}}$ which we denote by I. Let $m_1 \leq m - t$ be minimal such that $2^{-m_1} < x_1$. If no such m_1 exists, we define $I = \emptyset$ and note $|B_{\boldsymbol{x}} \setminus I| \leq 2^{-m+t}$. Otherwise, define $m_2 := m - t - m_1 > 0$. If $2^{-m_2} \geq x_2$, we again set $I = \emptyset$ and note $|B_{\boldsymbol{x}} \setminus I| \leq 2 \cdot 2^{-m+t}$.

We see that we can fit the intervals $I_{(m_1,m_2),\boldsymbol{a}}$ in $B_{\boldsymbol{x}}$ for all $\boldsymbol{a} = (0, j)$ with $2^{j-m_2} < x_2$. Lets call the union of these intervals I_1. By construction there holds

$$|[0, 2^{-m_1}) \times [0, x_2) \setminus I_1| \leq 2^{-m_1-m_2} = 2^{-m+t}.$$

For the remaining box $[2^{-m_1}, x_1) \times [0, x_2)$ we may repeat the argument until the remainder has a volume less than 2^{-m}. This is reached after at most m iterations. Thus, we can constructed a disjoint union $I \subseteq B_{\boldsymbol{x}}$ of intervals $I_{\boldsymbol{m},\boldsymbol{a}}$ such that

$$|B_{\boldsymbol{x}} \setminus I| \leq m 2^{-m+t}.$$

By Lemma 8, this implies that $\Delta_P(\boldsymbol{x}) \geq |I| - |B_{\boldsymbol{x}}| \geq -m2^{-m+t}$. Analogously, we can construct a disjoint union $\overline{I}$ with $\overline{I} \supseteq B_{\boldsymbol{x}}$ such that $|\overline{I} \setminus B_{\boldsymbol{x}}| \leq m2^{-m+t}$ and hence show that $|\Delta_P(\boldsymbol{x})| \leq m2^{-m+t}$. This concludes the proof. □

The combination of Lemma 9 and the Koksma-Hlawka inequality (Theorem 6 with $q = q = \infty$ and $p = p' = 1$) shows that the quadrature error of a digital net is bounded by

$$|I_s(f) - Q_n(f)| \lesssim \log(n)^{s-1} n^{-1} \sum_{\mathfrak{u} \subseteq \{1,\dots,s\}} \int_{[0,1]^{|\mathfrak{u}|}} \left|\partial_{\boldsymbol{x}_\mathfrak{u}} f(x_\mathfrak{u}; 1)\right| d\boldsymbol{x}_\mathfrak{u}.$$

On the first glance, this is a big improvement over the MC method (double the convergence rate) with the drawback of a much higher regularity requirement on f. The added regularity is not really a restriction but almost necessary (if high-dimensional functions are not regular in some sense, there is no chance in doing any computations at all. The unit cube $[0,1]^s$ is just way to big for large s). However, the factor $\log(n)^{s-1}$ is still bad in high dimensions. The quantity $\log(n)^{s-1}n^{-1}$ decreases only for $n \geq e^{s-1}$. Hence, we need some additional regularity which tells us the importance of different dimensions. This is achieved by introducing weighted Hilbert spaces. While this theory also exists for digital nets, we will consider a simpler approach.

1.8. **Lattice rules and weighted spaces.** It turns out that the Hilbert space $\mathcal{X}_s$ does not contain enough regularity information to obtain dimension independent quadrature rules. To that end, we introduce unanchored Sobolev spaces. In one dimension, it is clear that

$$\langle f\,,\, g\rangle_\gamma := \int_0^1 f(x)\,dx \int_0^1 g(x)\,dx + \frac{1}{\gamma}\int_0^1 f'(x)g'(x)\,dx,$$

is a scalar product of a RKHS. The corresponding kernel is given by

$$K_{1,\gamma}(x, y) := 1 + \gamma\eta(x, y),$$

where $\eta(x, y) = B_2(|x-y|)/2+(x-1/2)(y-1/2)$ and $B_2(x) := x^2-x+1/6$ is the Bernoulli polynomial of degree 2, which has the useful property $\int_0^1 B_2(|x-y|)\,dy = \int_0^1 B_2(y)\,dy = 0$. Similar to before, we introduce the s-dimensional kernel via

$$K_{s,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{fu \subseteq \{1,\dots,s\}} \gamma_\mathfrak{u} \prod_{j \in \mathfrak{u}} \eta(\boldsymbol{x}, \boldsymbol{y})$$

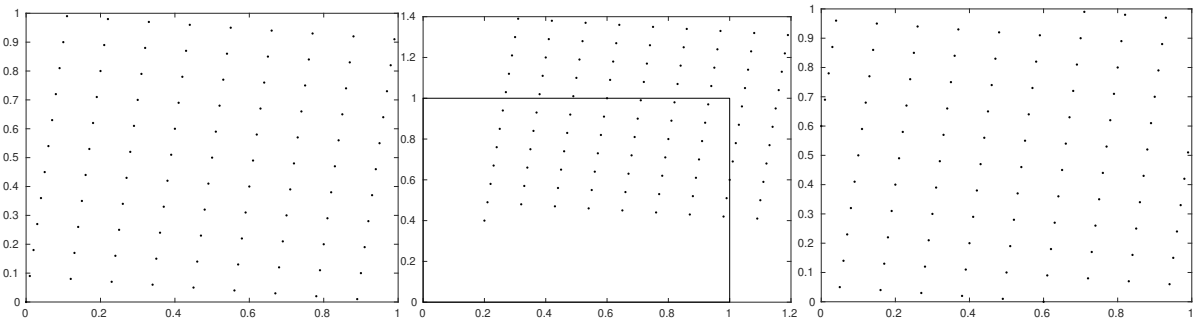

FIGURE 3. Lattice rule with $n = 100$ and $\boldsymbol{z} = (1, 9)$. The shifted version is generated by applying a random shift (here $S = (0.2, 0.3)$) and then taking the fractional part of the points.

(Note that in case $\gamma_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \gamma_j$ for some $\gamma_j > 0$, there holds $K_{s,\boldsymbol{\gamma}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^{s} K_{1,\gamma_j}(\boldsymbol{x}, \boldsymbol{y})$.) The coefficients $\gamma_{\mathfrak{u}} > 0$ are called *weights*. Heuristically speaking, the weights denote the *importance* of a dimension (or combination of dimensions). Again, we can check that $K_{s,\boldsymbol{\gamma}}(\cdot, \cdot)$ induces the RKHS $\mathcal{X}_{s,\boldsymbol{\gamma}}$ with the inner product

$$\langle f\ ,\ g\rangle_{s,\boldsymbol{\gamma}} = \sum_{fu\subseteq\{1,\ldots,s\}} \gamma_{\mathfrak{u}}^{-1} \int_{[0,1]^{|\mathfrak{u}|}} \Big(\int_{[0,1]^{s-|\mathfrak{u}|}} \partial_{\boldsymbol{x}_{\mathfrak{u}}} f(\boldsymbol{x})\, d\boldsymbol{x}_{\mathfrak{u}^c} \Big) \Big(\int_{[0,1]^{s-|\mathfrak{u}|}} \partial_{\boldsymbol{x}_{\mathfrak{u}}} g(\boldsymbol{x})\ d\boldsymbol{x}_{\mathfrak{u}^c} \Big)\, d\boldsymbol{x}_{\mathfrak{u}},$$

where $\mathfrak{u}^c := \{1, \ldots, s\} \setminus \mathfrak{u}$. The worst-case error for integration in $\mathcal{X}_{s,\boldsymbol{\gamma}}$ is again given by Theorem 5. Now, the quantities in the formula are given by

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

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

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

These formulae already hint at the possibility of bounding the worst-case error independently of the dimension, if only that weights converge to zero sufficiently fast. To prove this, we introduce *Lattice rules*.

An n-point rank-one lattice rule in s-dimensions is a QMC method with points P generated by

$$\boldsymbol{z}_i := \frac{i\boldsymbol{z}}{n} \bmod 1 := \frac{i\boldsymbol{z}}{n} - \left\lfloor \frac{i\boldsymbol{z}}{n} \right\rfloor, \quad \text{for all } i = 0, 1, \ldots, n-1,$$

where $\boldsymbol{z} \in \mathbb{Z}^s$ is an s-dimensional integer vector having no factor in common with n (the generating vector). To obtain a statistical error estimate and to avoid pathological cases, it is often useful to apply a random shift to a lattice rule. This means that a shift $S \in [0, 1]^s$ is picked uniformly and randomly and we consider the shifted lattice

$$P + S := \big\{ \boldsymbol{z} + S \bmod 1\ :\ \boldsymbol{z} \in P \big\}.$$

The corresponding quadrature rule is denoted by $Q_n(S, f)$. This procedure is illustrated in Figure 3.

From the theory of worst-case errors, we obtain immediately

$$|I_s(f) - Q_n(S, f)| \le e_n(P + S, \mathcal{X}) \|f\|_{\mathcal{X}}.$$

This means that the expected worst-case error (expectation is taken over all possible random shifts $S \in [0,1]^s$) is given by

$$\sqrt{\mathbb{E}|I_s(f) - Q_n(S,f)|^2} \le e_n^{\mathrm{sh}}(P,\mathcal{X})\|f\|_{\mathcal{X}},$$

where

$$e_n^{\mathrm{sh}}(P,\mathcal{X})^2 := \int_{[0,1]^s} e_n(P+S,\mathcal{X})^2 \, dS.$$

Remark 10. *The shift-averaged error estimate might seem useless for practical applications as one still has to choose a concrete shift S. However, the Chebyshev inequality shows for the random variable $X(S) := |I_s(f) - Q_n(S,f)|$ that*

$$\mathbb{P}(|X - \mathbb{E}(X)| \ge k) \le \frac{\mathrm{Var}(X)}{k^2} \le \frac{\mathbb{E}(X^2)}{k^2} \le \frac{e_n^{\mathrm{sh}}(P,\mathcal{X})^2\|f\|_{\mathcal{X}}^2}{k^2}.$$

With $k := Ce_n^{\mathrm{sh}}(P,\mathcal{X})\|f\|_{\mathcal{X}}$ for some $C > 0$ and $\mathbb{E}(X) \le \sqrt{\mathbb{E}(X^2)} \le e_n^{\mathrm{sh}}(P,\mathcal{X})\|f\|_{\mathcal{X}}$, the triangle inequality and the above estimate show

$$\mathbb{P}(X \ge (C+1)e_n^{\mathrm{sh}}(P,\mathcal{X})\|f\|_{\mathcal{X}}) \le \mathbb{P}(|X - \mathbb{E}(X)| \ge k) \le 1/C^2.$$

Hence, for a random shift S, we have a probability of at least $1 - C^{-2}$ that

$$|I_s(f) - Q_n(S,f)| \le (C+1)e_n^{\mathrm{sh}}(P,\mathcal{X})\|f\|_{\mathcal{X}}.$$

Lemma 11. *The shift-averaged worst-case error e_n^{sh} is given by*

$$e_n^{\mathrm{sh}}(P,\mathcal{X})^2 = -\int_{[0,1]^s}\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y} + \frac{1}{n^2}\sum_{\boldsymbol{z}\in P}\sum_{\boldsymbol{z}'\in P} K^{\mathrm{sh}}(\boldsymbol{z},\boldsymbol{z}'),$$

where

$$K^{\mathrm{sh}}(\boldsymbol{x},\boldsymbol{y}) := \int_{[0,1]^s} K(\boldsymbol{x}+S \bmod 1, \boldsymbol{y}+S \bmod 1)\,dS.$$

Proof. With the formula for e_n from Theorem 5, we obtain

$$\begin{aligned}\int_{[0,1]^s} e_n(P+S,\mathcal{X})^2\,dS = \int_{[0,1]^s}\int_{[0,1]^s} K(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{y} - \frac{2}{n}\sum_{\boldsymbol{z}\in P}\int_{[0,1]^s}\int_{[0,1]^s} K(\boldsymbol{z}+S \bmod 1,\boldsymbol{y})\,d\boldsymbol{y}\,dS \\ + \frac{1}{n^2}\sum_{\boldsymbol{z}\in P}\sum_{\boldsymbol{z}'\in P}\int_{[0,1]^s} K(\boldsymbol{z}+S \bmod 1,\boldsymbol{z}'+S \bmod 1)\,dS.\end{aligned}$$

A change of variables in the second term shows that it equals the integral of the first term. This concludes the proof. □

The function K^{sh} is actually a kernel of a RKHS with the additional shift-invariance property

$$K^{\mathrm{sh}}(\boldsymbol{x}+S \bmod 1, \boldsymbol{y}+S \bmod 1) = K(\boldsymbol{x},\boldsymbol{y}).$$

Lemma 12. *The shift-invariant kernel $K_{s,\boldsymbol{\gamma}}^{\mathrm{sh}}$ satisfies*

$$K_{s,\boldsymbol{\gamma}}^{\mathrm{sh}}(\boldsymbol{x},\boldsymbol{y}) := \sum_{\mathfrak{u}\subseteq\{1,\ldots,s\}} \gamma_{\mathfrak{u}} \prod_{j\in\mathfrak{u}} B_2(|x_j - y_j|).$$

Proof. The statement can be proved by straightforward calculation of the integrals. □

Since the lattice-rule pointset $P = P(\boldsymbol{z})$ depends only on the generating vector $\boldsymbol{z}$, we also use the notation

$$e_n^{\mathrm{sh}}(\boldsymbol{z})^2 := e_n^{\mathrm{sh}}(P(\boldsymbol{z}),\mathcal{X}_{s,\boldsymbol{\gamma}})$$

Lemma 13. *The shift-averaged worst-case error for lattice rules in the unachored space $\mathcal{X}_{s,\gamma}$ satisfies*

$$e_n^{\mathrm{sh}}(\boldsymbol{z})^2 = \sum_{\emptyset \neq \mathfrak{u} \subseteq \{1,\ldots,s\}} \gamma_{\mathfrak{u}} \Big(\frac{1}{n} \sum_{k=0}^{n-1} \prod_{j \in \mathfrak{u}} \Big(B_2\big(\frac{k z_j}{n} \bmod 1\big)\Big)\Big)$$

Proof. Again, the statement can be proved by straightforward calculations. □

It is obviously clear that not every generating vector $\boldsymbol{z} \in \mathbb{Z}^s$ will lead to a good lattice rule. For example the vector $\boldsymbol{z} := (1,1)$ just results in points which lie on the diagonal of the unit square. A general method for constructing good generating vectors would be to optimize $e_n^{\mathrm{sh}}(\boldsymbol{z})$ over all $\boldsymbol{z} \in \mathbb{Z}^s$. This, however, is prohibitively expensive. A feasible method to achieve similar performance is the *component-by-component* algorithm:

Algorithm 1. ***Input:*** *number of points $n \in \mathbb{N}$, dimension $s \in \mathbb{N}$, and weights $\boldsymbol{\gamma}$.*

(1) *Set $z_1 = 1$.*
(2) *For $j = 2, 3, \ldots, s$, choose $z_j \in \{1, \ldots, n-1\}$ with $\gcd(z_j, n) = 1$ such that*

$$e_n^{\mathrm{sh}}(z_1, \ldots, z_j) \text{ is minimal.}$$

Output: *generating vector $\boldsymbol{z}$.*

While for general weights $\boldsymbol{\gamma}$, even the component-by-component (CBC) algorithm is to expensive to compute, for special weights, there are very efficient implementations, which we will discuss later.

First, we want to prove that Algorithm 1 actually produces a good generating vector $\boldsymbol{z}$. To that end, we require a well-known fact about Fourier series.

Lemma 14. *There holds for all $m \in \mathbb{N}$*

$$\frac{1}{n} \sum_{k=0}^{n-1} e^{2\pi i k m/n} = \begin{cases} 1 & m \bmod n = 0, \\ 0 & \text{else.} \end{cases}$$

Proof. If $m \bmod n = 0$, we have $e^{2\pi i k m/n} = 1$ and hence the statement follows. Otherwise, we have

$$\Big(\sum_{k=0}^{n-1} e^{2\pi i k m/n} \Big) e^{2\pi i m/n} = \sum_{k=0}^{n-1} e^{2\pi i (k+1) m/n} = \sum_{k=0}^{n-1} e^{2\pi i k m/n},$$

since $e^{2\pi i n m/n} = 1 = e^{2\pi i 0 m/n}$. Since $e^{2\pi i m/n} \neq 1$, it follows $\sum_{k=0}^{n-1} e^{2\pi i k m/n} = 0$. This concludes the proof. □

Before we prove Theorem 16, we require a reverse Jensen-type inequality.

Lemma 15. *For $\lambda \leq 1$ and $\alpha_k \geq 0$, there holds*

$$\Big(\sum_{k=0}^{\infty} \alpha_k \Big)^{\lambda} \leq \sum_{k=0}^{\infty} \alpha_k^{\lambda}.$$

Proof. We note that it suffices to show $(a+b)^{\lambda} \leq a^{\lambda} + b^{\lambda}$ for $a, b \geq 0$ and $\lambda \leq 1$, the rest follows by induction on the number of summands. Consider the functions $f(a) := (a+b)^{\lambda}$ and $g(a) := a^{\lambda} + b^{\lambda}$ for $a, b \geq 0$. We note that $f'(a) = \lambda (a+b)^{\lambda - 1} \leq g'(a) = \lambda a^{\lambda - 1}$ (note that $\lambda - 1 \leq 0$). Since $f(0) = g(0)$ and $f'(a) \leq g'(a)$ for all $a \geq 0$, we conclude $f(a) \leq g(a)$ for all $a \geq 0$. This concludes the proof. □

We first prove a rather abstract convergence result, which give us a hint on how to choose the weights $\boldsymbol{\gamma}$.

Theorem 16. *The component-by-component algorithm (Algorithm 1) constructs a generating vector $\boldsymbol{z}$ which satisfies*

$$e_n^{\mathrm{sh}}(\boldsymbol{z})^2 \le \Big(\frac{1}{\phi(n)} \sum_{\emptyset \ne \mathfrak{u} \subseteq \{1,\dots,s\}} \gamma_{\mathfrak{u}}^{\lambda} \Big(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}}\Big)^{|\mathfrak{u}|}\Big)^{1/\lambda}$$

for all $\lambda \in (1/2, 1]$, where $\zeta(\cdot)$ is the Riemann zeta function and $\phi(n) := \#\{1 \le k \le n : \gcd(k,n) = 1\}$ is the Euler totient function.

Remark 17. *Theorem 16 shows that, if the weights $\gamma_{\mathfrak{u}}$ converge to zero sufficiently fast, we achieve a convergence rate of $\phi(n)^{-1+\delta}$ for all $\delta > 0$ for the shift-averaged worst-case error $e_n^{\mathrm{sh}}(\boldsymbol{z})$ if we construct $\boldsymbol{z}$ with the CBC Algorithm (Algorithm 1). This means that*

$$\sqrt{\mathbb{E}|I_s(f) - Q_n(S,f)|^2} \lesssim \phi(n)^{-1+\delta} \|f\|_{\mathcal{X}_{\boldsymbol{\gamma},s}},$$

where the expectation is taken over the shifts $S \in [0,1]^s$. This error estimate is independent of the dimension $s \in \mathbb{N}$.

The Euler totient function $\phi(n)$ satisfies

$$\phi(n) > \frac{n}{e^{\mu}\log(\log(n)) + \frac{3}{\log(\log(n))}}$$

for all $n > 2$ and with Euler's constant $\mu \approx 0.577215665\ldots$. Thus, we obtain for the error estimate above

$$\sqrt{\mathbb{E}|I_s(f) - Q_n(S,f)|^2} \lesssim n^{-1+\delta} \|f\|_{\mathcal{X}_{\boldsymbol{\gamma},s}},$$

for all $\delta > 0$ with the hidden constant depending on δ. In terms of convergence rate, this is almost as good as digital nets which achieve n^{-1} up to logarithmic terms. However, the present convergence estimate is completely independent of the number of dimensions $s \in \mathbb{N}$.

Proof of Theorem 16. We prove this by induction on s. The first step $s = 1$ is straightforward to verify for all $\lambda \in (1/2, 1]$ since $\sum_{k=0}^{n-1} B_2(k/n) = 1/(6n)$ and hence

$$e_n^{\mathrm{sh}}(1)^2 \le \frac{\gamma_1}{6n^2}.$$

With $\phi(n) \le n$ and $\zeta(2\lambda) \ge \zeta(2) = \pi^2/6$, we conclude the step $s = 1$.

We assume that we have chosen the first $s-1$ components $z_1, \dots, z_{s-1}$ such that the statement of the theorem holds with s replaced by $s-1$. Next, we derive the identity

$$e_n^{\mathrm{sh}}(\boldsymbol{z})^2 = e_n^{\mathrm{sh}}(z_1,\dots,z_{s-1})^2 + \theta(\boldsymbol{z})$$

with

$$\theta(\boldsymbol{z}) := \sum_{s \in \mathfrak{u} \subseteq \{1,\dots,s\}} \gamma_{\mathfrak{u}} \Big(\frac{1}{n} \sum_{k=0}^{n-1} \prod_{j \in \mathfrak{u}} \Big(B_2\big(\frac{k z_j}{n} \bmod 1\big)\Big)\Big).$$

Note that the worst-case error in $s-1$ dimensions is always smaller or equal to the worst-case error in s dimensions (one can extend the *worst possible* integrand in $s-1$ dimensions to be constant in dimension s). Hence $\theta(\boldsymbol{z}) \ge 0$ for any choice of $\boldsymbol{z}$.

The term $kz_j/n \bmod 1$ is quite cumbersume to tackle directly, however, we notice that it is periodic in z_j. Hence, a Fourier expansion of $B_2(kz_j/n \bmod 1)$ might be helpful. The Fourier series reads $B_2(x) = 1/(2\pi^2)\sum_{h\in\mathbb{Z}\setminus\{0\}} e^{2\pi i h x}/h^2$ for all $x \in \mathbb{R}$ and we can write

$$\prod_{j\in\mathfrak{u}} \Big(B_2\big(\frac{kz_j}{n} \bmod 1\big)\Big) = \frac{1}{(2\pi^2)^{|\mathfrak{u}|}} \sum_{\boldsymbol{h}_\mathfrak{u}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|}} \frac{e^{2\pi i k \boldsymbol{h}_\mathfrak{u}\cdot\boldsymbol{z}_\mathfrak{u}/n}}{\prod_{j\in\mathfrak{u}} h_j^2},$$

where $\boldsymbol{h}_\mathfrak{u}\cdot\boldsymbol{z}_\mathfrak{u} = \sum_{j\in\mathfrak{u}} h_j z_j$. With Lemma 14, this simplifies to

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{n-1}\prod_{j\in\mathfrak{u}} \Big(B_2\big(\frac{kz_j}{n} \bmod 1\big)\Big) &= \frac{1}{(2\pi^2)^{|\mathfrak{u}|}} \sum_{\substack{\boldsymbol{h}_\mathfrak{u}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|}\\ \boldsymbol{h}_\mathfrak{u}\cdot\boldsymbol{z}_\mathfrak{u} \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}} h_j^2}\\ &= \frac{1}{(2\pi^2)^{|\mathfrak{u}|}} \sum_{h_s\in\mathbb{Z}\setminus\{0\}} h_s^{-2} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}\\ \boldsymbol{h}_\mathfrak{u}\cdot\boldsymbol{z}_\mathfrak{u} \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} h_j^2}.\end{aligned}$$

Let $z_s^\star$ denote the minimum chosen by Algorithm 1 in step s. Since the minimum is always smaller than the average, we have for all $\lambda \in (0,1]$ that

$$\theta(z_1,\dots,z_{s-1},z_s^\star)^\lambda \le \frac{1}{\phi(n)} \sum_{\substack{1\le z_s\le n-1\\ \gcd(z_s,n)=1}} \theta(z_1,\dots,z_{s-1},z_s)^\lambda.$$

Altogether, and by use of Lemma 15, we obtain

$$\begin{aligned}&\theta(z_1,\dots,z_{s-1},z_s^\star)^\lambda\\ &\quad\le \frac{1}{\phi(n)} \sum_{\substack{1\le z_s\le n-1\\ \gcd(z_s,n)=1}} \sum_{s\in\mathfrak{u}\subseteq\{1,\dots,s\}} \frac{\gamma_\mathfrak{u}^\lambda}{(2\pi^2)^{|\mathfrak{u}|\lambda}} \sum_{h_s\in\mathbb{Z}\setminus\{0\}} |h_s|^{-2\lambda} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}\\ \boldsymbol{h}_\mathfrak{u}\cdot\boldsymbol{z}_\mathfrak{u} \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}}\\ &\quad\le \frac{1}{\phi(n)} \sum_{s\in\mathfrak{u}\subseteq\{1,\dots,s\}} \frac{\gamma_\mathfrak{u}^\lambda}{(2\pi^2)^{|\mathfrak{u}|\lambda}} \underbrace{\sum_{h_s\in\mathbb{Z}\setminus\{0\}} |\widetilde{h}_s|^{-2\lambda} \sum_{\substack{1\le z_s\le n-1\\ \gcd(z_s,n)=1}} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}\\ \boldsymbol{h}_\mathfrak{u}\cdot\boldsymbol{z}_\mathfrak{u} \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}}}_{:=\mathrm{SUM}},\end{aligned}$$

where $\widetilde{h}_s := n\lfloor h_s/n\rfloor$.

$$\mathrm{SUM} = \sum_{c=0}^{n-1} \sum_{\substack{1\le z_s\le n-1\\ \gcd(z_s,n)=1}} \sum_{\substack{h_s\in\mathbb{Z}\setminus\{0\}\\ h_s z_s \bmod n=c}} |\widetilde{h}_s|^{-2\lambda} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}\\ \boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\cdot\boldsymbol{z}_{\mathfrak{u}\setminus\{s\}}+c \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}}.$$

Since $h_s z_s \bmod n = c$ implies $h_s = kn + cz_s^{-1}$ for the muliplicative inverse z_s^{-1} of z_s modulo n, we have $\widetilde{h}_s = kn$ for all c and z_s. Hence, the sum above simplifies to

$$\begin{aligned}
\mathrm{SUM} &= \sum_{k\in\mathbb{Z}\setminus\{0\}} |kn|^{-2\lambda} \sum_{c=0}^{n-1} \sum_{\substack{1\le z_s\le n-1\\ \gcd(z_s,n)=1}} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}\\ \boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\cdot\boldsymbol{z}_{\mathfrak{u}\setminus\{s\}}+c \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}} \\
&= \sum_{k\in\mathbb{Z}\setminus\{0\}} |kn|^{-2\lambda}\phi(n) \sum_{c=0}^{n-1} \sum_{\substack{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}\\ \boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\cdot\boldsymbol{z}_{\mathfrak{u}\setminus\{s\}}+c \bmod n=0}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}} \\
&= \sum_{k\in\mathbb{Z}\setminus\{0\}} |kn|^{-2\lambda}\phi(n) \sum_{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}}.
\end{aligned}$$

Since $\lambda > 1/2$, $\phi(n) \le n^{2\lambda}$ and we have

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |kn|^{-2\lambda}\phi(n) \le \frac{\phi(n)}{n^{2\lambda}} \sum_{k\in\mathbb{Z}\setminus\{0\}} |k|^{-2\lambda} \le 2\zeta(2\lambda).$$

Hence, SUM is bounded by

$$\mathrm{SUM} \le 2\zeta(2\lambda) \sum_{\boldsymbol{h}_{\mathfrak{u}\setminus\{s\}}\in(\mathbb{Z}\setminus\{0\})^{|\mathfrak{u}|-1}} \frac{1}{\prod_{j\in\mathfrak{u}\setminus\{s\}} |h_j|^{2\lambda}} = 2\zeta(2\lambda)\Big(\sum_{h\in\mathbb{Z}\setminus\{0\}} |h|^{-2\lambda}\Big)^{|\mathfrak{u}|-1} = (2\zeta(2\lambda))^{|\mathfrak{u}|}.$$

Altogether, and by use of the induction hypothesis, we prove

$$\begin{aligned}
e_n^{\mathrm{sh}}(\boldsymbol{z})^2 &= e_n^{\mathrm{sh}}(z_1,\dots,z_{s-1})^2 + \theta(\boldsymbol{z}) \\
&\le \Big(\frac{1}{\phi(n)} \sum_{\emptyset\ne\mathfrak{u}\subseteq\{1,\dots,s-1\}} \gamma_{\mathfrak{u}}^{\lambda}\Big(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}}\Big)^{|\mathfrak{u}|}\Big)^{1/\lambda} + \Big(\frac{1}{\phi(n)} \sum_{s\in\mathfrak{u}\subseteq\{1,\dots,s\}} \frac{\gamma_{\mathfrak{u}}^{\lambda}}{(2\pi^2)^{|\mathfrak{u}|\lambda}} (2\zeta(2\lambda))^{|\mathfrak{u}|}\Big)^{1/\lambda}.
\end{aligned}$$

By use of $a^x + b^x \le (a+b)^x$ for all $x \ge 1$ and $a, b \ge 0$, we conclude the proof.

□

In the following, we will require a small combinatorial identity.

Lemma 18. *For $s \in \mathbb{N}$, there holds $\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} \prod_{j\in\mathfrak{u}} a_j = \prod_{j=1}^{s}(1+a_j)$ for $a_j \in \mathbb{R}$.*

Proof. We prove the statement by induction. For $s = 1$, the statement is trivial. Assume the statement holds for $s-1$. Then, we have

$$\begin{aligned}
\sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} \prod_{j\in\mathfrak{u}} a_j &= \sum_{\mathfrak{u}\subseteq\{1,\dots,s-1\}} \prod_{j\in\mathfrak{u}} a_j + \sum_{\mathfrak{u}\subseteq\{1,\dots,s-1\}} \prod_{j\in\mathfrak{u}} a_j a_s \\
&= \prod_{j=1}^{s-1}(1+a_j) + a_s \prod_{j=1}^{s-1}(1+a_j) = \prod_{j=1}^{s}(1+a_j).
\end{aligned}$$

This concludes the proof. □

1.9. **Fast CBC construction for special weights.** As mentioned above, Algorithm 1 is still unreasonably expensive for general weights $\boldsymbol{\gamma}$. We first consider the special case of product weights, i.e.,

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

for all $\mathfrak{u} \subseteq \{1,\dots,s\}$ and some $\gamma_j > 0$. The naive way of Algorithm 1 works as follows in the case of product weights: For all $j = 2,\dots,s$, look for $z_j \in U_n := \{1 \le z \le n-1 : \gcd(z,n) = 1\}$ such that

$$\begin{aligned} e_n^{\mathrm{sh}}(z_1,\dots,z_j)^2 &= \sum_{\emptyset\neq\mathfrak{u}\subseteq\{1,\dots,j\}} \gamma_{\mathfrak{u}} \Big(\frac{1}{n} \sum_{k=0}^{n-1} \prod_{i\in\mathfrak{u}} \Big(B_2\big(\frac{kz_i}{n} \bmod 1\big)\Big)\Big) \\ &= -1 + \frac{1}{n}\sum_{k=0}^{n-1} \prod_{i=1}^{j} \Big(1 + \gamma_i \big(B_2\big(\frac{kz_i}{n} \bmod 1\big)\big)\Big) \end{aligned}$$

is minimal (we used Lemma 18 in the last equality). The straighforward implementation of this has a cost of $\mathcal{O}(s^2n^2)$ (for each j, we have to compute n-terms of the sum and a product of j-terms, moreover we have to search through $\phi(n)$ values of z_s). To simplify the computation, we may write

$$e_n^{\mathrm{sh}}(z_1,\dots,z_j)^2 = e_n^{\mathrm{sh}}(z_1,\dots,z_{j-1})^2 + \frac{\gamma_j}{n} \sum_{k=0}^{n-1} \underbrace{B_2\big(\frac{kz_j}{n} \bmod 1\big)}_{\Omega(z_j,k)} \underbrace{\prod_{i=1}^{j-1} \Big(1 + \gamma_i \big(B_2\big(\frac{kz_i}{n} \bmod 1\big)\big)\Big)}_{p_{j-1}(k)}.$$

The n values of $p_{j-1}(k)$, $k = 0,\dots,n-1$ do not depend on z_s and can be stored during the search. This reduces the cost to $\mathcal{O}(sn^2)$ at the expense of $\mathcal{O}(n)$ storage. Next, we may vectorize the implementation in the following sense: We store the values $(e_n^{\mathrm{sh}}(z_1,\dots,z_j)^2)_{z_j\in U_n}$ in the vector $\boldsymbol{e}_j \in \mathbb{R}^{\phi(n)}$. Moreover, we store the matrix $\boldsymbol{\Omega} \in \mathbb{R}^{\phi(n)\times n}$ defined by

$$\Omega_{z,k} := \Omega(z,k) = B_2\big(\frac{kz}{n} \bmod 1\big) \quad \text{for } z \in U_n,\ k = 0,\dots,n-1.$$

Finally, we need the vector $\boldsymbol{p}_{j-1} \in \mathbb{R}^n$ defined by $p_{j-1,k} := p_{j-1}(k)$. With this, we may rewrite Algorithm 1 as (in Matlab notation):
For all $j = 2,\dots,s$:

(1) Compute

$$\boldsymbol{e}_j = e_n^{\mathrm{sh}}(z_1,\dots,z_{j-1})^2 + \frac{\gamma_j}{n}\boldsymbol{\Omega}\boldsymbol{p}_{j-1}.$$

(2) Select $z_j \in U_n$ such that e_{j,z_s} is the minimal entry of $\boldsymbol{e}_j$.
(3) Set $e_n^{\mathrm{sh}}(z_1,\dots,z_j)^2 = e_{j,z_s}$.
(4) Update $\boldsymbol{p}_j = (1 + \gamma_s \boldsymbol{\Omega}_{z_s,:}) .* \boldsymbol{p}_{j-1}$.

(Note that according to Matlab notation, $\boldsymbol{\Omega}_{z_s,:}$ denotes the row with index z_s of the matrix $\boldsymbol{\Omega}$ and $.*$ is the elementwise multiplication.) This does not improve the cost estimate yet, however, the trick is to order the indices $z \in U_n$ and $k = 0,\dots,n-1$ such that the matrix-vector product $\boldsymbol{\Omega}\boldsymbol{p}_{j-1}$ can be computed quickly.

We start with the simpler case of $n \in \mathbb{N}$ being prime. Then, $\phi(n) = n-1$. Let $g \in U_n$ be the generator of the cyclic group U_n, i.e., $U_n = \{g^i \bmod n : i \in \mathbb{N}_0\}$ (this can be done in $\mathcal{O}(n\log(n))$). We reorder $\boldsymbol{\Omega}$ to $\widetilde{\boldsymbol{\Omega}}$ by

$$\widetilde{\Omega}_{i+1,j+1} := \begin{cases} \Omega_{g^i \bmod n,(g^{-1})^j \bmod n} & \text{if } 0 \le i,j \le n-2, \\ 0 & \text{if } j = n-1. \end{cases}$$

Note that $\Omega_{i,0} = 0$ for all $i \in U_n$. Now, there holds

$$\begin{aligned}\widetilde{\Omega}_{i,j+r} &= \Omega_{g^{i-1} \bmod n,(g^{-1})^{j-1+r} \bmod n} = B_2\Big(\frac{(g^{-1})^{j-1+r} g^{i-1}}{n} \bmod 1\Big) \\ &= B_2\Big(\frac{(g^{-1})^{j-1+2r} g^{i-1+r}}{n} \bmod 1\Big) = \widetilde{\Omega}_{i+r \bmod n,j+r}\end{aligned}$$

for all $0 \le r \le n-1-r$. Hence, $\widetilde{\boldsymbol{\Omega}}$ has an $(n-1) \times (n-1)$-submatrix which is circulant, i.e.,

$$\widetilde{\boldsymbol{\Omega}} = \begin{pmatrix} \boldsymbol{C} & B_2(\boldsymbol{0}) \end{pmatrix},$$

for $\boldsymbol{C} \in \mathbb{R}^{n-1 \times n-1}$ being a circulant matrix.

Lemma 19. *Let $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ denote a circulant matrix defined by $\boldsymbol{C}_{ij} := \boldsymbol{c}_{i-j \bmod n}$ for some $\boldsymbol{c} \in \mathbb{R}^n$. Then, the matrix-vector product $\boldsymbol{C}\boldsymbol{x} \in \mathbb{R}^n$ can be computed via $\boldsymbol{C}\boldsymbol{x} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{c}) \odot \mathcal{F}(\boldsymbol{x}))$, where $\mathcal{F}\colon \mathbb{C}^n \to \mathbb{C}^n$ denotes the discrete Fourier transform defined by $\mathcal{F}(\boldsymbol{x})_j := \sum_{k=0}^{n-1} \boldsymbol{x}_k e^{-2\pi i j k/n}$ and $\mathcal{F}^{-1}(\boldsymbol{x})_j := \frac{1}{n}\sum_{k=0}^{n-1} \boldsymbol{x}_j e^{2\pi i j k/n}$ and $\odot\colon \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}^n$ denote the pointwise multiplication of vectors.*

Proof. Multiplication with a circulant matrix satisfies

$$(\boldsymbol{C}\boldsymbol{x})_i = \sum_{j=0}^{n-1} \boldsymbol{C}_{ij}\boldsymbol{x}_j = \sum_{j=0}^{n-1} \boldsymbol{c}_{i-j \bmod n}\boldsymbol{x}_j$$

This is a discrete convolution, and we my write $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{c} \star \boldsymbol{x}$. The discrete Fourier transform $\mathcal{F}\colon \mathbb{C}^n \to \mathbb{C}^n$ satisfies

$$\mathcal{F}(\boldsymbol{C}\boldsymbol{x}) = \mathcal{F}(\boldsymbol{c}) \odot \mathcal{F}(\boldsymbol{x}).$$

Since $\mathcal{F}$ is isomorphic, this concludes the proof. □

Remark 20. *The matrix vector product with circulant matrices can be computed in $\mathcal{O}(n \log n)$ using Fast-Fourier-Transform (FFT). This observation reduces the cost of the CBC algorithm to $\mathcal{O}(n \log(n) s)$.*

Note that the fast implementation of the CBC algorithm depends crucially on the product structure of the weights. A similar construction is possible for so-called product-and-order dependent weights (POD weights) of the form

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

for $\gamma_j > 0$ and $\mu_r > 0$. The fast CBC algorithm for POD-weights costs $\mathcal{O}(n \log(n) s^2)$.

1.10. **Higher-order convergence.** Given $\alpha \in \mathbb{N}$, any digital $(t, m, s\alpha)$-net P can be transformed into a higher-order net, a digital (t, m, s, α)-net P_α. To do that, we use the digit interlacing function $\mathcal{D}_\alpha\colon [0,1)^\alpha \to [0,1)$

$$D_\alpha(x_1, \dots, x_\alpha) := \sum_{i=1}^{\infty} \sum_{j=1}^{\alpha} x_{j,i} 2^{-j-(i-1)\alpha},$$

where $x_j = \sum_{i=1}^{\infty} x_{j,i} 2^{-i}$ for $j = 1, \dots, \alpha$. We also define $D_\alpha\colon [0,1)^{s\alpha} \to [0,1)^s$ by $D_\alpha(\boldsymbol{x}) := (D_\alpha(x_1, \dots, x_\alpha), D_\alpha(x_{\alpha+1}, \dots, x_{2\alpha}), \dots)$. The function is called *interlacing* function since, in base 2, there holds

$$D_\alpha(x_1, \dots, x_\alpha) = 0.x_{1,1}x_{2,1} \dots x_{\alpha,1}x_{1,2}x_{2,2} \dots x_{\alpha,2} \dots$$

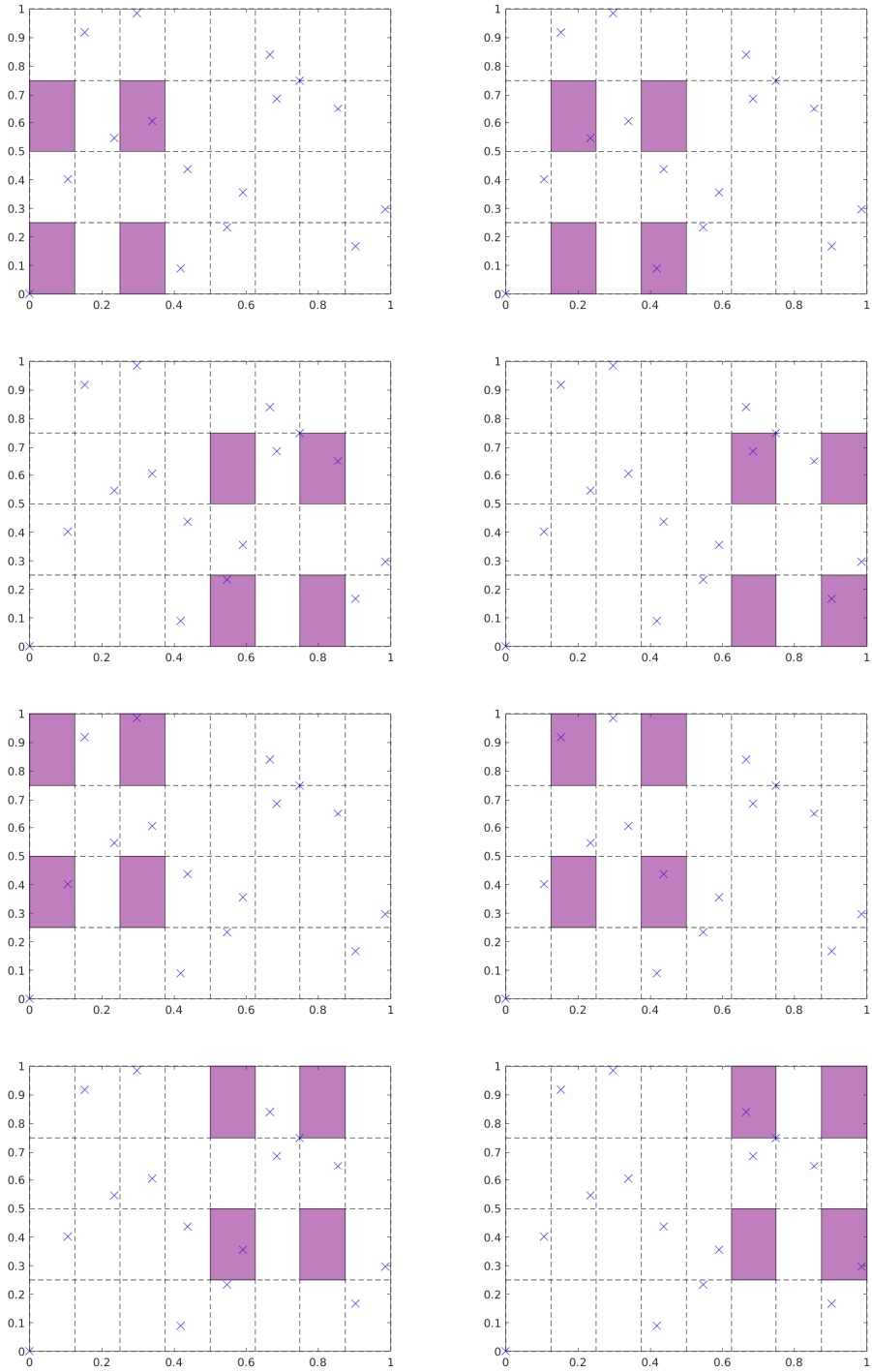

FIGURE 4. The geometric properties of higher-order digital nets. We see a digital $(1, 4, 2, 2)$-net (interlacing factor $\alpha = 2$). Each of the shaded intervals contains exactly two points of P_α. The conditions hold additionally to the classical (t, m, s)-net conditions.

The digitally interlaced net $P_\alpha := \left\{ D_\alpha(\boldsymbol{x}) \ : \ \boldsymbol{x} \in P \right\}$ satisfies additional geometric properties. An in depth look into those properties is beyond the scope of this lecture, however, Figure 4 illustrates them.

For sufficiently smooth integrand f with $\partial_{x_1}^{r_1} \dots \partial_{x_s}^{r_s}$ and $1 \le r_i \le \alpha$ bounded, higher-order digital nets can achieve a quadrature error of

$$|I_s(f) - Q(P_\alpha, f)| \lesssim m^{s\alpha} 2^{-\alpha m} \simeq \log(N)^{s\alpha}/N^\alpha.$$

Again, we have the problem that $\log(N)^{s\alpha}$ spoils the convergence for small N.

1.11. **Polynomial lattice rules.** A polynomial lattice rule is the digital net analogue of a lattice rule. Let $p \in \mathbb{Z}_2[x]$ denote a polynomial of degree m (with coefficients in $\{0,1\}$). Let $q_1, \dots, q_s \in \mathbb{Z}_2[r]$ denote polynomials of degree $m-1$. Since p/q_i is an analytic function in $\mathbb{C}$ apart from the zeros of p, we may consider the Laurent series expansion

$$\frac{q_i(x)}{p(x)} = \sum_{j=1}^{\infty} \alpha_{i,j} x^{-j}$$

for coefficients $\alpha_{i,j} \in \mathbb{Z}_2$. We define the generating matrices

$$\boldsymbol{C}_i := \begin{pmatrix} \alpha_{i,1} & \alpha_{i,2} & \alpha_{i,3} & \dots & \alpha_{i,m} \\ \alpha_{i,2} & \alpha_{i,3} & \alpha_{i,4} & \dots & \\ \alpha_{i,3} & \alpha_{i,4} & \dots & & \\ \vdots & & & & \\ \alpha_{i,m} & \dots & & & \alpha_{i,2m-1} \end{pmatrix} \in \mathbb{Z}_2^{m \times m}$$

for $i = 1, \dots, s$.

Lemma 21. *The $\boldsymbol{C}_i$ generate a (t, m, s)-net, if and only if, there holds*

$$\sum_{i=1}^{s} \beta_i(x) q_i(x) \neq 0 \bmod p$$

for all $\beta_i \in \mathbb{Z}_2[x]$ with degree $m_i - 1$ and $\sum_{i=1}^s m_i \le m - t$.

Remark 22. *The use of the negative part of the Laurent series corresponds to the* $\bmod 1$ *operation in the definition of lattice rules.*

Exactly as for lattice rules, we may consider the shifted digital net $P \oplus S$ for any shift $S \in [0,1]$. Now, we consider the digital shift

$$x \oplus y := \sum_{i=1}^{\infty} (x_i + y_i \bmod 2) 2^{-i}$$

for $x = \sum_{i=1}^\infty x_i 2^{-i}$ and $y = \sum_{i=1}^\infty y_i 2^{-i}$, $x_i, y_i \in [0,1]$. Again, we consider the digital shift averaged worst-case error

$$\sqrt{\mathbb{E}|I_s(f) - Q(P \oplus S, f)|^2} \le e_n^{\mathrm{dsh}}(P, \mathcal{X}) \|f\|_{\mathcal{X}}.$$

And, there also exists a CBC-type algorithm for polynomial lattice rules

Algorithm 2. ***Input:*** *number of points $n \in \mathbb{N}$, dimension $s \in \mathbb{N}$, and weights $\boldsymbol{\gamma}$.*

(1) *Set $q_1 = 1$.*
(2) *For $j = 2, 3, \dots, s$, choose $q_j \in \big\{q \in \mathbb{Z}_2[x] \setminus \{0\} \,:\, \deg(q) < m\big\}$*

$$e_n^{\mathrm{sh}}(q_1, \dots, q_j) \text{ is minimal.}$$

Output: *generating polynomials q_j.*

2. Poisson problem with random diffusion coefficient

We consider the Poisson problem

$$\begin{aligned} -\operatorname{div}(A(x,\boldsymbol{\omega})\nabla u(x,\boldsymbol{\omega})) &= f(x) \quad \text{for all } x \in D,\, \boldsymbol{\omega} \in \Omega \\ u(x,\boldsymbol{\omega}) &= 0 \quad \text{for all } x \in \partial D,\, \boldsymbol{\omega} \in \Omega. \end{aligned}$$

Here, $D \subset \mathbb{R}^d$ is the physical domain of the equation and Ω is the stochastic domain. This equation models the (stationary) end result of diffusion of pollutants in uncertain materials. The right-hand side f determines the amount of pollutant and the coefficient $A(x,\omega)$ the permeability of the material. The unknown material can be, e.g., soil or rock in a water pollution simulation or living tissue in a drug absorption simulation.

We assume that the random coefficient $A(x,\boldsymbol{\omega})$ is given in form of a Karhunnen-Loeve expansion, i.e.,

$$A(x,\boldsymbol{\omega}) := \phi_0(x) + \sum_{i=1}^{\infty} \phi_i(x)\omega_i$$

for all $x \in D$ and all $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots) \in \Omega := [-1/2, 1/2]^{\mathbb{N}}$ with coefficient functions $\phi_i \colon D \to \mathbb{R}$.

Remark 23. *We may introduce randomness by imagining the ω_i to be uniform random variables on $[-1/2, 1/2]$. Mathematically, one would need to separate the random variable as a function $\omega_i \colon \widetilde{\Omega} \to [-1/2, 1/2]$ on some underlying probability space $\widetilde{\Omega}$ from the parameter $\omega_i \in [-1/2, 1/2]$. However, there is no real gain in doing this here and it just complicates the notation.*

We will see later, why this assumption makes sense, for now we will accept it as a given. The functions $\phi_i \colon D \to \mathbb{R}$ are deterministic (do not depend on $\boldsymbol{\omega}$). We consider the problem in the weak form, i.e., we find $u \in H_0^1(D) := \{v \in L^2(D) \,:\, \nabla v \in L^2(D)\}$ such that

$$a_{\boldsymbol{\omega}}(u(\boldsymbol{\omega}), v) := a(A(\boldsymbol{\omega}); u(\boldsymbol{\omega}), v) := \int_D A(x,\boldsymbol{\omega})\nabla u(x,\boldsymbol{\omega}) \cdot \nabla v(x)\, dx = \int_D f(x)v(x)\, dx \tag{5}$$

for all $v \in H_0^1(D)$. To guarantee the existence of a unique solution, we need to assume

$$0 < A_{\min} \le \operatorname{real}\big(A(x,\boldsymbol{\omega})\big) \le \big|A(x,\boldsymbol{\omega})\big| \le A_{\max} < \infty$$

for all $x \in D$ and almost all $\boldsymbol{\omega} \in \Omega$ (note that $A(x,\omega) \in \mathbb{R}$ for the moment, but we will reuse this condition later with complex coefficients). Then, the Lax-Milgram lemma shows that there exists a unique weak solution $u \in H_0^1(D)$ such that

$$\|\nabla u\|_{L^2(D)} \le \frac{A_{\max}}{A_{\min}} \|f\|_{H^{-1}(D)}.$$

The goal is to compute the expectation of some quantity of interest (QOI) of the exact solution u. The quantity of interest is in our case a linear functional $G \in H_0^1(D)^\star = H^{-1}(D)$ and we want to approximate

$$\mathbb{E}(G(u)) = \int_\Omega G(u(\cdot,\boldsymbol{\omega}))\, d\boldsymbol{\omega} \in \mathbb{R},$$

where the expectation is taken with respect to $\boldsymbol{\omega} \in \Omega$. To do this, we assume the product measure $d\boldsymbol{\omega}$ on Ω as a product of uniform measures in each component justified by the following theorem.

Theorem 24. *Given a familiy of probability spaces $(\Omega_i, \Sigma_i, \mu_i)$, there exists a unique probability space (Ω, Σ, μ) such that $\Omega = \prod_{i\in\mathbb{N}} \Omega_i$, $\Sigma = \otimes_{i\in\mathbb{N}}\Sigma_i$, and*

$$\mu(\prod_{i\in\mathbb{N}} A_i) = \prod_{i\in\mathbb{N}} \mu_i(A_i)$$

for all $A_i \in \Sigma_i$ such that $A_i \neq \Omega_i$ for only finitely many $i \in \mathbb{N}$.

Applying Theorem 24 to the uniform measures on $[-1/2, 1/2]$ gives the product measure $d\boldsymbol{\omega}$ on Ω with the corresponding σ-Algebra generated by sets of the form

$$\Big\{ \prod_{i=1}^{\infty} A_i \,:\, A_i \in \mathcal{B}([-1/2,1/2]),\, A_i = [-1/2,1/2] \text{ for all but finitely many } i \in \mathbb{N} \Big\}$$

and $\mathcal{B}([-1/2, 1/2])$ denotes the Borel σ-algebra generated by the Euclidean topology on $[-1/2, 1/2]$. In general, we define the Borel σ-algebra $\mathcal{B}(S)$ as the smallest σ-algebra containing all open sets in S.

We need to check that the map $\boldsymbol{\omega} \to G(u(\boldsymbol{\omega}))$ is measurable. We do that by proving that it is even continuous.

Lemma 25. *Given $A, B\colon \Omega \to L^\infty(D)$, there holds the estimate*

$$\|a(A(\boldsymbol{\omega}),\cdot,\cdot) - a(B(\boldsymbol{\omega}),\cdot,\cdot)\| \leq \|A(\boldsymbol{\omega}) - B(\boldsymbol{\omega})\|_{L^\infty(D)} \quad \text{for all } \boldsymbol{\omega} \in \Omega.$$

as well as for $\widetilde{u} \in H^1_0(D)$ denoting the solution of (5) with B instead of A

$$\|\nabla(u - \widetilde{u})\|_{L^2(D)} \leq {A_{\min}}^{-1} \|A(\boldsymbol{\omega}) - B(\boldsymbol{\omega})\|_{L^\infty(D)} \|u\|_{H^1(D)}.$$

Proof. The first estimate follows since we have for almost all $\boldsymbol{\omega} \in \Omega$ that

$$\begin{aligned} |a(A(\boldsymbol{\omega}), u, v) - a(B(\boldsymbol{\omega}), u, v)| &\leq \int_D |A(x,\boldsymbol{\omega}) - B(x,\boldsymbol{\omega})||\nabla u||\nabla v|\, dx \\ &\leq \|A(\boldsymbol{\omega}) - B(\boldsymbol{\omega})\|_{L^\infty(D)} \|u\|_{H^1(D)} \|v\|_{H^1(D)}. \end{aligned}$$

For the second statement, we use ellipticity

$$\begin{aligned} A_{\min} \|\nabla(u - \widetilde{u})\|^2_{H^1(D)} &\leq a(A(\boldsymbol{\omega}); u - \widetilde{u}, u - \widetilde{u}) \\ &= \langle f\,,\, u - \widetilde{u} \rangle_D - a(A(\boldsymbol{\omega}); \widetilde{u}, u - \widetilde{u}) \\ &= a(B(\boldsymbol{\omega}); \widetilde{u}, u - \widetilde{u}) - a(A(\boldsymbol{\omega}); \widetilde{u}, u - \widetilde{u}) \\ &= a(A(\boldsymbol{\omega}) - B(\boldsymbol{\omega}); \widetilde{u}, u - \widetilde{u}) \\ &\leq \|A(\boldsymbol{\omega}) - B(\boldsymbol{\omega})\|_{L^\infty(D)} \|\widetilde{u}\|_{H^1(D)} \|u - \widetilde{u})\|_{H^1(D)} \end{aligned}$$

for all $\boldsymbol{\omega} \in \Omega$. This concludes the proof. □

To show that continuity implies measurability in the product σ-algebra, we need to show that the product σ-algebra is generated by the product topology. This is true for separable spaces.

Lemma 26. *The countable product of separable spaces is a separable space (with respect to the product topology).*

Proof. Let $(S_i)_{i\in\mathbb{N}}$ denote a family of separable spaces with dense countable subsets $D_i \subseteq S_i$. Fix a point $z \in S := \prod_{i\in\mathbb{N}} S_i$ and define

$$E_k := \big\{ y \in S \,:\, y_i \in D_i \text{ for all } i \leq k,\, y_i = z_i \text{ for all } i > k \big\}.$$

Clearly, the E_k are countable subsets of S. Let $x \in S$ be arbitrary and let $U \subseteq S$ be an open neighborhood of x. Since U is open in the product topology, there exists a number $k \in \mathbb{N}$ and open sets $U_i \subseteq S_i$ with $\prod_{i\leq k} U_i \times \prod_{i>k} S_i \subseteq U$. Since D_i is dense in S_i, we

find $y_i \in D_i$ with $y_i \in U_i$. The point $y = (y_1, \dots, y_k, z_{k+1}, z_{k+2}, \dots)$ is in E_k and $y \in U$. This shows that $\bigcup_{k\in\mathbb{N}} E_k$ is dense in S and concludes the proof. □

Lemma 27. *Let S_i, $i \in \mathbb{N}$ denote separable metric spaces and let $S := \prod_{i=1}^{\infty} S_i$ be endowed with the product topology. Then, there holds $\mathcal{B}(S) = \prod_{i=1}^{N} \mathcal{B}(S_i)$, where the latter denotes the product σ-algebra.*

Proof. First, $\prod_{i=1}^{N} \mathcal{B}(S_i)$ is generated by sets of the form $\prod_{i\neq k} S_i \times U_k$ for some open $U_k \subseteq S_i$. Those sets are part of the topology on S and hence $\mathcal{B}(S) \supseteq \prod_{i=1}^{N} \mathcal{B}(S_i)$. On the other hand, since all S_i are separable metric spaces, also S is a separable metric space and for any topological base, the open sets in S are countable unions of that base. Again, we may choose as a base the sets of the form $\prod_{i=1}^{\infty} U_i$ with $U_i = S_i$ for all but finitely many $i \in \mathbb{N}$. Those sets contained in $\prod_{i=1}^{N} \mathcal{B}(S_i)$ and hence the topology of S is contained in $\prod_{i=1}^{N} \mathcal{B}(S_i)$. Since $\mathcal{B}(S)$ is the smallest σ-algebra to contain the topology, we prove $\mathcal{B}(S) = \prod_{i=1}^{N} \mathcal{B}(S_i)$. □

Lemma 28. *Assume that $\sum_{i=1}^{\infty} \|\phi_i\|_{L^\infty(D)} < \infty$. Then, $\boldsymbol{\omega} \mapsto G(u(\boldsymbol{\omega}))$ is measurable with respect to the product measure $d\boldsymbol{\omega}$ on $[-1/2, 1/2]^{\mathbb{N}}$. Particularly, $\mathbb{E}G(u)$ is well-defined.*

Proof. Given $\varepsilon > 0$, Lemma 25 implies

$$|G(u(\boldsymbol{\omega})) - G(u(\boldsymbol{\omega}'))| \lesssim \sum_{i=1}^{\infty} \|\phi_i\|_{L^\infty(D)} |\omega_i - \omega_i'| \leq \sum_{i=1}^{n} \|\phi_i\|_{L^\infty(D)} |\omega_i - \omega_i'| + \sum_{i=n+1}^{\infty} \|\phi_i\|_{L^\infty(D)}.$$

We may choose $n \in \mathbb{N}$ sufficiently large, such that the second sum satisfies $\sum_{i=n+1}^{\infty} \|\phi_i\|_{L^\infty(D)} < \varepsilon/2$. Hence, we find a neighborhood $U = \prod_{i=1}^{n} U_i \times \prod_{i=n+1}^{\infty} [-1/2, 1/2] \subseteq [-1/2, 1/2]^{\mathbb{N}}$ in the product topology with $\boldsymbol{\omega}, \boldsymbol{\omega}' \in U$ implies $|G(u(\boldsymbol{\omega})) - G(u(\boldsymbol{\omega}'))| < \varepsilon$. Hence $\boldsymbol{\omega} \mapsto G(u(\boldsymbol{\omega}))$ is continuous in product topology and therefore measurable with respect to the product measure. □

It turns out that the function $\boldsymbol{\omega} \mapsto G(u(\boldsymbol{\omega}))$ is even holomorphic in each argument in a domain of the complex plane.

Lemma 29. *If $\sum_{j=0}^{\infty} \|\phi_j\|_{L^\infty(D)} < \infty$, the map $\boldsymbol{\omega} \mapsto G(u(A(\boldsymbol{\omega})))$ can be extended to a function which is holomorphic in each argument in the domain $\Omega' := \big\{\boldsymbol{\omega} \in \mathbb{C}^{\mathbb{N}} : A_{\min} < \operatorname{real} A(x, \boldsymbol{\omega}) < A_{\max}$ for all $x \in D\big\} \subseteq \mathbb{C}^{\mathbb{N}}$.*

Proof. We verify complex differentiability of the parametric solutions. Fix $j \in \mathbb{N}$. Given $z \in \mathbb{C}$, define $\boldsymbol{\omega} + z \in \mathbb{C}^{\mathbb{N}}$ by $(\boldsymbol{\omega} + z)_i = \omega_i$ for all $i \neq j$ and $(\boldsymbol{\omega} + z)_j = \omega_j + z$. Let z be sufficiently small such that there exists $\varepsilon \geq |z|$ with $B_\varepsilon(\boldsymbol{\omega}) \subseteq \Omega'$. Let $g(\boldsymbol{\omega}) \in H_0^1(D)$ denote the representer of $G(\cdot)$ with respect to $a_{\boldsymbol{\omega}}$, i.e., $G(\cdot) = a_{\boldsymbol{\omega}}(g(\boldsymbol{\omega}), \cdot)$. This and the above allows us to compute

$$\begin{aligned}
\frac{G(u(\boldsymbol{\omega}+z)) - G(u(\boldsymbol{\omega}))}{z} &= \frac{a_{\boldsymbol{\omega}}(u(\boldsymbol{\omega}+z) - u(\boldsymbol{\omega}), g(\boldsymbol{\omega}))}{z} \\
&= \frac{a_{\boldsymbol{\omega}}(u(\boldsymbol{\omega}+z), g(\boldsymbol{\omega})) - \int_D f g(\boldsymbol{\omega})\,dx}{z} \\
&= \frac{a_{\boldsymbol{\omega}}(u(\boldsymbol{\omega}+z), g(\boldsymbol{\omega})) - a_{\boldsymbol{\omega}+z}(u(\boldsymbol{\omega}+z), g(\boldsymbol{\omega}))}{z} \\
&= -\int_D \frac{A(x, \boldsymbol{\omega}+z) - A(x, \boldsymbol{\omega})}{z} \nabla u(\boldsymbol{\omega}+z) \cdot \nabla g(\boldsymbol{\omega})\,dx.
\end{aligned} \tag{6}$$

Since A is affine in each component of $\boldsymbol{\omega}$, we obtain

$$\frac{A(x,\boldsymbol{\omega}+z)-A(x,\boldsymbol{\omega})}{z}=\phi_j(x).$$

Lemma 25 and $\|A(\boldsymbol{\omega}+z)-A(\boldsymbol{\omega})\|_{L^\infty(D)}\le\|\phi_j\|_{L^\infty(D)}|z|$ shows $\lim_{z\to0}u(\boldsymbol{\omega}+z)=u(\boldsymbol{\omega})$ in $H^1(D)$ and hence

$$\lim_{z\to0}\frac{G(u(\boldsymbol{\omega}+z))-G(u(\boldsymbol{\omega}))}{z}=-\int_D\phi_j(x)\nabla u(\boldsymbol{\omega})\cdot\nabla g(\boldsymbol{\omega})\,dx.$$

This shows the existence of the complex derivative and hence concludes the proof. □

2.1. **FEM approximation.** To approximate the solution $u(\boldsymbol{\omega})$ of the deterministic problem, we use a standard finite element method. For a given triangulation $\mathcal{T}_h$ of D, we construct that space $V_h\subset H^1_0(D)$ of elementwise affine functions, i.e.,

$$V_h:=\big\{v\in H^1_0(D)\,:\,v|_T(x)=a_T\cdot x+b_T\text{ for some }a_T\in\mathbb{R}^d,b_T\in\mathbb{R},\,T\in\mathcal{T}_h\big\}.$$

The index $h>0$ denotes the maximal mesh-size, i.e. $\operatorname{diam}(T)\le h$ for all $T\in\mathcal{T}_h$. The FEM approximation $u_h^s(\boldsymbol{\omega})\in\mathcal{X}$ is defined by:

$$a^s_{\boldsymbol{\omega}}(u_h^s(\boldsymbol{\omega}),v)=\int_D fv\,dx\quad\text{ for all }v\in V_h,$$

where

$$a^s_{\boldsymbol{\omega}}(u,v):=\int_D\big(\phi_0(x)+\sum_{j=1}^s\phi_j(x)\omega_j\big)\nabla u(x)\cdot\nabla v(x)\,dx$$

is the truncated version of $a_{\boldsymbol{\omega}}(\cdot,\cdot)$. This allows us to approximate

$$\mathbb{E}(G(u))\approx Q_n(G(u_h^s)),$$

where $Q_n(\cdot)$ is a quadrature formula on $[-1/2,1/2]^s$.

Lemma 30. *Assume that $\sum_{j=1}^\infty\|\phi_j\|_{L^\infty(D)}<2\inf_{x\in D}\phi_0(x)$ and define $u^s\in H^1_0(D)$ as the unique solution of*

$$a^s_{\boldsymbol{\omega}}(u^s,v)=\int_D fv\,dx\quad\textit{for all }v\in H^1_0(D).$$

Then, also $u_h^s(\boldsymbol{\omega})$ is well-defined for all $\boldsymbol{\omega}\in\Omega$ and satisfies

$$\|u(\boldsymbol{\omega})-u^s(\boldsymbol{\omega})\|_{H^1(D)}\le C\|u(\boldsymbol{\omega})\|_{H^1(D)}\sum_{j=s+1}^\infty\|\phi_j\|_{L^\infty(D)},$$
$$\|u^s(\boldsymbol{\omega})-u_h^s(\boldsymbol{\omega})\|_{H^1(D)}\le C\inf_{v\in V_h}\|u^s(\boldsymbol{\omega})-v\|_{H^1(D)},$$

for all $C>0$ with $\sum_{j=1}^\infty\|\phi_j\|_{L^\infty(D)}+C^{-1/2}<2\inf_{x\in D}\phi_0(x)$.

Proof. The assumption on the ϕ_j implies that there exists $\delta>0$ such that

$$0<\delta\le\phi_0(x)+\sum_{j=1}^\infty\phi_j(x)\omega_j\le\delta^{-1}$$

for all $x \in D$ and all $\boldsymbol{\omega} \in [-1/2, 1/2]^{\mathbb{N}}$. Hence, the Lax-Milgram lemma guarantees the existence of the unique solutions $u^s \in H_0^1(D)$ and $u_h^s \in V_h$. We obtain (Céa lemma) that

$$\begin{aligned}\|\nabla(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}))\|_{L^2(D)}^2 &\le \delta^{-1} a_{\boldsymbol{\omega}}^s(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}), u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega})) \\ &\le \delta^{-1} a_{\boldsymbol{\omega}}^s(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}), u^s(\boldsymbol{\omega}) - v_h) \\ &\le \delta^{-2} \|\nabla(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}))\|_{L^2(D)} \|\nabla(u^s(\boldsymbol{\omega}) - v_h)\|_{L^2(D)}\end{aligned}$$

for all $v_h \in V_h$ (since $a_{\boldsymbol{\omega}}^s(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}), u_h^s(\boldsymbol{\omega})) = 0 = a_{\boldsymbol{\omega}}^s(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}), u^s - v_h)$ by definition). This implies

$$\|\nabla(u^s(\boldsymbol{\omega}) - u_h^s(\boldsymbol{\omega}))\|_{L^2(D)} \le \delta^{-2} \|\nabla(u^s(\boldsymbol{\omega}) - v_h)\|_{L^2(D)}.$$

Lemma 25 shows

$$\|\nabla(u(\boldsymbol{\omega}) - u^s(\boldsymbol{\omega}))\|_{L^2(D)} \lesssim \delta^{-1} \| \sum_{j=s+1}^{\infty} \phi_j \|_{L^\infty(D)} \|u(\boldsymbol{\omega})\|_{H^1(D)}.$$

Altogether, we conclude the proof with the Friedrich's inequality

$$\|v\|_{H^1(D)} \lesssim \|\nabla(v)\|_{L^2(D)} \quad \text{for all } v \in H_0^1(D).$$

□

The last result shows that in order to estimate the approximation error, it suffices to estimate

$$\begin{aligned}&\sqrt{\mathbb{E}_S|\mathbb{E}(G(u)) - Q_n(S, G(u_h^s))|^2} \\ &\quad\le |\mathbb{E}(G(u - u^s))| + \sqrt{\mathbb{E}_S|\mathbb{E}(G(u^s)) - Q_n(S, G(u^s))|^2} \\ &\qquad + \sqrt{\mathbb{E}_S|Q_n(S, G(u^s - u_h^s))|^2} \\ &\quad\le \sqrt{\mathbb{E}_S|I_s(G(u^s)) - Q_n(S, G(u^s))|^2} + \|G(u^s - u_h^s)\|_{L^\infty(\Omega)} \\ &\qquad + C\|G\| \sum_{j=s+1}^{\infty} \|\phi_j\|_{L^\infty(D)}.\end{aligned} \tag{7}$$

Here, $\mathbb{E}_S$ denotes the expectation over the random shifts S of the lattice rule from the previous section and $\mathbb{E}$ is the expectation over the random parameter space Ω. While $|I_s(G(u^s)) - Q_n(G(u^s))|$ is a high-dimensional integration error, the remaining error contributions depend only on the triangulation $\mathcal{T}_h$ and the decay of the coefficients ϕ_j.

2.2. **The quadrature error.** The goal is to control the integration error in (7). To that end, we employ randomly shifted lattice rules from the previous section. In order to control the error, we need to ensure that the integrand is in a weighted space $\mathcal{X}_{s,\boldsymbol{\gamma}}$. In the following lemma, we use the notion of open balls $B_r(x) \subset \mathbb{C}$ with radius $r > 0$ and center $x \in \mathbb{C}$. We show that holomorphy directly implies that the derivatives are bounded.

Lemma 31. *Let $(\varrho_j)_{j\in\mathbb{N}}$ be a positive sequence such that*

$$\Omega \subset \prod_{j\in\mathbb{N}} \overline{B_{1/2+\varrho_j}(0)} \subseteq \Omega'$$

and that $0 < A_{\min} < \operatorname{real} A(x, \boldsymbol{\omega}) < A_{\max}$ for all $x \in D$ and all $\boldsymbol{\omega} \in \Omega'$. Then, we have

$$\|\partial_{\boldsymbol{\omega}_{\mathfrak{u}}} G(u^s)\|_{L^\infty(\Omega)} \le C\|f\|_{H^{-1}(D)}\|G\| \prod_{i\in\mathfrak{u}} \varrho_i^{-1}$$

for all $\mathfrak{u} \subseteq \{1, \dots, s\}$ and some constant $C > 0$ that does not depend on s.

Proof. We define $F(\boldsymbol{\omega}) := G(u^s(\boldsymbol{\omega}))$. Lemma 29 shows that F can be extended to a function $F\colon \Omega' \to \mathbb{C}$, which is holomorphic in each coordinate ω_j. Moreover, Lemma 25 proves that F is uniformly continuous in Ω'. Therefore, we obtain immediately by induction that F satisfies the multidimensional analog of Cauchy's integral formula for all $\boldsymbol{\omega} \in \Omega'$: Choose n-distinct coordinates $\mathfrak{u} := \{d_1, \dots, d_n\} \subseteq \{1, \dots, s\}$, $\boldsymbol{z} \in \mathbb{C}^s$, and define $(\boldsymbol{z}; \boldsymbol{\omega}; \mathfrak{u}) \in \mathbb{R}^{\mathbb{N}}$ via

$$(\boldsymbol{z}; \boldsymbol{\omega}; \mathfrak{u})_i = \begin{cases} z_i & i \in \mathfrak{u} \\ \omega_i & i \notin \mathfrak{u}. \end{cases}$$

Given $\boldsymbol{\omega} \in \Omega'$, choose $\varepsilon_i > 0$ for all $i \in \mathfrak{u}$ such that any $\boldsymbol{\omega}' \in \mathbb{C}^{\mathbb{N}}$ with $|\boldsymbol{\omega}_i - \boldsymbol{\omega}'_i| \leq \varepsilon_i$ for $i \in \mathfrak{u}$ and $\boldsymbol{\omega}_i = \boldsymbol{\omega}'_i$ for $i \notin \mathfrak{u}$ satisfies $\boldsymbol{\omega}' \in \Omega'$. Then, there holds (note that uniform continuity allows us to interchange the order of integration)

$$\begin{aligned} F(\boldsymbol{\omega}) &= (2\pi i)^{-1} \int_{\partial B_{\varepsilon_1}(\omega_{d_1})} \frac{F(\boldsymbol{z}; \boldsymbol{\omega}; \{d_1\})}{(z_1 - \omega_{d_1})}\, dz_1 \\ &= (2\pi i)^{-2} \int_{\partial B_{\varepsilon_1}(\omega_{d_1})} \int_{\partial B_{\varepsilon_2}(\omega_{d_2})} \frac{F(\boldsymbol{z}; \boldsymbol{\omega}; \{d_1, d_2\})}{(z_1 - \omega_{d_1})(z_2 - \omega_{d_2})}\, dz_1\, dz_2 \\ &= \dots \\ &= (2\pi i)^{-n} \int_{\partial B_{\varepsilon_1}(\omega_{d_1})} \cdots \int_{\partial B_{\varepsilon_n}(\omega_{d_n})} \frac{F(\boldsymbol{z}; \boldsymbol{\omega}; \mathfrak{u})}{(z_1 - \omega_{d_1}) \dots (z_n - \omega_{d_n})}\, dz_1 \dots dz_n. \end{aligned} \tag{8}$$

Choosing $\varepsilon_j = \varrho_j$ and $\boldsymbol{\omega} \in \Omega$ we ensure that the contour integrals are contained in Ω'. Thus, differentiation with respect to $\boldsymbol{\omega}_{\mathfrak{u}}$ shows

$$\partial_{\boldsymbol{\omega}_{\mathfrak{u}}} F(\boldsymbol{\omega}) = (2\pi i)^{-|\mathfrak{u}|} \int_{\prod_{i \in \mathfrak{u}} \partial B_{\varrho_i}(\omega_i)} \frac{F(\boldsymbol{z}; \boldsymbol{\omega}; \mathfrak{u})}{\prod_{i \in \mathfrak{u}} (z_i - \omega_i)^2}\, dz$$

and hence

$$|\partial_{\boldsymbol{\omega}_{\mathfrak{u}}} F(\boldsymbol{\omega})| \leq (2\pi)^{-|\mathfrak{u}|} \Big(\prod_{i \in \mathfrak{u}} \frac{2\pi \varrho_i}{\varrho_i^2} \Big) \|F\|_{L^\infty(\Omega')} \leq \Big(\prod_{i=1}^{\infty} \varrho_i^{-1} \Big) \|F\|_{L^\infty(\Omega')}.$$

The norm of F can be bounded by $\|F\|_{L^\infty(\Omega')} \leq \|G\| \sup_{\boldsymbol{\omega} \in \Omega'} \|u^s(\boldsymbol{\omega})\|_{H^1(D)}$. This concludes the proof. □

Lemma 32. *Assume that $\sum_{j=1}^{\infty} \|\phi_j\|_{L^\infty(D)} + 2\delta \leq 2 \inf_{x \in D} \phi_0(x)$ for some $\delta > 0$ as well as $\sum_{j=1}^{\infty} \|\phi_j\|_{L^\infty(D)}^{1/2} < \infty$. Define for sufficiently small $\delta > 0$*

$$\beta_i := 2/\delta \|\phi_i\|_{L^\infty(D)}^{1/2} \sum_{j=1}^{\infty} \|\phi_j\|_{L^\infty(D)}^{1/2}.$$

Then, all $\mathfrak{u} \subseteq \{1, \dots, s\}$ satisfy

$$\|\partial_{\boldsymbol{\omega}_{\mathfrak{u}}} F\|_{L^\infty(\Omega)} \leq C \Big(\prod_{i \in \mathfrak{u}} \beta_i \Big) \|f\|_{H^{-1}(D)} \|G\|$$

The constant $C > 0$ is independent of s.

Proof. Let $r = \sum_{j=1}^{\infty} \|\phi_j\|_{L^\infty(D)}^{1/2} < \infty$. Given $\mathfrak{u} \subseteq \{1, \dots, s\}$ and $\varepsilon > 0$, an admissible sequence $(\varrho_j)_{j \in \mathbb{N}}$ in Lemma 31 is

$$\varrho_j := \beta_j^{-1} = \delta / 2 r^{-1} \|\phi_j\|_{L^\infty(D)}^{-1/2}.$$

Since $|\omega_j| \le 1/2 + \varrho_j$, this sequence satisfies

$$\inf_{\boldsymbol{\omega}\in\Omega'} \operatorname{real}\big(\phi_0 + \sum_{i=1}^{\nu} \omega_i\phi_i\big) \ge \phi_0 - 1/2 \sum_{j=1}^{\infty} \|\phi_j\|_{L^\infty(D)} - \delta/2r^{-1} \sum_{i=1}^{\infty} \|\phi_j\|_{L^\infty(D)}^{1/2} \ge \delta/2.$$

Analogously, we obtain

$$\sup_{\boldsymbol{\omega}\in\Omega'} \big|\|\phi_0\|_{L^\infty(D)} + \sum_{i=1}^{\nu} |\omega_i| \|\phi_i\|_{L^\infty(D)}\big| \lesssim 1 + r^{-1} \sum_{i=1}^{\infty} \|\phi_j\|_{L^\infty(D)}^{1/2} < \infty.$$

Hence, Lemma 31 applies and shows

$$\|\partial_{\boldsymbol{\omega}_\mathfrak{u}} G(u^s)\|_{L^\infty(\Omega)} \le C\|f\|_{H^{-1}(D)}\|G\| \prod_{i\in\mathfrak{u}} \varrho_i^{-1} = C\|f\|_{H^{-1}(D)}\|G\| \prod_{i\in\mathfrak{u}} \beta_i.$$

This concludes the proof. □

Now, we have all the necessary ingredients to apply the QMC theory to the problem at hand. The last theorem suggests to use product weights, i.e.,

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

This leads to the following theorem.

Theorem 33. *Let $\sum_{j=1}^{\infty} \|\phi_j\|_{L^\infty(D)}^{\lambda/2} < \infty$. Then, a randomly shifted lattice rule with generating vector constructed by the CBC algorithm (Algorithm 1) with weights $\gamma_\mathfrak{u}$ applied to the Poisson problem with random coefficients satisfies*

$$\sqrt{\mathbb{E}_S |I_s(G(u^s)) - Q_n(S, G(u^s))|^2} \le C\phi(n)^{-1/(2\lambda)}$$

for all $n \in \mathbb{N}$. The constant $C > 0$ is independent of the dimension $s \in \mathbb{N}$.

Proof. We already know that the integration error is bounded by

$$\sqrt{\mathbb{E}_S |I_s(G(u^s)) - Q_n(S, G(u^s))|^2} \le e_n^{\mathrm{sh}}(\boldsymbol{z}) \|G(u^s)\|_{\mathcal{X}_{s,\boldsymbol{\gamma}}}.$$

Theorem 16 shows that the shift-averaged worst-case error is then bounded by

$$\begin{aligned} e_n^{\mathrm{sh}}(\boldsymbol{z})^2 &\le \Big(\frac{1}{\phi(n)} \sum_{\emptyset\ne\mathfrak{u}\subseteq\{1,\dots,s\}} \prod_{j\in\mathfrak{u}} \beta_j^\lambda \Big(\frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}\Big)^{|\mathfrak{u}|}\Big)^{1/\lambda} \\ &= \frac{1}{\phi(n)^{1/\lambda}} \Big(\sum_{\emptyset\ne\mathfrak{u}\subseteq\{1,\dots,s\}} \prod_{j\in\mathfrak{u}} (C_\lambda \beta_j^\lambda)\Big)^{1/\lambda} = \frac{1}{\phi(n)^{1/\lambda}} \Big(-1 + \prod_{j=1}^{s} (1 + C_\lambda \beta_j^\lambda)\Big)^{1/\lambda} \end{aligned}$$

for $C_\lambda := \frac{2\zeta(2\lambda)}{(2\pi^2)^\lambda}$ (here we used Lemma 18 for the final equality). The product can be estimated by

$$\log\Big(\prod_{j=1}^{s} (1 + C_\lambda \beta_j^\lambda)\Big) = \sum_{j=1}^{s} \log(1 + C_\lambda \beta_j^\lambda) \le C_\lambda \sum_{j=1}^{s} \beta_j^\lambda \le C_\lambda \sum_{j=1}^{\infty} \beta_j^\lambda < \infty$$

by assumption on the $\|\phi_j\|_{L^\infty(D)}$. Hence, we obtain $e_n^{\mathrm{sh}}(\boldsymbol{z}) \le \widetilde{C}_\lambda \phi(n)^{-1/(2\lambda)}$, where $\widetilde{C}_\lambda$ depends only on λ. It remains to estimate the norm $\|G(u)\|_{\mathcal{X}_{s,\boldsymbol{\gamma}}}$. To that end, Lemma 32

shows

$$\|G(u^s)\|^2_{\mathcal{X}_{s,\gamma}} = \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} \gamma_{\mathfrak{u}}^{-1} \int_{[-1/2,1/2]^{|\mathfrak{u}|}} \Big(\int_{[-1/2,1/2]^{s-|\mathfrak{u}|}} \partial_{\boldsymbol{\omega}_{\mathfrak{u}}} G(u^s(\boldsymbol{\omega}))\, d\boldsymbol{\omega}_{\mathfrak{u}^c} \Big)^2 d\boldsymbol{\omega}_{\mathfrak{u}}$$

$$\lesssim \|u^s\|^2_{H^1(D)} \|G\|^2 \sum_{\mathfrak{u}\subseteq\{1,\dots,s\}} \gamma_{\mathfrak{u}} = \|u^s\|^2_{H^1(D)} \|G\|^2 \prod_{j=1}^{s} (1+\beta_j).$$

Analogously as before, we show that the product is bounded independently of s. This concludes the proof. □

Remark 34. *Note that the result of Theorem 33 is not optimal. A more careful analysis shows that the condition $\sum_{j=1}^\infty \|\phi_j\|^{2/3}_{L^\infty(D)} < \infty$ would suffice (instead of $\sum_{j=1}^\infty \|\phi_j\|^{1/2}_{L^\infty(D)} < \infty$). In this case, however, one requires POD-weights of the form*

$$\gamma_{\mathfrak{u}} = |\mathfrak{u}|! \prod_{j\in\mathfrak{u}} \widetilde{\gamma}_j$$

to bound the integration error. This makes the CBC algorithm more expensive, it costs $\mathcal{O}(s^2 n)$ to find a good generating vector (although there is a recent work which speeds up the CBC algorithm also for POD-weights in certain cases, see `https://arxiv.org/abs/1902.11068` *or* [5]*)*

2.3. **The remaining error contributions.** The error estimate in (7) together with Theorem 33 shows

$$\sqrt{\mathbb{E}_S |\mathbb{E}(G(u)) - Q_n(S, G(u_h^s))|^2} \lesssim n^{-1/(2\lambda)} + |G(u^s - u_h^s)| + \sum_{j=s+1}^\infty \|\phi_j\|_{L^\infty(D)}.$$

Approximation of the random coefficient: Theorem 33 requires convergence of $\sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)}$ to get the above result. Assuming that $\|\phi_j\|_{L^\infty(D)}$ is decreasing for $j\to\infty$, we obtain

$$s\|\phi_s\|^{\lambda/2}_{L^\infty(D)} \le \sum_{j=1}^s \|\phi_j\|^{\lambda/2}_{L^\infty(D)} \le \sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)}$$

and hence

$$\|\phi_s\|_{L^\infty(D)} \le s^{-2/\lambda} \Big(\sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)} \Big)^{2/\lambda}$$

as well as

$$\sum_{j=s+1}^\infty \|\phi_j\|_{L^\infty(D)} \le \|\phi_s\|^{1-\lambda/2}_{L^\infty(D)} \sum_{j=s+1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)} \le \|\phi_s\|^{1-\lambda/2}_{L^\infty(D)} \sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)}.$$

The combination of the two estimates shows

$$\begin{aligned} \sum_{j=s+1}^\infty \|\phi_j\|_{L^\infty(D)} &\le s^{-2(1-\lambda/2)/\lambda} \Big(\sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)} \Big)^{2(1-\lambda/2)/\lambda} \sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)} \\ &\le s^{-(2/\lambda-1)} \Big(\sum_{j=1}^\infty \|\phi_j\|^{\lambda/2}_{L^\infty(D)} \Big)^{2/\lambda} \end{aligned} \tag{9}$$

Approximation by FEM: Straightforward estimation of $|G(u^s - u_h^s)|$ leads to a convergence rate of h. However, we can use the so-called Aubin-Nitsche trick to obtain a better rate of h^2.

Lemma 35. *Assume $G(v) = \int_D g(x)v(x)\,dx$ for some $g \in L^2(D)$. Let $\widehat{g}(\boldsymbol{\omega}) \in H^1_0(D)$ denote the unique solution of $a^s_{\boldsymbol{\omega}}(\widehat{g}, v) = \int_D gv\,dx$. Then, there holds*

$$|G(u^s(\boldsymbol{\omega}) - u^s_h(\boldsymbol{\omega}))| \le C \inf_{v_h\in V_h} \|u(\boldsymbol{\omega}) - v_h\|_{H^1(D)} \inf_{g_h\in V_h} \|\widehat{g}(\boldsymbol{\omega}) - g_h\|_{H^1(D)},$$

where the constant $C > 0$ is independent of h and $\boldsymbol{\omega}$.

Proof. There holds

$$G(u^s - u^s_h) = \int_D g(u^s - u^s_h)\,dx = a^s_{\boldsymbol{\omega}}(\widehat{g}, u^s - u^s_h).$$

Since $a^s_{\boldsymbol{\omega}}(g_h, u^s - u^s_h) = a^s_{\boldsymbol{\omega}}(u^s - u^s_h, g_h) = 0$ for all $g_h \in V_h$, we have

$$|G(u^s - u^s_h)| = \inf_{g_h\in V_h} |a^s_{\boldsymbol{\omega}}(u^s - u^s_h, \widehat{g} - g_h)| \lesssim \|u^s - u^s_h\|_{H^1(D)} \inf_{g_h\in V_h} \|\widehat{g} - g_h\|_{H^1(D)}.$$

Lemma 30 improves the above estimate to

$$|G(u^s(\boldsymbol{\omega}) - u^s_h(\boldsymbol{\omega}))| \lesssim \inf_{v_h\in V_h} \|u(\boldsymbol{\omega}) - v_h\|_{H^1(D)} \inf_{g_h\in V_h} \|\widehat{g}(\boldsymbol{\omega}) - g_h\|_{H^1(D)}$$

and concludes the proof. □

To estimate $\inf_{v\in V_h} \|u(\boldsymbol{\omega}) - v\|_{H^1(D)}$, we require some FEM theory.

Lemma 36. *If ∂D is convex, or sufficiently smooth, $f \in L^2(D)$ and $\|\phi_0\|_{W^{1,\infty}(D)} + \sum_{j=1}^\infty \|\phi_j\|_{W^{1,\infty}(D)} < \infty$, there holds*

$$\inf_{v\in V_h} \|u(\boldsymbol{\omega}) - v\|_{H^1(D)} \le Ch\|f\|_{L^2(D)}.$$

Proof. The assumptions show that $\|A(\boldsymbol{\omega}, \cdot)\|_{W^{1,\infty}(D)} < \infty$ uniformly for all $\boldsymbol{\omega} \in \Omega$. Hence, we may rewrite

$$-\mathrm{div}(A(\boldsymbol{\omega}, x)\nabla u(\boldsymbol{\omega}, x)) = -\nabla A(\boldsymbol{\omega}, x) \cdot \nabla u(x) - A(\boldsymbol{\omega}, x)\Delta u(\boldsymbol{\omega}, x) = f(x)$$

which leads to

$$-\Delta u(\boldsymbol{\omega}, x) = \mathrm{rhs}(\boldsymbol{\omega}, x) := \frac{1}{A(\boldsymbol{\omega}, x)}\Big(f(x) + \nabla A(\boldsymbol{\omega}, x) \cdot \nabla u(x)\Big).$$

Since $u \in H^1(D)$, we have $\nabla A(\boldsymbol{\omega}, x) \cdot \nabla u(x) \in L^2(D)$ and hence the right-hand side above is in $L^2(D)$. Thus, elliptic regularity theory (see, e.g., [2, Section 9.6]) shows

$$\|u\|_{H^2(D)} \lesssim \|\mathrm{rhs}(\boldsymbol{\omega}, x)\|_{L^2(D)} \lesssim \|f\|_{L^2(D)} + \|u\|_{H^1(D)},$$

where the hidden constants depend only on $\|\phi_0\|_{W^{1,\infty}(D)} + \sum_{j=1}^\infty \|\phi_j\|_{W^{1,\infty}(D)} < \infty$ and $A_{\min} > 0$. The Lax-Milgram lemma shows $\|u\|_{H^1(D)} \lesssim \|f\|_{L^2(D)}$ and hence concludes $\|u(\boldsymbol{\omega})\|_{H^2(D)} \lesssim \|f\|_{L^2(D)}$ with $\boldsymbol{\omega}$-independent hidden constant.

The theory of quasi-interpolation operators (see FEM lecture notes or [9]) proves for a quasi-interpolation operator $I_h\colon H^1(D) \to V_h$ that

$$\inf_{v\in V_h} \|u(\boldsymbol{\omega}) - v\|_{H^1(D)} \le \|(1 - I_h)u(\boldsymbol{\omega})\|_{H^1(D)} \lesssim h\|u(\boldsymbol{\omega})\|_{H^2(D)} \lesssim h\|f\|_{L^2(D)}.$$

This concludes the proof. □

Altogether, we have an approximation error bounded by

$$\sqrt{\mathbb{E}_S|\mathbb{E}(G(u)) - Q_n(S, G(u^s_h))|^2} \lesssim n^{-1/(2\lambda)} + h\|f\|_{L^2(D)} + s^{-2/\lambda+1}.$$

If the assumptions of Lemma 35 are satisfied, we even obtain

$$\sqrt{\mathbb{E}_S|\mathbb{E}(G(u)) - Q_n(S, G(u^s_h))|^2} \lesssim n^{-1/(2\lambda)} + h^2\|g\|_{L^2(D)}\|f\|_{L^2(D)} + s^{-2/\lambda+1}.$$

This follows from the fact that Lemma 36 also applies to $\|\widehat{g} - g_h\|_{H^1(\Omega)}$ since $\widehat{g}$ is the solution of the same problem with different right-hand side.

2.4. **Cost of the approximation.** The approximate number of elements in $\mathcal{T}_h$ is $\mathcal{O}(h^{-d})$. On each element, we have to evaluate s-terms of $A(\boldsymbol{\omega}, x)$, i.e., $\mathcal{O}(sh^{-d})$. If we use a good preconditioner as well as an iterative solver, we can assume that the solution of the linear system to compute $u_h^s(\boldsymbol{\omega})$ costs $\mathcal{O}(h^{-d})$. Finally, the quadrature has a setup cost of $\mathcal{O}(sn\log(n))$ (the CBC-algorithm) and an online cost of $\mathcal{O}(mnsh^{-d})$, where $m \in \mathbb{N}$ is the number of random shifts (we assume $m \simeq 1$ for simplicity). Thus, to achieve an error of $\varepsilon > 0$, we need to choose

$$n \simeq \varepsilon^{-2\lambda}, \quad h \simeq \varepsilon^{-1/2}, \quad s \simeq \varepsilon^{1/(1-2/\lambda)}.$$

This results in a cost estimate of $\mathcal{O}(\varepsilon^{-(2\lambda+d/2)+1/(1-2/\lambda)}$.

2.5. **Multi-level QMC.** In the previous section, we saw that the different error contributions are additive in the total error (FEM error + QMC error + truncation error). The total cost, however, is multiplicative (FEM cost × QMC cost × truncation cost). This can lead to very high, and in practice, unmanagable computational cost to reach a prescribed accuracy.

On way out of this are so-called Multi-level methods. The idea is the following. We rewrite the approximation as

$$Q_n(G(u_h^s)) \approx \sum_{\ell=0}^{N} Q_{n_\ell}(G(u_{h_\ell}^s - u_{h_{\ell-1}}^s)), \tag{10}$$

where $u_{h_{-1}}^s := 0$, $h_{\ell+1} = h_\ell/2$, and $h_N := h$. For $n_\ell = n$, we have equality in the above approximation due to the telescoping sum. The big advantage of this reformulation is, that we can adaptively chose $n_\ell \ll n$ in case $|G(u_{h_\ell}^s - u_{h_{\ell-1}}^s)| \ll 1$. This is illustrated in the following lemma.

Lemma 37. *Under the assumptions of Theorem 33 and Lemma 36, where we additionally assume $V_{h_{\ell-1}} \subseteq V_{h_\ell}$ for all $\ell \in \mathbb{N}$, there holds*

$$\sqrt{\mathbb{E}_S|(I_s - Q_{n_\ell}(S,\cdot))(G(u_{h_\ell}^s - u_{h_{\ell-1}}^s))|^2} \le C\frac{h_\ell^2}{\phi(n_\ell)^{1/(2\lambda)}},$$

for $\lambda \in (1/2, 1]$ and some constant $C > 0$ that is independent of $\ell \in \mathbb{N}$.

Proof. First note that Lemma 29–32 remain valid if we replace u^s by u_h^s. This is due to the fact that we only used that u^s is the weak solution of a PDE in a Hilbert space. This is also true for u_h^s. Moreover, the constants are independent of $h > 0$ (Excercise: verify that yourself). Define $F(\boldsymbol{\omega}) := G(u_{h_\ell}^s - u_{h_{\ell-1}}^s)$, which is analytic in each argument due to Lemma 29. The proofs of Lemma 31–32 show that

$$|\partial_{\boldsymbol{\omega}_\mathfrak{u}} F(\boldsymbol{\omega})| \le \Big(\prod_{i=1}^{\infty} \beta_i\Big)\|F\|_{L^\infty(\Omega')}.$$

Lemma 35 shows

$$\begin{aligned}\|F\|_{L^\infty(\Omega')} &\le \sup_{\boldsymbol{\omega}\in\Omega'} |G(u^s(\boldsymbol{\omega}) - u_{h_\ell}^s(\boldsymbol{\omega}))| + |G(u^s(\boldsymbol{\omega}) - u_{h_{\ell-1}}^s(\boldsymbol{\omega}))| \\ &\lesssim \inf_{v_h \in V_{h_{\ell-1}}} \|u(\boldsymbol{\omega}) - v_h\|_{H^1(D)} \inf_{g_h \in V_{h_{\ell-1}}} \|\widehat{g}(\boldsymbol{\omega}) - g_h\|_{H^1(D)},\end{aligned}$$

where we used that $V_{h_{\ell-1}} \subseteq V_{h_\ell}$ and hence $\inf_{v_h \in V_{h_{\ell-1}}} \ldots \geq \inf_{v_h \in V_{h_\ell}} \ldots$. Lemma 36 implies

$$\|F\|_{L^\infty(\Omega')} \lesssim h_\ell^2 \|g\|_{L^2(D)} \|f\|_{L^2(D)}.$$

Thus, with the arguments from the proof of Theorem 33, we obtain

$$\sqrt{\mathbb{E}_S |(I_s - Q_{n_\ell}(S,\cdot))F|^2} \lesssim \frac{h_\ell^2}{\phi(n)^{1/(2\lambda)}}.$$

This concludes the proof. □

This leads to the following theorem.

Theorem 38. *Assume that Lemma 37 holds for all $\lambda \in (1/2, 1]$. Choosing $n_\ell \simeq (h_\ell/h_L)^{4\lambda}$ in each term of the multi-level expansion* (10) *gives*

$$\sqrt{\mathbb{E}_S |(I_s(G(u^s)) - \sum_{\ell=0}^{L} Q_{n_\ell}(S, G(u^s_{h_\ell} - u^s_{h_{\ell-1}}))|^2} \leq C L h_L^2,$$

where $L \simeq \log_2(h_L)$ and $C > 0$ is independent of L.

Proof. There holds

$$\begin{aligned}(I_s(G(u^s)) &- \sum_{\ell=0}^{L} Q_{n_\ell}(S, G(u^s_{h_\ell} - u^s_{h_{\ell-1}})) \\ &= I_s(G(u^s - u^s_{h_L})) + \sum_{\ell=0}^{L} (I_s - Q_{n_\ell}(S,\cdot)) G(u^s_{h_\ell} - u^s_{h_{\ell-1}})).\end{aligned}$$

The first term is bounded by Lemma 35 and Lemma 36 in the sense

$$|I_s(G(u^s - u^s_{h_L}))| \lesssim h_L^2 \|g\|_{L^2(D)} \|f\|_{L^2(D)}.$$

The remaining term is bounded by Lemma 37 in the sense

$$\sqrt{\mathbb{E}_S \Big| \sum_{\ell=0}^{L} (I_s - Q_{n_\ell}(S,\cdot)) G(u^s_{h_\ell} - u^s_{h_{\ell-1}})) \Big|^2} \lesssim \sum_{\ell=0}^{L} \frac{h_\ell^2}{n_\ell^{1/(2\lambda)}},$$

where we used $\phi(n) \gtrsim n^{-1+\delta}$ for all $\delta > 0$ and $n \in \mathbb{N}$. With $h_\ell \simeq 2^{-\ell}$, the choice of n_ℓ leads to

$$\frac{h_\ell^2}{n_\ell^{1/(2\lambda)}} \simeq h_L^2.$$

This concludes the proof. □

Combining the above error estimate with the estimates for the remaining error contributions, we obtain the following: If all assumptions are satisfied, there holds

$$\sqrt{\mathbb{E}_S \Big| \mathbb{E}(G(u)) - \sum_{\ell=0}^{L} Q_{n_\ell}(S, G(u^s_{h_\ell} - u^s_{h_{\ell-1}})) \Big|^2} \lesssim L h_L^2 + s^{-2/\lambda+1}.$$

Thus, to achieve an error of $\varepsilon > 0$, we need to choose

$$h_L \simeq \varepsilon^{1/2} \quad \text{and} \quad s \simeq \varepsilon^{1/(-2\lambda+1)}$$

and hence $L \simeq \frac{1}{2}\log(\varepsilon)$ Given h_L, we compute that $h_\ell \simeq 2^{L-\ell}\varepsilon^{1/2}$ and hence $n_\ell \simeq 2^{2\lambda(L-\ell)}$. The cost for each term of the multi-level expansion is again given by

$$\text{FEM cost} \times \text{truncation cost} \times \text{QMC cost}$$

In our case, this leads $\mathcal{O}(h_\ell^{-d} n_\ell s) = \mathcal{O}(\varepsilon^{-d/2+1/(1-2/\lambda)} 2^{(L-\ell)(4\lambda-d)})$. Summing up over all $L \simeq |\log_2(h_L)| \simeq |\log_2(\varepsilon)|$ levels gives a total cost estimate of

$$\mathcal{O}\Big(\varepsilon^{-d/2+1/(1-2/\lambda)} \sum_{\ell=0}^{L} 2^{(L-\ell)(4\lambda-d)}\Big) = \begin{cases} \mathcal{O}(\varepsilon^{-d/2+1/(1-2/\lambda)}) & 4\lambda < d, \\ \mathcal{O}(|\log_2(\varepsilon)|\varepsilon^{-2\lambda+1/(1-2/\lambda)}) & \text{else}, \end{cases}$$

where we used $\sum_{\ell=0}^{L} 2^{(L-\ell)(4\lambda-d)} \lesssim 1$ for $4\lambda < d$ and $\sum_{\ell=0}^{L} 2^{(L-\ell)(4\lambda-d)} \leq (L+1)2^{L(4\lambda-d)} \simeq |\log(\varepsilon)|\varepsilon^{2\lambda-d/2}$ otherwise. Compared to the single-level approach of the previous section, the exponent in the cost estimate contains only the maximum of $d/2$ and λ instead of their sum. This might not seem like much initially, however, for $\lambda = 1/2 + \delta$ with small $\delta > 0$, $d = 3$ and $\varepsilon = 10^{-2}$ (an error of one percent is a standard engineering requirement), the cost advantage of the multi-level algorithm is

$$\approx 6 \cdot 10^5 \quad \text{versus} \quad \approx 10^7 \quad \text{computational operations.}$$

This can mean the difference between hours vs. days of computational time.

3. The random parameter $A(x, \boldsymbol{\omega})$

The main goal of this section (based on [10]) is to introduce the concept of stochastic processes, i.e., random variables which depend on one or more parameters and to derive the Karhuen-Loeve expansion.

3.1. **Borel sets.** Let $(\mathcal{X}, d)$ denote a metric space. The Borel σ-algebra (σ-field) $\mathcal{B} = \mathcal{B}(\mathcal{X})$ is the smallest σ-algebra in $\mathcal{X}$ that contains the topology (all open subsets) of $\mathcal{X}$. A set $\mathcal{A} \in \mathcal{B}(\mathcal{X})$ is also called a *Borel set*.

Remark 39 (What is a σ-algebra?)**.** *A sigma-algebra on a set $\mathcal{X}$ is a subset $\Sigma \subseteq \mathcal{P}(\mathcal{X})$ of the power set of $\mathcal{X}$, which satisfies:*

1. *$\mathcal{X} \in \Sigma$,*
2. *Σ is closed under complementation, i.e., $\mathcal{A} \in \Sigma \implies \mathcal{X} \setminus \mathcal{A} \in \Sigma$,*
3. *Σ is closed under countable unions, i.e., $\mathcal{A}_1, \mathcal{A}_2, \ldots \in \Sigma \implies \bigcup_{i=1}^{\infty} \mathcal{A}_i \in \Sigma$.*

The smallest σ-algebra is $\Sigma := \{\emptyset, \mathcal{X}\}$. The pair $(\mathcal{X}, \Sigma)$ is called a measurable space*. For a subset $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$, we denote by $\sigma(\mathcal{A})$ the smallest σ-algebra which contains $\mathcal{A}$.*

Lemma 40. *If $\mathcal{X}$ is a separable metric space, then $\mathcal{B}(\mathcal{X})$ equals the σ-algebra generated by the open (or closed) balls of $\mathcal{X}$.*

Proof. Let $\mathcal{A} := \big\{B \subseteq \mathcal{X} \,:\, B \text{ is open (or closed) ball in } \mathcal{X}\big\}$. Then, obviously, there holds $\sigma(\mathcal{A}) \subseteq \mathcal{B}(\mathcal{X})$. Since $\mathcal{X}$ is seperable, we find a countable and dense set $\mathcal{D} \subseteq \mathcal{X}$. Moreover, let $B_r(x)$ denote the open (or closed) ball in $\mathcal{X}$ with center $x \in \mathcal{X}$ and radius $r > 0$ (or $r \geq 0$). Given an open (or closed) set $\mathcal{U} \subseteq \mathcal{X}$ and $x \in \mathcal{U}$, we define $y_x \in D \cap \mathcal{U}$ and $r_x > 0$ (or $r_x \geq 0$) sufficiently small with $r_x \in \mathbb{Q}$ being rational such that $x \in B_{r_x}(y_x) \subseteq \mathcal{U}$. From this, we obtain

$$\mathcal{U} = \bigcup_{x \in \mathcal{U}} B_{r_x}(y_x),$$

which is a countable union. Hence $\mathcal{U} \in \sigma(\mathcal{A})$ and we conclude $\sigma(\mathcal{A}) = \mathcal{B}(\mathcal{X})$. □

A function $f\colon \mathcal{X}_1 \to \mathcal{X}_2$ between metric spaces with corresponding σ-algeras Σ_1 and Σ_2 is called measurable iff

$$f^{-1}(\mathcal{A}) = \big\{x \in \mathcal{X}_1 \,:\, f(x) \in \mathcal{A}\big\} \in \Sigma_1 \quad \text{for all } \mathcal{A} \in \Sigma_2.$$

A function $\mu\colon \mathcal{B}(x) \to [0,\infty)$ such that

(1) $\mu(\emptyset) = 0$,
(2) $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are mutually disjoint $\implies \mu(\bigcup_{i=1}^{\infty} \mathcal{A}_i) = \sum_{i=1}^{\infty} \mu(\mathcal{A}_i)$

is called a *finite Borel measure*. If additionally, there holds $\mu(\mathcal{X}) = 1$, μ is called a *Borel probability measure*. Given a probability measure μ on $\mathcal{X}_1$ and a measureable function $f\colon \mathcal{X}_1 \to \mathcal{X}_2$, we may define the push-forward measure μ^f on $(\mathcal{X}_2, \Sigma_2)$ via

$$\mu^f(\mathcal{A}) := \mu(f^{-1}(\mathcal{A})) \quad \text{for all } \mathcal{A} \in \Sigma_2.$$

Lemma 41. *The push-forward measure μ^f is a Borel (probability) measure.*

Proof. Obviously, $f^{-1}(\emptyset) = \emptyset$ and hence $\mu^f(\emptyset) = 0$. Moreover, if $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are mutually disjoint, there holds

$$f^{-1}(\mathcal{A}_i) \cap f^{-1}(\mathcal{A}_j) = \emptyset \quad \text{for all } i \neq j$$

and hence

$$\mu^f(\bigcup_{i=1}^{\infty} \mathcal{A}_i) = \mu(f^{-1}(\bigcup_{i=1}^{\infty} \mathcal{A}_i)) = \sum_{i=1}^{\infty} \mu(f^{-1}(\mathcal{A}_i)) = \sum_{i=1}^{\infty} \mu^f(\mathcal{A}_i).$$

Finally, $f^{-1}(\mathcal{X}_2) = \mathcal{X}_1$ and hence $\mu^f(\mathcal{X}_2) = \mu(\mathcal{X}_1) = 1$ (for probability measures) concludes the proof. □

Remark 42. *In probability theory, the probability measure μ is often denoted by $\mathbb{P}$ and the function f is then called a* random variable. *Then, the common notation for the push-forward measure is*

$$\mathbb{P}(f \in \mathcal{A}) := \mu^f(\mathcal{A}) \tag{11}$$

for all $\mathcal{A} \in \Sigma_2$. This notation allows to forget *about the domain of definition of f and just consider the realizations (the elements in the range) of f.*

The common definition of expectation $\mathbb{E}(\cdot)$ can be reduced to the domain of f in the following sense: If μ^f is integrable, we obtain

$$\mathbb{E}(f) := \int_{\mathcal{X}_1} f(x)\, d\mu(x) = \int_{\mathcal{X}_2} x\, d\mu^f(x).$$

3.2. **Gaussian processes.** Before we proceed, we will require some definitions.

Definition 43 (second order)**.** *A real-valued stochastic process $X\colon D \times \Omega \to \mathbb{R}$ is called* second-order *if $X(x)\colon \Omega \to \mathbb{R} \in L^2(\Omega)$ for all $x \in D$. This allows us to define the mean function $\mu(x) := \mathbb{E}(X(x)) = \int_\Omega X(x,\omega)\, d\mathbb{P}(\omega)$ as well as the covariance function $\varrho(x,y) := \mathbb{E}(X(x)X(y))$ for all $x, y \in D$.*

Definition 44 (real-valued Gaussian process)**.** *A second-order process $X\colon D \times \Omega \to \mathbb{R}$ is called Gaussian if $(X(x_1), \ldots, X(x_n))\colon \Omega \to \mathbb{R}^n$ follows a multivariate Gaussian distribution for any $x_1, \ldots, x_n \in D$ and all $n \in \mathbb{N}$.*

A central stochastic process is the *Brownian motion*, which was discovered when Robert Brown studied the seemingly random movement of pollen in a fluid (for more detailed historic background the corresponding Wikipedia page is always a good resource). A Brownian motion is also often called a *Wiener process* (after Norbert Wiener) and hence often denoted by W.

Definition 45 (Brownian motion (Wiener process))**.** *A Brownian motion is a real-valued Gaussian process with $D = [0, \infty)$, continuous sample paths, mean function $\mu(t) = 0$, and covariance function* $\mathrm{Cov}(s, t) = \min\{s, t\}$.

We will show later, that a stochastic process satisfying the definition of a Brownian motion actually exists. For now, assume that $W(t)$ is a Brownian motion. It is easy to see that

$$\begin{aligned}
&\mathbb{E}((W(t_2) - W(t_1))(W(s_2) - W(s_1))) \\
&\qquad = \mathrm{Cov}(W(t_2), W(s_2)) - \mathrm{Cov}(W(t_1), W(s_2)) \\
&\qquad\quad - \mathrm{Cov}(W(t_2), W(s_1)) + \mathrm{Cov}(W(t_1), W(s_1)) \\
&\qquad = \min\{t_2, s_2\} - \min\{t_1, s_2\} - \min\{t_2, s_1\} + \min\{t_1, s_1\} = 0
\end{aligned}$$

for $t_1 \leq t_2 \leq s_1 \leq s_2$. This shows that the increments $W(t_2) - W(t_1)$ are uncorrelated. Since two increments have joint Gaussian distribution by definition, the increments are even independent. With $t_i = s_i$, $i = 1, 2$, we also find that

$$\mathrm{Var}(W(t_2) - W(t_1)) = |t_2 - t_1|$$

and hence $W(t) - W(s) \sim N(0, |t - s|)$.

Remark 46 (Reminder: Independence of random variables)**.** *Two subsets $\Omega_1, \Omega_2 \subseteq \Omega$ of a probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$ are called independent iff $\mathbb{P}(\Omega_1 \cap \Omega_2) = \mathbb{P}(\Omega_1)\mathbb{P}(\Omega_2)$. Two sub-$\sigma$-algebras $\Sigma_1, \Sigma_2 \subseteq \Sigma_\Omega$ are called independent if all pairs of elements $(\Omega_1, \Omega_2) \in \Sigma_1 \times \Sigma_2$ are independent. Two random variables $f_i \colon \Omega \to \mathcal{X}$ for a measurable space $(\mathcal{X}, \Sigma)$ are called independent if the σ-algebras $f_i^{-1}(\Sigma)$, $i = 1, 2$ are independent.*

Due to the above considerations, we are able to give another definition of a Brownian motion:

Definition 47 (Brownian motion (second definition))**.** *A Brownian motion is a real-valued stochastic process W on $D = [0, \infty)$ such that*

(1) *$W(0) = 0$ almost surely,*
(2) *the increments satisfy $W(t) - W(s) \sim N(0, |t - s|)$ and increments over disjoint intervals are independent,*
(3) *W has continuous sample paths.*

Lemma 48. *The two definitions of the Brownian motion are equivalent.*

Proof. We already argued that Definition 45 implies Definition 47 ($W(0)$ is Gaussian with zero variance and zero mean and hence $W(0) = 0$. To see that the second definition implies the first one, we first observe that

$$\mu(t) = \mathbb{E}(W(t)) = \mathbb{E}(W(t) - W(0)) + \mathbb{E}(W(0)) = 0$$

since $W(0) = 0$ and $W(t) - W(0) \sim N(0, t)$. Second, there holds for $t > s$

$$\begin{aligned}
\mathrm{Cov}(W(t), W(s)) &= \mathrm{Cov}(W(t) - W(s), W(s) - W(0)) + \mathrm{Cov}(W(s), W(s)) \\
&= 0 + \mathbb{E}((W(s) - W(0))^2) = s,
\end{aligned}$$

since $W(s) - W(0) \sim N(0, s)$. By symmetry of the covariance functions, we conclude $\mathrm{Cov}(W(t), W(s)) = \min\{t, s\}$. □

The second definition leads to a straightforward algorithm for sampling a Brownian motion:

Algorithm 3. ***Input:*** *evaluation points $t_1 < t_2 < \ldots < t_n$ with $t_1 = 0$.*
Set $W(0) = 0$. For $j = 2, \ldots, n$ do:

(1) *Generate standard normal random number z_j (`randn()` in Matlab).*
(2) *Define $W(j) := W(j-1) + \sqrt{t_j - t_{j-1}} z_j$.*

3.3. **Gaussian processses and the covariance function.** Given the domain D, let $\mathbb{R}^D$ denote the set of all functions $f\colon D \to \mathbb{R}$, and let $\mathcal{B}(\mathbb{R}^D)$ denote the smallest σ-algebra that contains all sets

$$\mathcal{A} = \left\{ f \in \mathbb{R}^D \,:\, [f(x_1), \ldots, f(x_N)] \in F \right\} \tag{12}$$

for $N \in \mathbb{N}$, $x_1, \ldots, x_N \in D$, and $F \in \mathcal{B}(\mathbb{R}^N)$. Note that $\mathcal{B}(\mathbb{R}^D)$ is the Borel σ-algebra with respect to the topology of pointwise convergence (not a metric space, however). By definition, the sample paths of a real-valued stochastic process $X(\cdot, \omega)$ belong to $\mathbb{R}^D$.

Lemma 49. *Let $(\Omega, \Sigma, \mathbb{P})$ denote the underlying probability space. The map $\omega \mapsto X(\cdot, \omega)$ from (Ω, Σ) to the measurable space $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$ is measurable. Thus, the sample path is a $\mathbb{R}^D$ valued random variable.*

Proof. Let $\mathcal{A} \in \mathcal{B}(\mathbb{R}^D)$ as defined in (12) and the corresponding $F \in \mathcal{B}(\mathbb{R}^N)$. The topology of $\mathbb{R}^N$ is countably generated and hence it suffices to consider $F = F_1 \times \ldots \times F_N$ for open $F_i \subseteq \mathbb{R}$. We consider

$$\begin{aligned} \left\{ \omega \in \Omega \,:\, X(\cdot, \omega) \in \mathcal{A} \right\} &= \left\{ \omega \in \Omega \,:\, [X(x_1, \omega), \ldots, X(x_N, \omega)] \in F \right\} \\ &= \bigcap_{i=1}^{N} X(x_i, \cdot)^{-1}(F_i). \end{aligned}$$

Since the $X(x_i, \cdot)$ are Σ-measurable functions by definition, we show that $\left\{ \omega \in \Omega \,:\, X(\cdot, \omega) \in \mathcal{A} \right\} \in \Sigma$ and hence conclude the proof. □

With this, we may define independence of processes.

Definition 50. (1) *Two real-valued processes $X(\cdot), Y(\cdot)$ are independent processes if the associated $(\mathbb{R}^D, \mathcal{B}(\mathcal{R}^D))$ random variables are independent. This is equivalent to the fact that*

$$[X(x_1), \ldots, X(x_M)] \quad \text{and} \quad [Y(y_1), \ldots, Y(y_M)]$$

are independent multi-variate random variables for all $M \in \mathbb{N}$ and all $x_i, y_i \in D$.

(2) *We say $f_1, f_2\colon D \to \mathbb{R}$ are independent sample paths of a real-valued process $X(\cdot)$ if $f_i(x) = X_i(x, \omega)$ for some $\omega \in \Omega$, where $X_i(\cdot)$ are i.i.d. processes with the same distribution as $X(\cdot)$ (if they have the same push-forward measure $\mathbb{P}^X(\cdot) = \mathbb{P} \circ X^{-1}$ on $\mathcal{B}(\mathbb{R}^D)$).*

The following theorem states that stochastic processes are essentially characterized by their projections on finitely many evaluation points.

Theorem 51 (Daniell-Kolmogorov Theorem)**.** *Suppose that for each set $\{x_1, \ldots, x_N\} \subset D$, there exists a probability measure $\mu_{x_1, \ldots, x_N}$ on $\mathbb{R}^N$ such that*

(1) *for any permutation π of $\{1, \ldots, N\}$ and all $F_1, \ldots, F_N \in \mathcal{B}(\mathbb{R})$,*

$$\mu_{x_{\pi(1)}, \ldots, x_{\pi(N)}}(F_{\pi(1)} \times \ldots \times F_{\pi(N)}) = \mu_{x_1, \ldots, x_N}(F_1 \times \ldots \times F_N),$$

(2) *for $M < N$ and any $F \in \mathcal{B}(\mathbb{R}^M)$*

$$\mu_{x_1, \ldots, x_N}(F \times \mathbb{R}^{N-M}) = \mu_{x_1, \ldots, x_M}(F).$$

Then, there exists a stochastic process $X(\cdot)$ with finite-dimensional distributions $\mathbb{P}^{[X(x_1),\dots,X(x_N)]} = \mu_{x_1,\dots,x_N}$ for all $x_1,\dots,x_N \in D$. If $X(\cdot)$ and $Y(\cdot)$ are two such processes, then there holds $\mathbb{P}^X = \mathbb{P}^Y$ on $\mathcal{B}(\mathbb{R}^D)$.

Remark 52. *We note that each stochastic process satisfies the conditions (1–2) of the above theorem. To see that, note that*

$$\begin{aligned}\mathbb{P}^{[X(x_1),\dots,X(x_N)]}(F_1 \times \ldots \times F_N) &= \mathbb{P}(\bigcap_{i=1}^N X(x_i)^{-1}(F_i)) = \mathbb{P}(\bigcap_{i=1}^N X(x_{\pi(i)})^{-1}(F_{\pi(i)})) \\ &= \mathbb{P}^{[X(x_{\pi(1)}),\dots,X(x_{\pi(N)})]}(F_{\pi(1)} \times \ldots \times F_{\pi(N)}).\end{aligned}$$

Moreover, there holds

$$\begin{aligned}\mathbb{P}^{[X(x_1),\dots,X(x_N)]}(F \times \mathbb{R}^{N-M}) &= \mathbb{P}\big([X(x_1),\dots,X(x_M)]^{-1}(F) \cap [X_{M+1},\dots,X_N]^{-1}(\mathbb{R}^{N-M})\big) \\ &= \mathbb{P}^{[X(x_1),\dots,X(x_M)]}(F).\end{aligned}$$

We recall that a positive (semi-)definite function $\varrho\colon D \times D \to \mathbb{R}$ satisfies for all $x_1,\dots,x_N \in D$ and all $a_1,\dots,a_N \in \mathbb{R}$ that

$$\sum_{i,j=1}^N a_i a_j \varrho(x_i,x_j) > (\geq) 0.$$

This is equivalent to the fact that the induced matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$, $\boldsymbol{C}_{ij} := \varrho(x_i,x_j)$ is positive (semi-)definite.

Theorem 53. *The following statements are equivalent:*

(i) *There exists a real-valued second order stochastic process $X(\cdot)$ with mean function μ and covariance function* Cov*.*

(ii) *Let $\mu\colon D \to \mathbb{R}$ and* Cov$\colon D \times D \to \mathbb{R}$ *with* Cov *being symmetric and positive semi-definite.*

Particularly, (ii) even implies existence of a Gaussian process.

Proof of $(i) \implies (ii)$. The functions are well-defined since the process is second-order. There holds for any $x_1,\dots,x_N \in D$ and $a_1,\dots,a_N \in \mathbb{R}$ that

$$\begin{aligned}\sum_{i,j=1}^N a_i a_j \mathrm{Cov}(x_i,x_j) &= \mathbb{E}\Big(\sum_{i,j=1}^N a_j a_k (X(x_i)-\mu(x_i))(X(x_j)-\mu(x_j))\Big) \\ &= \mathbb{E}\Big(\sum_{i=1}^N a_j (X(x_i)-\mu(x_i))\Big)^2 \geq 0.\end{aligned}$$

Obviously, $\mathrm{Cov}(x,y)$ is also symmetric. □

Proof of $(ii) \implies (i)$. Define the covariance matrix $\boldsymbol{C} \in \mathbb{R}^{N\times N}$, $\boldsymbol{C}_{ij} := \mathrm{Cov}(x_i,x_j)$. By definition, $\boldsymbol{C}$ is symmetric and positive semi-definite. Hence, we may consider a random variable Y obeying a multivariate Gaussian distribution $Y \sim N(0,\boldsymbol{C})$. We define $\mu_{x_1,\dots,x_N} := \mathbb{P}^Y$. Theorem 51 applies since $\mu_{x_{\pi(1)},\dots,x_{\pi(N)}}$ is the push-forward measure of $\widetilde{Y} \sim N(0,\widetilde{\boldsymbol{C}})$ with $\widetilde{\boldsymbol{C}}_{ij} := \boldsymbol{C}_{\pi(i),\pi(j)}$. Hence

$$\begin{aligned}\mu_{x_{\pi(1)},\dots,x_{\pi(N)}}(F_{\pi(1)} \times \ldots F_{\pi(N)}) &= \mathbb{P}(\widetilde{Y} \in F_{\pi(1)} \times \ldots F_{\pi(N)}) \\ &= \mathbb{P}(Y \in F_1 \times \ldots F_N) = \mu_{x_1,\dots,x_N}(F_1 \times \ldots F_N).\end{aligned}$$

Moreover, there holds

$$\mu_{x_1,\dots,x_N}(F \times \mathbb{R}^{N-M}) = \mathbb{P}([Y_1,\dots,Y_M] \in F) = \mu_{x_1,\dots,x_M}(F).$$

This follows from the fact, that the marginal distribution of a multi-variate Gaussian is obtained by just dropping the corresponding rows and columns in the covariance matrix, i.e., we consider

$$\mu_{x_1,\dots,x_N}(F \times \mathbb{R}^{N-M}) = \int_F \int_{\mathbb{R}^{N-M}} (2\pi)^{-N/2} \det(\boldsymbol{C})^{-1/2} e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{C}^{-1}\boldsymbol{x}}\, d\boldsymbol{x}.$$

Since $\boldsymbol{C}$ is symmetric, we obtain a block-Cholesky factorization, i.e., $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$, where

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2 \end{pmatrix}$$

with $\boldsymbol{D}_1 \in \mathbb{R}^{M\times M}$ and $\boldsymbol{D}_2 \in \mathbb{R}^{(N-M)\times(N-M)}$ and $\boldsymbol{L}$ is the corresponding block-lower-triangular matrix with two identity block in the diagonal. This allows us to transform the integral with $(\boldsymbol{y}_1, \boldsymbol{y}_2) = \boldsymbol{L}^{-1}\boldsymbol{x}$ to

$$\begin{aligned}\mu_{x_1,\dots,x_N}(F \times \mathbb{R}^{N-M}) &= \int_F (2\pi)^{-M/2} \det(\boldsymbol{D}_1)^{-1/2} e^{-\frac{1}{2}(\boldsymbol{y}_1)^T \boldsymbol{D}_1^{-1}\boldsymbol{y}_1}\, d\boldsymbol{y}_1 \\ &\qquad \times \int_{\mathbb{R}^{N-M}} (2\pi)^{-(N-M)/2} \det(\boldsymbol{D}_2)^{-1/2} e^{-\frac{1}{2}\boldsymbol{y}_2^T \boldsymbol{D}_2^{-1}\boldsymbol{y}_2}\, d\boldsymbol{y}_2 \\ &= \mu_{x_1,\dots,x_M}(F).\end{aligned}$$

Therefore, Theorem 51 guarantees the existence of a real valued stochastic process $Y(\cdot)$ with the finite-dimensional distributions $\mu_{x_1,\dots,x_N}$. In particular, the distribution of $[Y(x), Y(y)]$ is

$$[Y(x), Y(y)] \sim N\Big(0, \begin{pmatrix} \mathrm{Cov}(x,x), \mathrm{Cov}(x,y) \\ \mathrm{Cov}(y,x), \mathrm{Cov}(y,y) \end{pmatrix}\Big).$$

This shows that the convariance function of $Y(\cdot)$ is $\mathrm{Cov}(x,y)$. Since Y has zero mean, we conclude the proof with $X := Y + \mu$. □

Corollary 54. *The probability distribution $\mathbb{P}^X$ on $(\mathbb{R}^D, \mathcal{B}(\mathcal{R}^D))$ of a real-valued Gaussian process $X(\cdot)$ is uniquely determined by its mean $\mu\colon D \to \mathbb{R}$ and convariance function $\mathrm{Cov}\colon D \times D \to \mathbb{R}$.*

Proof. Let $X(\cdot)$ and $Y(\cdot)$ denote Gaussian processes with the same mean and covariance functions. We know that the finite-dimensional distributions $\mathbb{P}^{[X(x_1),\dots,X(x_N)]}$ and $\mathbb{P}^{[Y(x_1),\dots,Y(x_N)]}$ are multivariate Gaussian. A multivariate Gaussian distribution is uniquely determined by its mean and covariance information. Hence $X(\cdot)$ and $Y(\cdot)$ have the same finite dimensional distributions. Theorem 51 concludes that $\mathbb{P}^X = \mathbb{P}^Y$ on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$. □

The most prominent examples of covariance functions are the Gaussian and the exponential kernel, defined by:

$$\mathrm{Cov}(x,y)^{\text{Gaussian}} := e^{-|x-y|^2/\lambda} \quad \text{and} \quad \mathrm{Cov}(x,y)^{\text{exponential}} := e^{-|x-y|/\lambda},$$

where $\lambda \geq 0$ is the so-called *correlation length*. Those functions are limit cases of the more general Matérn class of covariance functions

$$\mathrm{Cov}(x,y)^{\text{Matérn}} := \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\sqrt{2\nu}\frac{|x-y|}{\lambda}\Big)^{\nu} K_\nu\Big(\sqrt{2\nu}\frac{|x-y|}{\lambda}\Big),$$

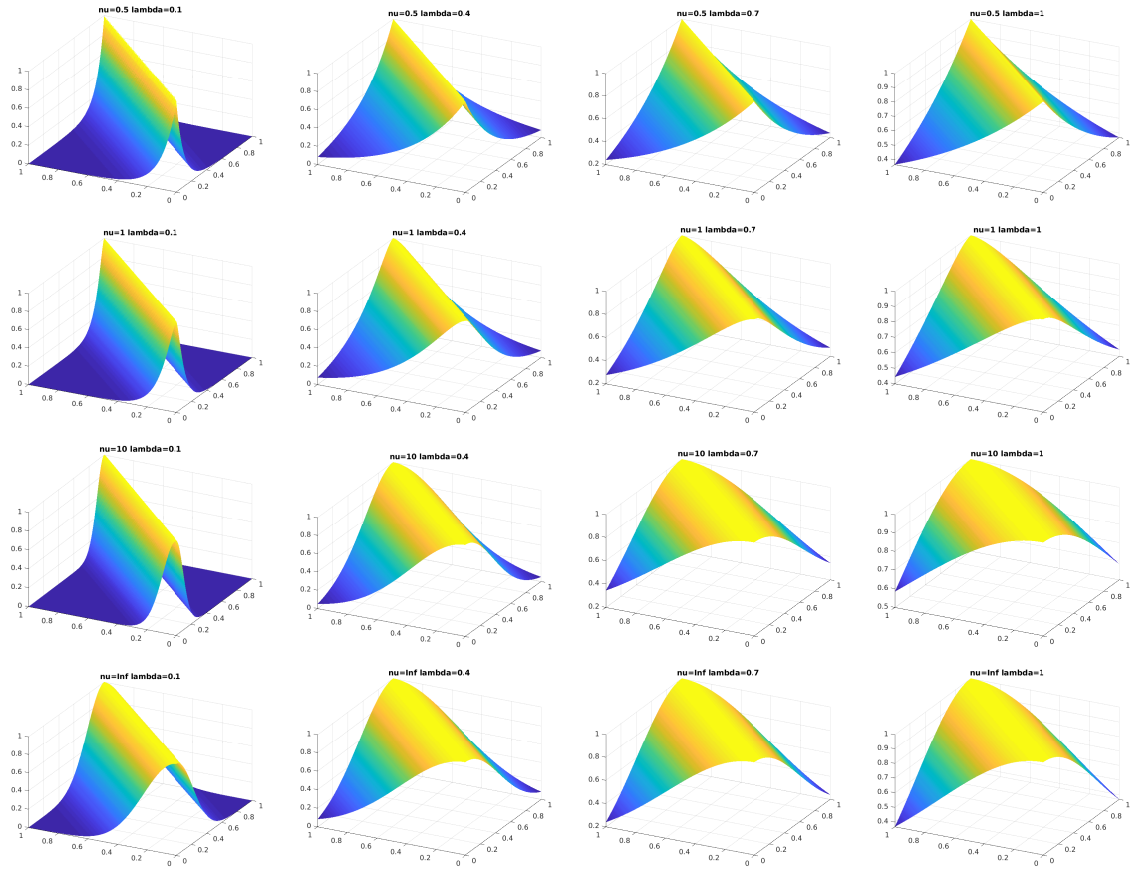

FIGURE 5. Matern covariance kernels with different parameters λ and ν.

where K_ν is the modified Bessel function of the second kind (again Wikipedia is your friend) and $\lambda, \nu \geq 0$. The parameter ν is a smoothness parameter and $\nu = 1/2$ gives the exponential covariance whereas $\nu \to \infty$ gives the Gaussian covariance. See also Figure 5 for the illustration of the kernels with different parameters. A standard result in measure theory (Bochner's theorem) shows that Cov is a positive definite function if and only if the Fourier transform is a positive (pointwise large than zero) function. This can be used to show that $\mathrm{Cov}(x,y)^{\text{Matérn}}$ is positive definite.

Lemma 55. *The covariance function* $\mathrm{Cov}(s,t) := \min\{s,t\}$ *is symmetric and positive semi-definite.*

Proof. While symmetry is obvious, we need to check for positivity. To that end let $t_1, \dots, t_n \geq 0$ and $a_1, \dots, a_n \in \mathbb{R}$. Without loss of generality, we may assume $t_i \leq t_{i+1}$ for all $i = 1, \dots, n-1$. Define the matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ with $\boldsymbol{C}_{ij} := \mathrm{Cov}(t_i, t_j) = t_{\min\{i,j\}}$. This results in

$$\boldsymbol{C} = \begin{pmatrix} t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & \dots & t_2 \\ \vdots & & & \\ t_1 & t_2 & \dots & t_n \end{pmatrix} = \sum_{m=1}^{n} c_m \boldsymbol{E}_m, \quad \text{with} \quad \boldsymbol{E}_m := \begin{pmatrix} \mathbf{0}_{(n-m)\times(n-m)} & \mathbf{0}_{(n-m)\times m} \\ \mathbf{0}_{m\times(n-m)} & \mathbf{1}_{m\times m}, \end{pmatrix}$$

with coefficients $c_m := t_{n-m+1} - t_{n-m} \geq 0$. It is easy to see that $x \cdot \boldsymbol{E}_m x = (\sum_{n+1-m}^{m} x_j)^2 \geq 0$ is positive semi-definite. Since also all c_m are non-negative, we obtain that $\boldsymbol{C}$ is positive semi-definite. This concludes the proof. □

Remark 56. *Lemma 55 together with Theorem 53 imply that there exists a stochastic process $X(\cdot)$ on $[0,\infty)$ which satisfies all properties of a Brownian motion except continuity. We will see later, that the sample path regularity can be deduced from the regularity of the covariance function and hence show that a Brownian motion really exists.*

A very straightforward way to sample Gaussian processes is via the covariance matrix. Say one wants to know the values of a Gaussian process $X(\cdot)$ at the nodes $x_1,\dots,x_n$. Then, one can assemle the covariance matrix $\boldsymbol{C}\in\mathbb{R}^{n\times n}$, $\boldsymbol{C}_{ij}:=\mathrm{Cov}(x_i,x_j)$ and the mean vector $\boldsymbol{\mu}\in\mathbb{R}^n$, $\boldsymbol{\mu}_i:=\mu(x_i)$. Next, one computes *a square-root* $\boldsymbol{C}=\boldsymbol{R}^T\boldsymbol{R}$ and evaluates

$$\boldsymbol{R}^T bx+\boldsymbol{\mu},$$

where $\boldsymbol{x}\in\mathbb{R}^n$ are i.i.d standard normal random numbers (e.g., generated by `randn()` in Matlab). It is very easy to see that the resulting process has mean function zero. Moreover, the covariance matrix

$$\mathbb{E}((\boldsymbol{R}^T\boldsymbol{x})_i(\boldsymbol{R}^T\boldsymbol{x})_j)=\sum_{k,m=1}^{n}(\boldsymbol{R}^T)_{ik}(\boldsymbol{R}^T)_{jm}\mathbb{E}(\boldsymbol{x}_k\boldsymbol{x}_m)=\sum_{k=1}^{n}(\boldsymbol{R}^T)_{ik}\boldsymbol{R}_{kj}=\boldsymbol{C}_{ij}$$

equals that of $X(x_1,\dots,x_n)$. Hence, we produced exact samples of $X(\cdot)$ at the nodes $x_1,\dots,x_n$. Note that $\boldsymbol{R}=\boldsymbol{R}^T=\boldsymbol{C}^{1/2}$ is only one of the possible choices, another one being the Cholesky factorization with upper triangular $\boldsymbol{R}$. When done in Matlab with `chol`, one has to be aware of the fact that Matlab often uses pivoting to increase the stability of the decomposition and hence implicitly permutes the vector $(x_1,\dots,x_n)$. Some Python libraries return the upper triangle of the Cholesky factorization by default, and hence one has to use the transposed return value. See Figure 6 for some samples of Gaussian processes.

3.4. **The Karhunen-Loève expansion.** The KL-expansion is an attempt to separate spatial and random dependencies of stochastic processes. Given a stochastic process $X(\cdot)$ with mean function $\mu(x)$, we are interested in writing the sample paths $X(x,\omega)-\mu(x)$ in a orthonormal basis, i.e.,

$$X(x,\omega)=\mu(x)+\sum_{j=1}^{\infty}\gamma_j(\omega)\phi_j(x), \tag{13}$$

where the $(\phi_j)_{j\in\mathbb{N}}$ are a $L^2(D)$-orthonormal basis and the coefficients γ_j are random variables given by

$$\gamma_j(\omega):=\int_D(X(x,\omega)-\mu(x))\phi_j(x)\,dx.$$

Let $\mathrm{Cov}(x,y)$ denote the covariance function of X and define the integral operator $\mathcal{C}\colon L^2(D)\to L^2(D)$ by

$$(\mathcal{C}f)(x):=\int_D\mathrm{Cov}(x,y)f(y)\,dy.$$

The expansion (13) is called KL-expansion if the ϕ_j are chosen as eigenfunctions of $\mathcal{C}$.

Lemma 57. *If $X\in L^2(\Omega,L^2(D))$, there holds $\mu\in L^2(D)$ as well as $\mathrm{Cov}(\cdot,\cdot)\in L^2(D\times D)$ and the sample paths $X(\cdot,\omega)\in L^2(D)$ for almost all $\omega\in\Omega$.*

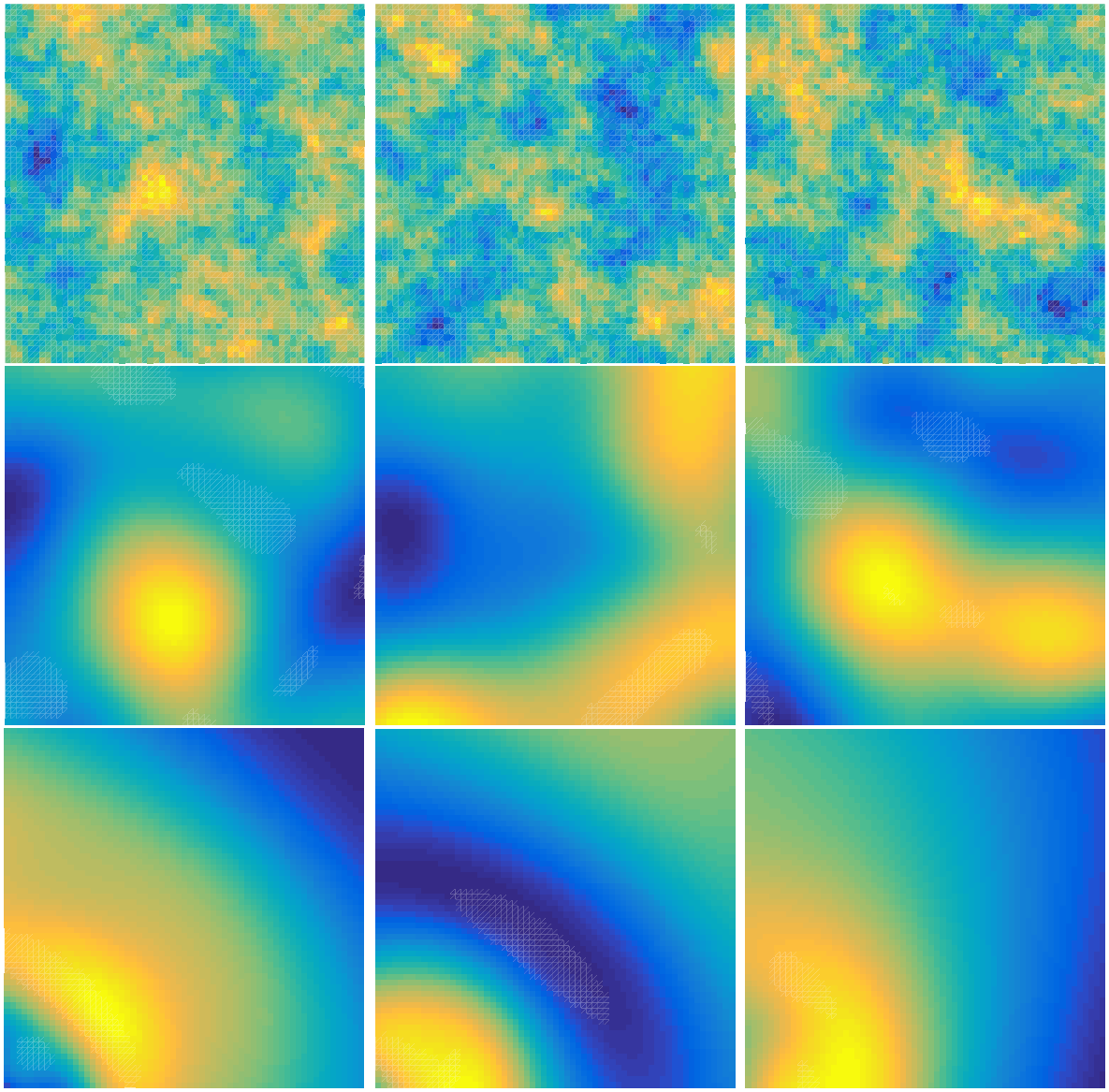

FIGURE 6. Samples of Gaussian processes on $D := [0,1]^2$ (Gaussian random fields). The first line of pictures corresponds to the exponential covariance, the second line corresponds to the Gaussian covariance, and the third line corresponds to a non-stationary covariance function $\varrho(x,y)$ which forces larger covariance in the lower-left corner. (Stationary covariance functions satisfy $\varrho(x,y) = \rho(|x-y|)$ for some function $\rho(\cdot)$ and thus produce samples with similar irregularities in the whole domain D.)

Proof. The assumption $\|X\|_{L^2(\Omega;L^2(D))} < \infty$ implies $\|X(\cdot,\omega)\|_{L^2(D)} < \infty$ almost everywhere in Ω and hence $X(\cdot,\omega) \in L^2(D)$ almost surely. Jensen's inequality implies

$$\begin{aligned}\|\mu\|^2_{L^2(D)} &= \int_D \mu(x)^2\,dx \le \int_D \mathbb{E}(X(x)^2)\,dx \\ &= \int_D \int_\Omega |X(x,\omega)|^2\,dx\,d\omega = \int_\Omega \int_D |X(x,\omega)|^2\,dx\,d\omega = \|X\|^2_{L^2(\Omega,L^2(D))} < \infty.\end{aligned}$$

Finally, a Hölder inequality shows

$$\int_{D\times D} \mathrm{Cov}(x,y)^2\,dx\,dy = \int_{D\times D} \Big(\mathbb{E}((X(x)-\mu(x))(X(y)-\mu(y)))\Big)^2\,dx\,dy \\ \leq \Big(\int_D \mathbb{E}((X(x)-\mu(x))^2)\,dx\Big)^2.$$

This is finite due to $X \in L^2(\Omega, L^2(D))$ and we see $\mathrm{Cov} \in L^2(D\times D)$. This concludes the proof. □

Lemma 58. *Consider a process $X \in L^2(\Omega, L^2(D))$. Then,*

$$X(x,\omega) = \mu(x) + \sum_{j=1}^{\infty} \sqrt{\nu_j}\phi_j(x)\xi_j(\omega),$$

where the sum converges in $L^2(\Omega, L^2(D))$,

$$\xi_j((\omega) := \nu_j^{-1/2} \int_D (X(x,\omega)-\mu(x))\phi_j(x)\,dx,$$

and the (ν_j, ϕ_j) denote the eigenvalues and eigenfunctions of the covariance operator $\mathcal{C}\colon L^2(D) \to L^2(D)$. The random variables ξ_j have mean zero, unit variance and are pairwise uncorrelated. If the process is Gaussian, then $\xi_j \sim N(0,1)$, i.i.d.

Proof. The theory of Hilbert-Schmidt operators shows that $\mathcal{C}\colon L^2(D) \to L^2(D)$ is a compact operator since $\mathrm{Cov} \in L^2(D\times D)$ as shown in Lemma 57. Therefore, the spectral theorem provides an orthonormal basis $(\phi_j)_{j\in\mathbb{N}}$ of eigenfunctions of $\mathcal{C}$. Since

$$\begin{aligned}\int_D \mathcal{C}(f)(x)f(x)\,dx &= \int_D\int_D \mathrm{Cov}(x,y)f(y)f(x)\,dx,\,dy \\ &= \int_D\int_D \mathbb{E}((X(x)-\mu(x))(X(y)-\mu(y)))f(y)f(x)\,dx\,dy \\ &= \mathbb{E}(\Big(\int_D (X(x)-\mu(x))f(x)\,dx\Big)^2) \geq 0,\end{aligned}$$

we see that the corresponding eigenvalues ν_j are non-negative. Define the truncated expansion

$$X_J(x,\omega) := \mu(x) + \sum_{j=1}^{J} \sqrt{\nu_j}\phi_j(x)\xi_j(\omega).$$

Since $(\phi_j)_{j\in\mathbb{N}}$ is an ONB of $L^2(D)$, we have $X_J(\cdot,\omega) \to X(\cdot,\omega)$ in $L^2(D)$ almost everywhere in Ω and moreover, we have

$$\|X_J(\cdot,\omega)\|_{L^2(D)} \leq \|X(\cdot,\omega)\|_{L^2(D)}$$

and hence $X_J \to X$ in $L^2(\Omega, L^2(D))$ by the dominated convergence theorem. Finally, we see that

$$\begin{aligned}\mathrm{Cov}(\xi_i,\xi_j) &= \nu_i^{-1/2}\nu_j^{-1/2}\mathbb{E}\Big(\int_D\int_D (X(x)-\mu(x))\phi_i(x)(X(y)-\mu(y))\phi_j(y)\,dx\,dy\Big) \\ &= \nu_i^{-1/2}\nu_j^{-1/2}\int_D\int_D \mathrm{Cov}(x,y)\phi_i(x)\phi_j(y)\,dx\,dy = \nu_i^{-1/2}\nu_j^{-1/2}\int_D \mathcal{C}(\phi_i)(x)\phi_j(x)\,dx.\end{aligned}$$

Since the ϕ_j are eigenfunctions of $\mathcal{C}$ and orthogonal in $L^2(D)$, the last expression is zero for $i \neq j$ and one for $i = j$. This shows the statement about correlation and variance. If X is Gaussian, the ξ_j are Gaussian since they are linear functions of X. Since the ξ_i

and ξ_j then have joint Gaussian distribution $\mathrm{Cov}(\xi_i, \xi_j) = 0$ implies that ξ and ξ_j are independent. This concludes the proof. □

3.5. **Sample path continuity.** For practical computations, the regularity of the random coefficient in the spatial variable $x \in D$ is important. After all, to compute the stiffness matrix of the FEM, one needs to compute integrals of the form

$$\int_T A(x, \boldsymbol{\omega}) p(x)\, dx$$

for a finite element $T \subseteq D$ and polynomials p of a certain degree (degree zero for lowest order FEM) by use of quadrature. In the following, we show that $x \mapsto A(x, \boldsymbol{\omega})$ is Hölder continuous almost surely and hence classical quadrature rules produce a good approximation. Recall the Hölder norm of exponent $0 < \alpha \leq 1$, i.e.,

$$|f|_{C^\alpha(D)} := \sup_{x,y \in D} \frac{|f(x) - f(y)|}{|x - y|^\alpha}.$$

We also recall Chebyshev's inequality bounding the deviation from the mean for a random variable X and $k > 0$, $p \in \mathbb{N}$

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq k) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^p}{k^p}.$$

Remark 59. *Note that any (sufficiently general) process $X(\cdot)$ can not have more regular sample paths than its mean function $\mu(x) := \mathbb{E}(X(x))$. Hence, we assume vanishing mean in the following. For sufficiently smooth $\mu(x)$, we can just consider sample path regularity of $X(\cdot) - \mu(\cdot)$.*

Definition 60. *We say that a process Y is a continuous version of X, if $Y = X$ for almost all $\omega \in \Omega$ and Y has continuous sample paths.*

Theorem 61. *Let $D \subseteq \mathbb{R}^d$ and $X(\cdot)$ a stochastic process with vanishing mean such that, for some $p, r, K > 0$, there holds*

$$\mathbb{E}\Big(|X(x) - X(y)|^p\Big) \leq K|x - y|^{d+r}$$

for all $x, y \in D$. Then, there exists a continuous version of $X(\cdot)$. If $r > p\alpha$ for some $\alpha > 0$, there even exists a Hölder continuous version of $X(\cdot)$ with Hölder exponent α.

Lemma 62. *Suppose that $X(\cdot)$ and $Y(\cdot)$ are stochastic processes with vanishing mean such that*

(i) *$X(x) = Y(x)$ almost surely for x in a dense subset of D,*
(ii) *all sample paths of $Y(\cdot)$ are continuous,*
(iii) *for some $p, r, K > 0$*

$$\mathbb{E}\Big(|X(x) - X(y)|^p\Big) \leq K|x - y|^r$$

for all $x, y \in D$.

Then, $Y(\cdot)$ is a continuous version of $X(\cdot)$.

Proof. Fix $x \in D$. Since $X(x) = Y(x)$ in a dense subset of D, we may choose $x_n \in D$ such that $X(x_n) = Y(x_n)$ and $|x_n - x| \leq 2^{-n}$. Chebyshev's inequality shows

$$\mathbb{P}(|Y(x_n) - X(x)| \geq \varepsilon) = \mathbb{P}(|X(x_n) - X(x)| \geq \varepsilon) \leq \frac{\mathbb{E}(|X(x_n) - X(x)|^p)}{\varepsilon^p} \leq K\varepsilon^{-p} 2^{-nr}.$$

Let $\varepsilon_n := 2^{-nr/(2p)}$ and $F_n := \{|Y(x_n) - X(x)| \geq \varepsilon_n\}$. Then, we have $\sum_{n=1}^{\infty} \mathbb{P}(F_n) < \infty$ and the Borel-Cantelli lemma shows

$$\mathbb{P}(\limsup_{n\to\infty} F_n) = \mathbb{P}(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} F_m) = 0.$$

Thus, for almost all $\omega \in \Omega$, we find $n(\omega) \in \mathbb{N}$ such that $|Y(\omega, x_n) - X(\omega, x)| \leq \varepsilon_n$ for all $n > n(\omega)$. By taking $n \to \infty$ and by continuity of $Y(\omega, \cdot)$, we see $Y(x) = X(x)$ almost surely in Ω and hence conclude the proof. □

Proof of Theorem 61, (Continuity). For $n \in \mathbb{N}$, let $\mathcal{T}_n$ denote a triangulation of D such that each triangle $T \in \mathcal{T}_n$ satisfies $|T| \simeq 2^{-dn}$ and $\operatorname{diam}(T) \simeq 2^{-n}$. Note that those assumptions already imply uniform shape regularity of the $\mathcal{T}_n$. Moreover, we assume that $\mathcal{T}_{n+1}$ is a refinement of $\mathcal{T}_n$ for all $n \in \mathbb{N}$. Let $Y_n(\cdot)$ be the piecewise linear interpolation of $X(\cdot)$ on $\mathcal{T}_n$. We show that $Y_n(\cdot)$ converges almost surely to a limit point $Y(\cdot)$ in $C(D)$. To that end, note that on an element $T \in \mathcal{T}_n$, the largest difference between Y_{n+1} and Y_n occurs on a new node $x_T \in T$ of $\mathcal{T}_{n+1}$ which is not a node of $\mathcal{T}_n$. Since $x_T = \sum_{i=1}^{d+1} \alpha_{T,i} x_{T,i}$ is a convex combination of the nodes $x_{T,i}$ of T, we obtain

$$\begin{aligned}\|Y_{n+1} - Y_n\|_{L^\infty(T)} &\leq |Y_{n+1}(x_T) - Y_n(x_T)| = |Y_{n+1}(x_T) - \sum_{i=1}^{d+1} \alpha_{T,i} Y_n(x_{T,i})| \\ &\leq \sum_{i=1}^{d+1} \alpha_{T,i} |Y_{n+1}(x_T) - Y_n(x_{T,i})| = \sum_{i=1}^{d+1} \alpha_{T,i} |X(x_T) - X(x_{T,i})|\end{aligned}$$

where we used that the Y_n are interpolations of X. Hence, we see that

$$\mathbb{P}(\|Y_{n+1} - Y_n\|_{L^\infty(T)} \geq \varepsilon) \leq \sum_{i=1}^{d+1} \mathbb{P}(|X(x_T) - X(x_{T,i})| \geq \varepsilon).$$

Chebyshev's inequality and the assumptions on $X(\cdot)$ show

$$\mathbb{P}(|X(x_T) - X(x)| \geq \varepsilon) \leq \frac{\mathbb{E}(|X(x_T) - X(x)|^p)}{\varepsilon^p} \leq K \frac{|x_T - x|^{d+r}}{\varepsilon^p}.$$

Altogether, this shows

$$\mathbb{P}(\|Y_{n+1} - Y_n\|_{L^\infty(T)} \geq \varepsilon) \leq (d+1) K \varepsilon^{-p} 2^{-n(d+r)}.$$

Summing up over all the $\mathcal{O}(2^{dn})$ elements of $\mathcal{T}_n$, we obtain

$$\mathbb{P}(\|Y_{n+1} - Y_n\|_{L^\infty(D)} \geq \varepsilon) \leq \sum_{T\in\mathcal{T}_n} \mathbb{P}(\|Y_{n+1} - Y_n\|_{L^\infty(T)} \geq \varepsilon) \leq (d+1) K \varepsilon^{-p} 2^{-nr}.$$

Choosing $\varepsilon_n = 2^{-(1-\delta)nr/p}$ for some $\delta > 0$ and $n \in \mathbb{N}$, we end up with $\mathbb{P}(\|Y_{n+1} - Y_n\|_{L^\infty(D)} \geq \varepsilon_n) \lesssim 2^{-\delta nr}$ and hence

$$\sum_{n=1}^{\infty} \mathbb{P}(\|Y_{n+1} - Y_n\|_{L^\infty(D)} \geq \varepsilon_n) < \infty.$$

Hence, for $F_n := \{\|Y_{n+1} - Y_n\|_{L^\infty(D)} \geq \varepsilon_n\}$, the Borel-Cantelli lemma shows that $\limsup_{n\to\infty} F_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} F_n$ has probability zero. Hence, for almost all $\omega \in \Omega$ exists $n(\omega)$ such that for $n > n(\omega)$

$$\|Y_{n+1}(\omega) - Y_n(\omega)\|_{L^\infty(D)} \leq \varepsilon_n$$

almost surely. Since $\varepsilon_n := 2^{-(1-\delta)nr/p}$, almost all $\omega \in \Omega$ satisfy for $n \geq n(\omega)$ that

$$\|Y - Y_n\|_{L^\infty(D)} \leq \sum_{k=n}^{\infty} \varepsilon_k \leq 2^{-(1-\delta)nr/p} \frac{1}{1 - 2^{-(1-\delta)r/p}}. \tag{14}$$

Thus, $Y_n = Y_0 + \sum_{k=1}^{n} (Y_k - Y_{k-1})$ converges absolutely in $C(D)$ for almost all ω. Finally, define

$$Y := \begin{cases} \lim_{n\to\infty} Y_n & \text{the sum converges,} \\ Y_0 & \text{otherwise.} \end{cases}$$

Then, Y is a well-defined $C(D)$-random variable such that $Y_n \to Y$ in $C(D)$ almost surely. Moreover, $Y = Y_n(x) = X(x)$ for all nodes x of $\mathcal{T}_n$ for all $n \in \mathbb{N}$. Since $\bigcup_{n\in\mathbb{N}} \text{nodes of}(\mathcal{T}_n) \subset D$ is dense, Lemma 62 applies and concludes the $Y = X$ is continuous almost surely. □

Proof of Theorem 61 (Hölder continuity). Just as in the first part of the proof (Continuity), we have for all $x, x' \in T \in \mathcal{T}_n$ that

$$\mathbb{P}\Big(\frac{|X(x) - X(x')|}{h_T^\alpha} \geq \varepsilon / h_T^\alpha\Big) = \mathbb{P}(|X(x) - X(x')| \geq \varepsilon) \leq K \frac{|x - x'|^{d+r}}{\varepsilon^p}.$$

Note that for the elementwise affine function Y_n, there holds

$$|Y_n|_{C^\alpha(T)} = \max_{x\neq y \in \text{nodes of } T} \frac{Y_n(x) - Y_n(y)}{|x - y|^\alpha} \lesssim \max_{x,y\in\text{nodes of } T} \frac{Y_n(x) - Y_n(y)}{h_T^\alpha},$$

where the hidden constant depends only on the shape-regularity of $\mathcal{T}_n$ and $\alpha > 0$. With $\varepsilon = h_T^\alpha$ and $Y_n(x) = X(x)$ for all nodes x of T, this shows

$$\mathbb{P}(|Y_n|_{C^\alpha(T)} \geq 1) \leq K h_T^{d+r-p\alpha}.$$

Summing up over all $T \in \mathcal{T}_n$ (note that $\#\mathcal{T} \simeq h_T^{-d}$) shows

$$\mathbb{P}(\max_{T\in\mathcal{T}_n} |Y_n|_{C^\alpha(T)} \geq 1) \lesssim h_T^{r-p\alpha} \simeq 2^{-n(r-p\alpha)}.$$

With $r > p\alpha$, we define $F_n := \{\max_{T\in\mathcal{T}_n} |Y_n|_{C^\alpha(T)} \geq 1\} \subseteq \Omega$ and observe

$$\sum_{n\in\mathbb{N}} \mathbb{P}(F_n) < \infty.$$

Again, the Borel-Cantelli Lemma implies $\mathbb{P}(\limsup_{n\to\infty} F_n) = 0$ and hence for almost all $\omega \in \Omega$, we find $m(\omega) \in \mathbb{N}$ such that $\omega \notin F_n$ for all $n \geq m(\omega)$. Recall the uniform convergence (14) for all $n \geq n(\omega)$ and almost all $\omega \in \Omega$. With $r > \alpha p$, we may choose $\delta > 0$ sufficiently small such that $(1 - \delta)r/p \geq \alpha$, and hence obtain

$$\|Y - Y_n\|_{L^\infty(D)} \leq C 2^{-n\alpha} \quad \text{for all } n \geq n(\omega).$$

Given $\omega \in \Omega$, let $x, y \in D$ sufficiently close such that there exists $n \geq \max\{n(\omega), m(\omega)\}$ and $T_x, T_y \in \mathcal{T}_n$ with $T_x \cap T_y \neq \emptyset$ and $x \in T_x$, $y \in T_y$. Let $z \in T_x \cap T_y$ such that $\max\{|x-z|, |z-y|\}$ is minimal. Shape regularity of $\mathcal{T}_n$ additionally implies that $\max\{|x-$

$z|, |z-y|\} \lesssim |x-y| \simeq 2^{-n}$. Together with the above, we have

$$\begin{aligned}\frac{|Y(x)-Y(y)|}{|x-y|^\alpha} &\leq \frac{|Y(x)-Y_n(x)|+|Y(y)-Y_n(y)|}{|x-y|^\alpha} + \frac{|Y_n(x)-Y_n(y)|}{|x-y|^\alpha} \\ &\lesssim \frac{2^{-n\alpha}}{|x-y|^\alpha} + \frac{|Y_n(x)-Y_n(z)|}{|x-y|^\alpha} + \frac{|Y_n(z)-Y_n(y)|}{|x-y|^\alpha} \\ &\lesssim 1 + \max_{T\in\mathcal{T}_n} |Y_n|_{C^\alpha(D)} \lesssim 1.\end{aligned}$$

Hence, $Y(\omega)$ is locally Hölder continuous with exponent α. A standard compactness argument shows that $Y(\omega)$ is Hölder continuous on D. This concludes the proof. □

Remark 63. *Note that a more careful argument allows one to show (at least for Gaussian processes) that $Z\colon \omega \mapsto |Y(\omega)|_{C^\alpha(D)}$ is a random variable with $e^Z \in L^q(\Omega)$ for all $1 \leq q < \infty$. To achieve this, one has to replace the Borel-Cantelli Lemma with a quantitative version, i.e., one needs to know how fast $n(\omega)$ and $m(\omega)$ will grow.*

Note that we can apply Theorem 61 to our Gaussian processes with known covariance functions $\varrho(x,y)$. Given $x, y \in D$, note that $Z := X(x) - X(y)$ is a Gaussian random variable with zero mean and variance $\varrho(x,x) - 2\varrho(x,y) + \varrho(y,y)$. Since the higher moments of Gaussian's are explicitly known, we obtain

$$\mathbb{E}\Big(|X(x)-X(y)|^{2p}\Big) = \mathbb{E}Z^{2p} = \Big(\prod_{i=0}^{p-1}(2p-2i)\Big)\Big(\varrho(x,x)-2\varrho(x,y)+\varrho(y,y)\Big)^p.$$

For example, the Brownian motion satisfies $\varrho(s,t) := \min\{s,t\}$ and hence

$$\mathbb{E}\Big(|B(s)-B(t)|^{2p}\Big) \simeq |s-t|^p.$$

Choosing $p \in \mathbb{N}$ sufficiently large, we satisfy $p \geq d + \alpha 2p$ for all $\alpha < 1/2$. Hence, Theorem 61 shows that B is Hölder continuous for all $0 < \alpha < 1/2$. Together with Theorem 53, we finally proved the existence of Brownian motions. (Note that one can show that $\alpha < 1/2$ is indeed the maximal regularity for Brownian motions.)

The exponential covariance $\varrho(x,y) := \exp(-|x-y|/\lambda)$ produces the same Hölder regularity.

For the Gaussian covariance $\varrho(x,y) := \exp(-|x-y|^2/\lambda)$, we have $\mathbb{E}\Big(|X(x)-X(y)|^{2p}\Big) \simeq |x-y|^{2p}$ and hence Hölder regularity for all $0 < \alpha < 1$. In fact, one can show that the derivative $\nabla_x X(x,\omega)$ has the covariance function $\nabla_x \nabla_y \varrho(x,y)$ and is again a Gaussian process. Thus, we may apply Theorem 61 to all derivatives and prove that $X(\omega) \in C^\infty(D)$ almost surely. The different regularity of samples can also be observed in Figure 6.

4. High-dimensional approximation: Neural Networks

The number of elements in a regular mesh $\mathcal{T}_h$ in which each element $T \in \mathcal{T}_h$ satisfies $\operatorname{diam}(T) \simeq |T|^{1/d} \simeq h$ scales roughly like $\mathcal{O}(h^{-d})$, i.e., exponentially in the dimension. We remember the a priori convergence of FEM of (quasi-) interpolation operators of the form

$$\|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(h)$$

in case u is sufficiently smooth. To compute u_h, we need to solve a linear system with $\#\mathcal{T}_h$ elements. The cost for this is at least $\mathcal{O}(\#\mathcal{T}_h) = \mathcal{O}(h^{-d})$. Thus, in terms of cost, we get the estimate

$$\|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(N^{-1/d})$$

with $N = \#\mathcal{T}_h$.

Just as for tensor quadrature, the convergence rate with respect to cost goes down in higher dimensions, the *curse of dimensionality* renders the approach impractical. We consider a couple of many different approaches to this problem and start with *neural networks*. They are extremely versatile and have been shown to posses almost all the approximation characteristics that classical methods (sparse grids, polynomial interpolation,...) enjoy. On the downside, they are mathematically harder to study and often lead to non-linear, non-convex interpolation problems.

4.1. **Definition of Artificial Neural Networks.** Artificial Neural Networks (or just networks or neural networks in the following) are a class of functions $F\colon \mathbb{R}^s \to \mathbb{R}^{s'}$ which can be represented by a number of parameters (also often called *weights*). In that regard, neural networks are no different to the class of polynomial functions, or the class of rational functions.

In the following, we define a certain kind of network known as *feed forward* network (note that there are many other kinds of artificial neural networks, and we only consider a simple class here): Given a depth $d \in \mathbb{N}$ and an architecture $s_0, \dots, s_d \in \mathbb{N}$, we define the weight matrices

$$\boldsymbol{W}_i \in \mathbb{R}^{s_i \times s_{i-1}} \quad \text{for all } i = 1, \dots, d$$

and the biases

$$\boldsymbol{b}_i \in \mathbb{R}^{s_i} \quad \text{for all } i = 1, \dots, d.$$

For a given activation function $\phi\colon \mathbb{R} \to \mathbb{R}$, which is applied entrywise to vectors, we define the network $F\colon \mathbb{R}^{s_0} \to \mathbb{R}^{s_d}$ by $F(x) := F_d(x)$ with $F_0(x) := x$ and

$$F_{i+1}(x) = \phi(\boldsymbol{W}_{i+1} F_i(x) + \boldsymbol{b}_{i+1}) \quad \text{for all } i = 0, \dots, d-2$$

and $F_d(x) := \boldsymbol{W}_d F_{d-1}(x) + \boldsymbol{b}_d$. The activation function is typically a non-linear function (if ϕ is linear, then F is just an affine function). Popular activation functions are

- *ReLU:* $\phi(x) = \max\{x, 0\}$
- *leaky-ReLU:* $\phi(x) = \max\{x, \delta x\}$ for some $0 < \delta \ll 1$.
- *sigmoid:* $\phi(x) = 1/(1 + e^x)$
- *swish:* $\phi(x) = x/(1 + e^{-x})$
- *softplus:* $\phi(x) = \log(1 + e^x)$

Remark 64. *Note that in practical applications, all sorts of (and combinations of) activation functions have proven themselves useful. In the mathematical analysis of neural networks, the choice of ϕ often doesn't make a real difference and one sticks with simple choices such as ReLU.*

We will work only with the ReLU-activation function $\phi(x) = \max\{x, 0\}$. We define the dimensions of a network F by

$$\operatorname{depth}(F) := d \quad \text{and} \quad \operatorname{width}(F) := \max_{0 \le i \le d} s_i$$

as well as $\dim(F) := \max\{\operatorname{depth}(F), \operatorname{width}(F)\}$. Note that the term *depth* is sometimes referred to as the number of layers of a network. A network F with $\operatorname{depth}(F) > 2$ is usually called a *deep* neural network, while $\operatorname{depth}(F) \le 2$ is called a *shallow* network. Similarly, the notion *deep learning* refers to the use of deep neural networks as opposed to shallow networks.

The number of parameters that specify a network F of given architecture $s_0, \dots, s_d$ is

$$\#F := \sum_{i=1}^{d} s_i s_{i-1} + s_i \leq d\,(\dim(F)^2 + \dim(F)).$$

Note that, contrary to polynomial spaces, the space of networks F of a given architecture is not linear and also the dependence of F on the weights $\boldsymbol{W}_1, \dots, \boldsymbol{W}_d$ is non-linear. To clarify this dependence, it is often useful to write

$$F(x) = F(\boldsymbol{W}, \boldsymbol{b}, x),$$

where $\boldsymbol{W} := (\boldsymbol{W}_1, \dots, \boldsymbol{W}_d) \in \mathbb{R}^{\sum_{i=1}^{d} s_i s_{i-1}}$ and $\boldsymbol{b} := (\boldsymbol{b}_1, \dots, \boldsymbol{b}_d) \in \mathbb{R}^{\sum_{i=1}^{d} s_i}$ are interpreted sometimes as matrices and sometimes as vectors for convenience.

4.2. **Gradient descent.** As with interpolation in polynomial spaces, one can try to approximate given data with neural networks. Given $x_1, \dots, x_N \in \mathbb{R}^s$ and $y_1, \dots, y_N \in \mathbb{R}^{s'}$ as well as an architecture $s = s_0, s_1, \dots, s_d = s'$, the approximation problem is to find weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$ such that

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}) := \sum_{i=1}^{N} |F(\boldsymbol{W}, \boldsymbol{b}, x_i) - y_i|^2 \to \min. \tag{15}$$

The function $\mathcal{L}(\boldsymbol{W}, \boldsymbol{b})$ is called the *loss function* (note that there are many other useful definitions of loss, but we will only consider the least-square loss).

Due to the non-linearity of $(\boldsymbol{W}, \boldsymbol{b}) \mapsto F(\boldsymbol{W}, \boldsymbol{b}, \cdot)$, we have to use a non-linear optimization method. One of these methods is *Gradient descent*.

Algorithm 4. *Input: function $f\colon \mathbb{R}^d \to \mathbb{R}$, starting value $w_0 \in \mathbb{R}^d$ and step-size $\alpha > 0$. For $\ell = 0, 1, 2, \dots$ do:*

(1) *Compute gradient $G_\ell := \nabla_w f(w_\ell) \in \mathbb{R}^d$.*
(2) *Update $w_{\ell+1} = w_\ell - \alpha G_\ell$.*

Output: sequence of approximations w_ℓ of the minimizer of f.

Remark 65. *Obviously, the gradient descent algorithm (Algorithm 4) can be applied to minimize $\mathcal{L}(\boldsymbol{W}, \boldsymbol{b})$ by embedding $(\boldsymbol{W}, \boldsymbol{b}) \in \mathbb{R}^{\#F}$ and setting $w_\ell = (\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)$.*

Remark 66. *Note that for non-linear, non-convex optimization, convergence of $(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)$ to the true minimizer is all but certain. This is the reason, why most results about approximation by neural networks fall into one of the following two categories: (1) Results that show that a neural network with certain approximation properties exists and (2) results that show that a certain optimization algorithm will find a network with certain approximation properties. Naturally, category 2 is much harder to prove than category 1.*

Remark 67. *Note that practical implementations of machine learning often use* Stochastic Gradient Descent *instead of plain gradient descent. The only difference to Algorithm 4 is that instead of computing $G_\ell = \nabla_{(\boldsymbol{W}, \boldsymbol{b})} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b})$, one randomly selects $m \ll n$ (the so-called* batch size*) indices $1 \leq i_1, \dots, i_m \leq N$ and computes*

$$G_\ell^{\text{stoch}} := \frac{n}{m} \nabla_{(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)} \sum_{j=1}^{m} |F(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell, x_{i_j}) - y_{i_j}|^2 \approx \nabla_{(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)} \mathcal{L}(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)$$

in Step (1) of Algorithm 4. Often, one uses $m = 32$. The algorithm has several practical advantages such as:

- *more optimization steps for the same cost,*

- *often more efficient as the batch size can be optimized such that G_ℓ^{stoch} can be computed in the fast memory close to the processor,*
- *stochastic nature of G_ℓ^{stoch} can prevent getting stuck at local minima.*

The mathematical analysis of stochastic gradient descent is very similar to plain gradient descent, since one can rely on the fact

$$\mathbb{E}G_\ell^{\text{stoch}} = \nabla_{(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)}\mathcal{L}(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell) = G_\ell.$$

We give one example of a function class which guarantees convergence of gradient descent.

Lemma 68. *Let $Q, q > 0$ and let $f\colon \mathbb{R}^d \to \mathbb{R}$ denote a non-negative function such that $qf(w) \leq |\nabla f(w)|^2 \leq Qf(w)$ for all $w \in \mathbb{R}^d$. Moreover, let*

$$|\nabla f(w) - \nabla f(v)| \leq Q|w-v||w| + \frac{q}{2Q^{1/2}} f(w)^{1/2} \quad \text{for all } w, v \in \mathbb{R}^d.$$

Then, there exists a step-size $\alpha > 0$ and some $0 < \kappa < 1$ such that Algorithm 4 produces a sequence $(w_\ell)_{\ell\in\mathbb{N}}$ with

$$f(w_\ell) \leq \kappa f(w_{\ell-1}) \leq \kappa^\ell f(w_0) \quad \text{for all } \ell \in \mathbb{N}.$$

Proof. There holds

$$\begin{aligned} f(w_{\ell+1}) - f(w_\ell) &= -\alpha \int_0^1 \nabla f(w_\ell + s(w_{\ell+1} - w_\ell)) G_\ell \, ds \\ &= -\alpha \nabla f(w_\ell) \cdot G_\ell - \alpha \int_0^1 \Big(\nabla f(w_\ell + s(w_{\ell+1} - w_\ell)) - \nabla f(w_\ell)\Big) G_\ell \, ds. \end{aligned}$$

Under the assumptions on f, we obtain

$$|\nabla f(w_\ell + s(w_{\ell+1} - w_\ell)) - \nabla f(w_\ell)| \leq Q\alpha |G_\ell||w_\ell| + q/2 f(w_\ell)^{1/2} \leq (Q^{3/2}\alpha|w_\ell| + q/(2Q^{1/2}))f(w_\ell)^{1/2}.$$

This concludes

$$f(w_{\ell+1}) - f(w_\ell) \leq -\alpha q f(w_\ell) + (Q^2\alpha|w_\ell| + q/2)\alpha f(w_\ell). \tag{16}$$

We prove by induction that $|w_\ell| \leq C$ and $f(w_\ell) \leq \kappa f(w_{\ell-1})$ for all $\ell \in \mathbb{N}$ with

$$C := |w_0| + 8Q^{1/2}/q f(w_0)^{1/2} \quad \text{and} \quad \kappa := 1 - q^2/(16Q^2C).$$

For $\ell = 0$, there is nothing to prove. Assume the induction assumption holds for all $0 \leq \ell \leq L$ and choose $\alpha > 0$ such that the reduction factor in (16)

$$1 - \alpha q/2 + Q^2 C \alpha^2 < 1$$

is minimal. Elementary optimization reveals that $\alpha := (q/4)/(Q^2C)$ is the optimal choice and the minimum is $\kappa = 1 - q^2/(16Q^2C)$. Then, (16) implies

$$f(w_{\ell+1}) \leq \kappa f(w_\ell) \quad \text{for all } 0 \leq \ell \leq L$$

and hence also

$$|w_{\ell+1} - w_\ell| = \alpha|G_\ell| \leq Q^{1/2}\alpha f(w_\ell)^{1/2} \leq Q^{1/2}\alpha\kappa^{\ell/2} f(w_0)$$

for all $0 \leq \ell \leq L$. This shows $|w_{\ell+1}| \leq |w_0| + Q^{1/2}\alpha f(w_0)^{1/2} \sum_{j=0}^\infty \kappa^{\ell/2} = |w_0| + Q^{1/2}\alpha f(w_0)/(1-\sqrt{\kappa})$. With $\sqrt{1-x} \leq 1 - x/2$, we obtain

$$\sqrt{\kappa} \leq 1 - \frac{q^2}{32Q^2C}$$

and hence

$$Q^{1/2}\alpha f(w_0)^{1/2}/(1-\sqrt{\kappa}) \leq Q^{1/2} f(w_0)^{1/2}\alpha \frac{32Q^2C}{q^2} = C - |w_0|.$$

Hence, we obtain $|w_\ell| \leq C$ for all $0 \leq \ell \leq L+1$. This concludes the induction and thus the proof. □

Remark 69. *Note that, e.g., strongly convex functions with linearly bounded second derivative satisfy the assumptions of the above lemma.*

4.2.1. *Numerical example.* As a practical example, we try to approximate the function $x \mapsto x^2$ on $[0,1]$ by neural networks of different depth. We consider the networks $F^i\colon \mathbb{R} \to \mathbb{R}$ with depth equal to i and $s_0 = s_i = 1$ and $s_j = 5$ for all $1 \leq j \leq i-1$. We use the ReLU activation function $\phi(x) = \max\{x, 0\}$. The loss function is defined as

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}) := \sum_{i=0}^{10^3} (F(\boldsymbol{W}, \boldsymbol{b}, x_i) - x_i^2)^2$$

for $x_i := i10^{-3}$. To solve the optimization problem (15), we use the Python library `TensorFlow 2.0`. In Figures 7–8 you can see the numerical results. Instead of plain gradient descent (Algorithm 4), the code uses a more sophisticated optimizer called *Adam*, which averages the gradient over a number of steps to obtain smoother updates. We include the Python code for running the experiment with F^2:

```
import numpy as np
import tensorflow as tf
# define data
n=1e3;
xdata = np.arange(0,1,1/n)
ydata = xdata**2
# define neural network
model = tf.keras.models.Sequential([tf.keras.layers.Input(shape=(1,)),
            tf.keras.layers.Dense(5,activation='relu'),
            tf.keras.layers.Dense(5,activation='relu'),
            tf.keras.layers.Dense(1)])
# define loss function
def loss(y_actual,y_pred):
  return  tf.reduce_sum(tf.square(y_actual-y_pred))
# set up gradient descent and train the network
model.compile(optimizer='adam',loss=loss)
model.fit(xdata,ydata,epochs=1000)
```

4.3. **Elementary approximation properties.** Note that many elementary function can be represented by a neural network directly. For example, for $\phi(x) := \max\{x, 0\}$ the following functions are neural networks:

- Identity: The identity $\mathrm{id}\colon \mathbb{R}^s \to \mathbb{R}^s$ can be represented by two or three-layer networks with width bounded by $2s$, e.g.,

$$\mathrm{id}(x) = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{I} \end{pmatrix} \phi\Big(\begin{pmatrix} \boldsymbol{I} \\ -\boldsymbol{I} \end{pmatrix} x \Big) = \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{I} \end{pmatrix} \phi\Big(\begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \phi\Big(\begin{pmatrix} \boldsymbol{I} \\ -\boldsymbol{I} \end{pmatrix} x \Big) \Big),$$

  where $\boldsymbol{I}, \boldsymbol{0} \in \mathbb{R}^{s\times s}$ denotes the identity matrix and the zero matrix.

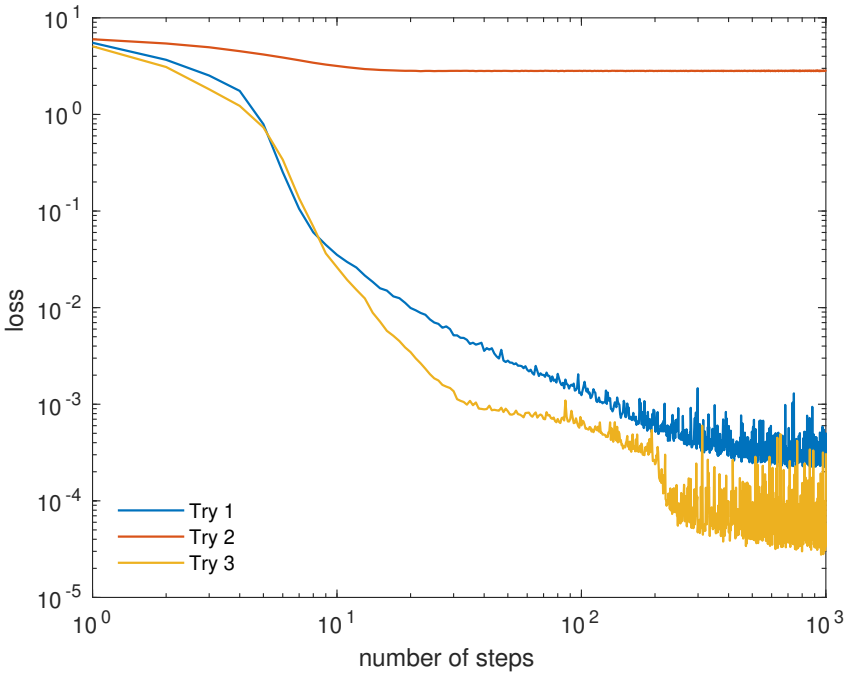

FIGURE 7. We plot three runs of gradient descent with the goal to find optimal weights for $F^5(\boldsymbol{W}, \boldsymbol{b}, x) \approx x^2$. Since the starting guess $(\boldsymbol{W}_0, \boldsymbol{b}_0)$ is random, the performance of gradient descent varies dramatically and sometimes the algorithm even fails to converge. The x-axis shows the number of iterations of gradient descent and the y-axis shows the value of $\mathcal{L}(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)$.

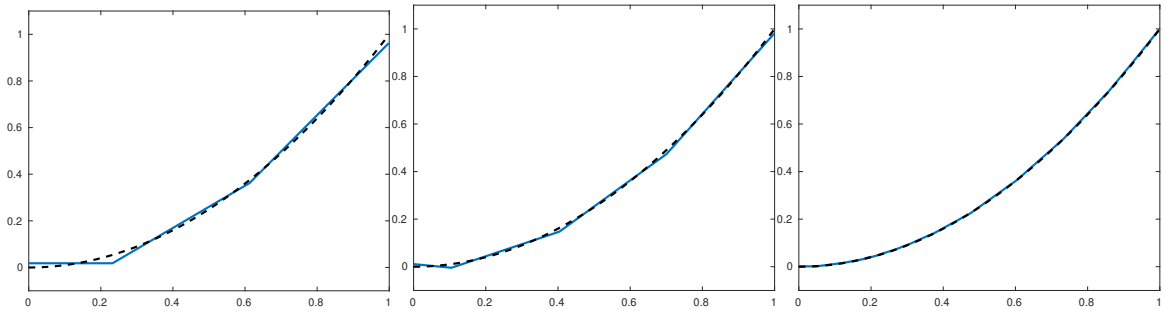

FIGURE 8. We plot (from left to right) the best approximations F^1, F^3, F^5 to $x \mapsto x^2$ which gradient descent finds after 10^3 optimization steps. We can clearly see, that a deep network F^5 achieves a better approximation than a shallow network F^1.

- Maximum: For $x, y \in \mathbb{R}$, there holds

$$\max\{x, y\} = \max\{x - y, 0\} + y = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix} \phi\Big(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \Big).$$

- Minimum: For $x, y \in \mathbb{R}$, there holds

$$\min\{x, y\} = -\max\{-x, -y\} = \begin{pmatrix} -1 & -1 & 1 \end{pmatrix} \phi\Big(\begin{pmatrix} -1 & 1 \\ 0 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \Big).$$

- Absolute value: For $x \in \mathbb{R}$, there holds

$$|x| = \max\{x, 0\} + \max\{-x, 0\} = \begin{pmatrix} 1 & 1 \end{pmatrix} \phi\Big(\begin{pmatrix} 1 \\ -1 \end{pmatrix} x \Big). \tag{17}$$

Given $e_1, \dots, e_s, f_1, \dots, f_s \in \mathbb{R}$, a one-layer network

$$F(x) = (e_1, \dots, e_s)\phi((f_1, \dots, f_s)^T x + (b_1, \dots, b_s)^T) \colon \mathbb{R} \to \mathbb{R}$$

with $\phi(x) := \max\{x, 0\}$ is a piecewise linear function. The kinks of this piecewise linear function are located at

$$x_i := -b_i/f_i \quad \text{for } i = 1, \dots s.$$

Hence, we have for $x \notin \{x_1, \dots, x_s\}$ that

$$F'(x) = \sum_{\substack{i=1 \\ f_i x + b_i \geq 0}}^{s} e_i f_i.$$

This shows that by carefully choosing the weights e_i, f_i and b_i, we can exactly represent any piecewise linear function on a given interval $[a, b] \subset \mathbb{R}$. Since piecewise linear functions are dense in continuous functions, this implies that we can uniformly approximate any continuous function $f \colon [a, b] \to \mathbb{R}$ by a sufficiently wide one-layer neural network. A similar theorem in multiple dimensions was already proven in the 90s.

Theorem 70 (Universal approximation theorem [7])**.** *Let $f \colon K \to \mathbb{R}^{d'}$ be continuous on the compact set $K \subset \mathbb{R}^d$ and $\varepsilon > 0$. If $\phi \colon \mathbb{R} \to \mathbb{R}$ is continuous and not a polynomial, there exists $s \in \mathbb{N}$ and $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^s$ such that*

$$\sup_{x \in K} |f(x) - (\boldsymbol{W}_2^T \phi(\boldsymbol{W}_1 x + \boldsymbol{b}_1) + \boldsymbol{b}_2)| \leq \varepsilon.$$

Remark 71. *Note that the condition of ϕ not being a polynomial is necessary. Otherwise, $F(\cdot)$ is also just a polynomial of the same degree on $\mathbb{R}^d$ and therefore can't approximate continuous functions.*

Remark 72. *The usefulness for numerical algorithms of the theorem above is limited, as it does not quantify the number of parameters necessary to obtain a certain accuracy. It turns out that deep networks can be more efficient by orders of magnitude (see Theorem 74 below). This reminds us of the situation of first-order approximation vs. higher-order approximation. For example, linear approximation (Interpolation, Scott-Zhang) converges with rate $\mathcal{O}(h^2)$ in L^2. Higher-order Cebyshev interpolation, however, can converge with exponential rate e^{-p} if sufficient smoothness is available.*

Lemma 73. *Given two networks $F \colon \mathbb{R}^s \to \mathbb{R}^r$ and $G \colon \mathbb{R}^r \to \mathbb{R}^t$, the composition $G \circ F$ is a network with $\operatorname{depth}(G \circ F) = \operatorname{depth}(F) + \operatorname{depth}(G)$ and $\operatorname{width}(G \circ F) = \max\{\operatorname{width}(G), \operatorname{width}(F)\}$. Given networks $F_1, \dots F_n \colon \mathbb{R}^s \to \mathbb{R}^r$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, there exists a network F with $\operatorname{width}(F) \leq \max\{\sum_{i=1}^n \operatorname{width}(F_i), 2s, r\}$ and $\operatorname{depth}(F) \leq \max_{1 \leq i \leq n}\{\operatorname{depth}(F_i)\} + 2$ such that $F(x_1, \dots, x_n) = (F(x_1), \dots, F(x_n))$ for all $(x_1, \dots, x_n) \in \mathbb{R}^{ns}$. Moreover, there exists a network G of the same dimensions of F such that $G(x) = \sum_{i=1}^n \alpha_i F_i(x)$ for all $x \in \mathbb{R}^s$.*

Proof. For the proof, we assume vanishing biases, i.e. $\boldsymbol{b} = 0$, for all the involved networks in order to simplify the notation.

Part 1) Composition: Assume that the architecture of F is $s = s_0, \dots, s_d = r$ with weight matrices $\boldsymbol{W}_i^F$ and that of G is $r = s'_0, \dots, s'_{d'} = t$ with matrices $\boldsymbol{W}_i^G$. Then, we define $G \circ F$ with the architecture

$$s_0, \dots, s_d, s'_0, \dots, s'_{d'}$$

and weight matrices $\boldsymbol{W}_i := \boldsymbol{W}_i^F$ for all $i = 1, \dots, d$ and $\boldsymbol{W}_i := \boldsymbol{W}_{i-d}^G$ for all $i = d+1, \dots d+d'$. The dimensions of $G \circ F$ can be derived directly from this construction.

Part 2) Weighted sum and vectorization: To sum two networks of different depth, we first have to bring them to equal length, using the identity blocks id. Given networks $F_1, \dots F_n \colon \mathbb{R}^s \to \mathbb{R}^r$ with depths $d_1, \dots, d_n \in \mathbb{N}$, we can use id to obtain networks $\widetilde{F}_i$ with depths $d := 2 + \max_{1 \le i \le n} d_i$ by writing $d - d_i = 2k + 3r$ with $k \in \mathbb{N}$ and $r \in \{0, 1\}$ and composing

$$\widetilde{F}_i := F_i \circ \mathrm{id}_2^{(k)} \circ \mathrm{id}_3^{(r)},$$

where id_i, $i \in \{2, 3\}$ denotes the two-layer and three-layer version of the identity network. Note that $\mathrm{width}(\widetilde{F}_i) \le \max\{\mathrm{width}(F_i), 2s\}$. Given the weight-matrices $\widetilde{\boldsymbol{W}}_{i,j} \in \mathbb{R}^{s_{i,j} \times s_{i,j-1}}$, $j = 1, \dots, d$ of the networks $\widetilde{F}_i$, we construct $\boldsymbol{W}_j \in \mathbb{R}^{\sum_{i=1}^n s_{i,j} \times \sum_{i=1}^n s_{i,j-1}}$ as a block-diagonal matrix, i.e.,

$$\boldsymbol{W}_j := \mathrm{diag}(\boldsymbol{W}_{1,j}, \dots, \boldsymbol{W}_{n,j})$$

for $j = 1, \dots, d$ Finally, we define

$$\boldsymbol{V}_s := \begin{pmatrix} \boldsymbol{I} & \dots & \boldsymbol{I} \end{pmatrix} \in \mathbb{R}^{s \times sn},$$
$$\boldsymbol{R}_r := \begin{pmatrix} \alpha_1 \boldsymbol{I} & \dots & \alpha_n \boldsymbol{I} \end{pmatrix} \in \mathbb{R}^{r \times rn},$$

The construction above shows that the networks $F := \boldsymbol{W}_d \phi(\dots \phi(\boldsymbol{W}_1 (x_1, \dots, x_n)^T) \dots)$ as well as $G := \boldsymbol{R}_r \boldsymbol{W}_d \phi(\dots \phi(\boldsymbol{W}_1 \boldsymbol{V}_s^T x) \dots)$ satisfy the statement. This concludes the proof. □

A fundamental property of deep neural networks, as opposed to shallow networks, is the fact that they can efficiently approximate the $x \mapsto x^2$ function. Despite the simple proof, this was discovered only quite recently in [11].

Theorem 74. *There exists a neural network F with* $\mathrm{depth}(F) = 2k$ *and* $\mathrm{width}(F) = 3k + 2$ *such that*

$$|F(x) - x^2| \le 4^{-k} \quad \textit{for all } 0 \le x \le 1.$$

The magnitude of the weights of F is bounded by four.

Proof. Part 1: Define the saw-tooth function (see Figure 9)

$$g(x) := \begin{cases} 2x & x < 1/2, \\ 2 - 2x & 1/2 \le x. \end{cases}$$

It is easy to see that

$$g(x) = \min\{2x, 2 - 2x\} = -\max\{2 - 4x, 0\} + 2 - 2x$$

and hence g can be represented exactly by the two layer network

$$g(x) = \begin{pmatrix} 1 & -1 & -1 \end{pmatrix} \phi\Big(\begin{pmatrix} 2 \\ -2 \\ -4 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \Big) + 2.$$

Let $g^{(j)} := g \circ \ldots \circ g$ denote the j-times composition of g with itself. Define the function

$$G_k(x) := \sum_{j=1}^{k} 4^{-j} g^{(j)}(x).$$

We claim that G_k is the piecewise linear interpolation of $f(x) := x - x^2$ on $[0,1]$ on the points $x = i2^{-k}$ for all $i = 0, \ldots, 2^k$, i.e.,

$$G_k(x) = x - x^2 \quad \text{for all } x = i2^{-k},\ i = 0, \ldots, 2^k. \tag{18}$$

See Figure 9 for some illustration.

Part 2: To show (18), we first prove that

$$\begin{aligned} g^{(k)}(i2^{-k}) &= 1 \quad \text{for all } i = 1, 3, 5, \ldots, 2^k - 1, \\ g^{(k)}(i2^{-k}) &= 0 \quad \text{for all } i = 0, 2, 4, \ldots, 2^k, \end{aligned} \tag{19}$$

and that $g^{(k)}|_{[i2^{-k},(i+1)2^{-k}]}$ is an affine function for all $i = 0, \ldots, 2^k - 1$. To that end, write $x \in [0,1)$ as $x = \sum_{r=1}^{k} a_r 2^{-r}$ with $a_r \in \{0,1\}$ for $r = 1, \ldots, k-1$ and $0 \le a_k < 2$. Note that $a_k \in \{0,1\}$ implies $x = i2^{-k}$ with even resp. odd i. Moreover, $a_k \in (0,1)$ represents all $x \in (i2^{-k}, (i+1)2^{-k})$ and $a_k \in (1,2)$ represents all $x \in ((i+1)2^{-k}, (i+2)2^{-k})$.

By definition of g, there holds

$$\begin{aligned} g(x) &= \begin{cases} \sum_{r=1}^{k} a_r 2^{-r+1} & a_1 = 0, \\ 2 - \sum_{r=1}^{k} a_r 2^{-r+1} & a_1 = 1 \end{cases} \\ &= \begin{cases} \sum_{r=1}^{k-1} a_{r+1} 2^{-r} & a_1 = 0, \\ 1 - \sum_{r=1}^{k-1} a_{r+1} 2^{-r} & a_1 = 1 \end{cases} \end{aligned}$$

and hence, with $g(x) = g(1-x)$, we have

$$g(g(x)) = g\Big(\sum_{r=1}^{k-1} a_{r+1} 2^{-r} \Big). \tag{20}$$

This immediately proves (19) for all $0 \le i < 2^k$, since from (20), we obtain $g^{(k)}(x) = g(a_k 2^{-1})$ and hence

$$g^{(k)}(x) = g(a_k 2^{-1}) = \begin{cases} g(1/2) = 1 & a_k = 1, \\ g(0) = 0 & a_k = 0, \\ g|_{(0,1/2)\cup(1/2,1)} = \text{piecewise affine function} & a_k \in (0,1) \cup (1,2). \end{cases}$$

For $i = 2^k$, we have $x = 1$ and a direct calculation shows $g^{(k)}(1) = 0$. This concludes (19).

Part 3: With the results from Part 2 at hand, we prove (18) by induction on k. First, note that (19) implies that G_k is a piecewise affine function on $(i2^{-k}, (i+1)2^{-k})$ for $i = 0, \ldots, 2^k - 1$. For $k = 1$, there holds $G_1(0) = 0$, $G_1(1) = 0$, and $G_1(1/2) = 1/4$ which interpolates $f(x) := x - x^2$. Moreover, G_k is linear on $[0, 1/2]$ and $[1/2, 1]$. Assume the G_k interpolates $f(x)$ in the respective points given in (18). Then, we know that $(f - G_k)|_{[i2^{-k},(i+1)2^{-k}]}$ is a parabola which is zero at its endpoints. Thus, the height of the parabola at the interval midpoint $m := i2^{-k} + 2^{-k-1}$ is uniquely determined by its second derivative (which is -2) and the distance of the endpoints (which is 2^{-k}). Elementary calculations show that $(f - G_k)(m) = 2^{-2k}/4 = 4^{-k-1}$. In (19), we proved that G_{k+1} is an affine function on $(i2^{-k}, m)$ and $(m, (i+1)2^{-k})$. Hence, in order to show $G_{k+1}(m) = f(m)$, it remains to confirm that $g^{(k+1)}(m) = 1$ and $g^{(k+1)}(i2^{-k}) = g^{(k+1)}((i+1)2^{-k}) = 0$. This

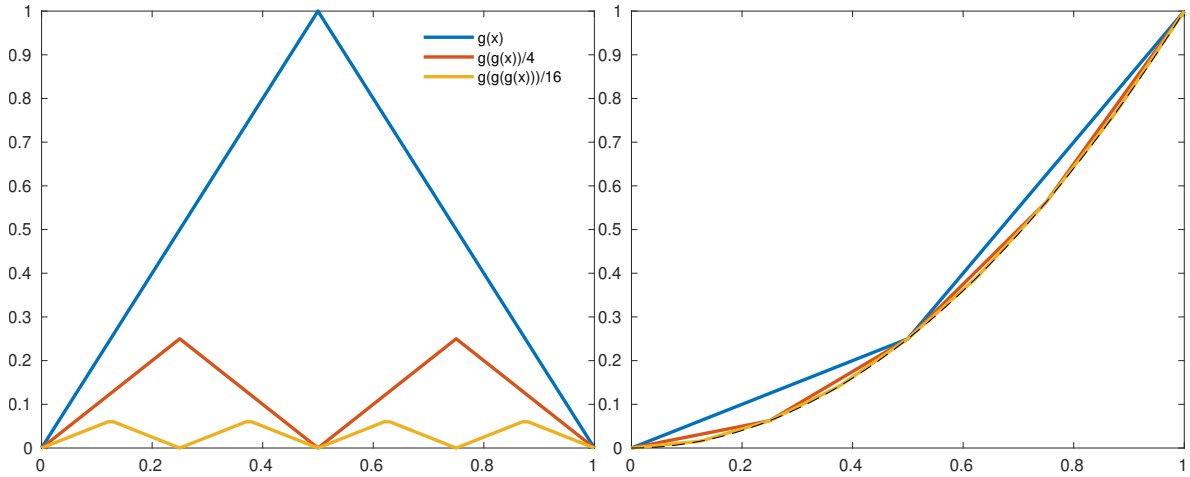

FIGURE 9. Scaled compositions of the sawtooth function $g(x)$ defined in the proof of Theorem 74 (left) and nodal interpolation of x^2 (right).

follows from (19), since $m = (2i+1)2^{-k-1}$. This concludes the induction and hence the proof of (18).

Part 4: The approximation error for the 1D linear interpolant on mesh-size 2^{-k} can be bounded by use of the fundamental theorem of calculus, i.e.,

$$\|f - G_k\|_{L^\infty([0,1])} \leq \frac{1}{2} 2^{-2k} \|f''\|_{L^\infty([0,1])} = 4^{-k}.$$

Lemma 73 shows that G_k can be represented by a neural network with depth $2k+2$ and width $3k$. Examining the proof of Lemma 73, we see that, since the $g^{(k)}$ have an even number of layers, we can actually construct G_k with depth $2k$ instead of $2k+2$. Another application of Lemma 73 shows that $x - G_k(x)$ is a neural network with depth $2k$ and width $3k+2$. This concludes the proof. □

Corollary 75. *Given $k, M \in \mathbb{N}$, there exists a network G with depth $2k+5$ and width $9k+6$ such that*

$$|xy - G(x,y)| \leq 6M^2 4^{-k} \quad \textit{for all } -M \leq x, y \leq M.$$

Proof. First, we note that $xy = 4M^2(x/(2M))(y/(2M))$. Hence, we may restrict ourselves to numbers $x, y \in [-1/2, 1/2]$. We already know how to square numbers in $[0,1]$ using the network F from Theorem 74 together with the absolute value from (17), there holds $\widetilde{F}(x) := (F \circ |\cdot|)(x) = F(|x|)$ for all $x \in [-1,1]$.

Therefore, we compute for all $x, y \in [-1/2, 1/2]$

$$\begin{aligned} 2xy &= ((x+y)^2 - x^2 - y^2) \\ &\approx \widetilde{G}(x,y) := \big(\widetilde{F}(x+y) - \widetilde{F}(x) - \widetilde{F}(y)\big) \end{aligned}$$

Lemma 73 shows that $x+y$ is a network with depth 2 and width 4 (it is easy to construct $x+y$ with width two directly). Hence, with Theorem 74, $\widetilde{F}(x+y)$ is a network with depth $2k+3$ and width $3k+2$, while $\widetilde{F}$ is a network with depth $2k+1$ and width $3k+2$. Thus, Lemma 73 implies that the right-hand side $\widetilde{G}(x,y)$ is a neural network with depth $2k+5$ and width $9k+6$. Finally, let $\widetilde{\boldsymbol{W}}_0, \ldots, \widetilde{\boldsymbol{W}}_{2k+5}$ denote the weight matrices of $\widetilde{G}$. The division by M and the multiplication by $2M^2$ can be achieved by setting the

weight matrices of G to $\boldsymbol{W}_1 = \widetilde{\boldsymbol{W}}_1/(2M)$, $\boldsymbol{W}_{2k+5} = 2M^2\widetilde{\boldsymbol{W}}_{2k+5}$ and $\boldsymbol{W}_i = \widetilde{\boldsymbol{W}}_i$ for all $i = 2, \dots, 2k+4$.

The approximation error satisfies for $x, y \in [-1/2, 1/2]$, by use of the triangle inequality and Theorem 74,

$$|2xy - \widetilde{G}(x,y)| \leq |(x+y)^2 - F(|x+y|)| + |x^2 - F(|x|)| + |y^2 - F(|y|)| \leq 3 \cdot 4^{-k}.$$

This shows for $x, y \in [-M, M]$ that

$$|xy - G(x,y)| \leq 6M^2 4^{-k}$$

and concludes the proof. □

Corollary 76. *Given a monomial $\prod_{i=1}^d x_i^{q_i}$, there exists a network $F\colon \mathbb{R}^d \to \mathbb{R}$ with $\operatorname{depth}(F) \simeq (d+q)k$ and $\operatorname{width}(F) \simeq dk$ such that*

$$|\prod_{i=1}^d x_i^{q_i} - F(x_1, \dots, x_d)| \leq C 4^{-k} (2M)^{2d-q}$$

for all $x_1, \dots, x_d$ with $|x_i| \leq M^{1/(dq)}$ and $q := \max_{1 \leq i \leq d} q_i$. The hidden constants depend only on M.

Proof. For simplicity, we assume $M \geq 3/2$.

Step 1: We use the multiplication network G from Corollary 75 with $2M$ instead of M and $k \in \mathbb{N}$ sufficiently large such that $6 \cdot 4^{-k}(2M)^{(2d+1)q} \leq M$. The approximation is constructed via

$$x^q \approx R_q(x) := \underbrace{G(x, G(x, \dots G}_{q-1\text{-times}}(x,x)\dots)).$$

We show that the approximation error satisfies

$$|x^q - R_q(x)| \leq 6 \cdot 4^{-k}(2M)^q \tag{21}$$

by induction on q. For $q = 2$, Corollary 75 implies $|x^2 - G(x,x)| \leq 6(2M)^2 4^{-k}$ and hence confirms (21). Assume that (21) holds for some $q \in \mathbb{N}$. Then, since $|R_q(x)| \leq |x^q| + 6 \cdot 4^{-k}(2M)^q \leq 2M$, Corollary 75 shows

$$\begin{aligned} |x^{q+1} - R_{q+1}(x)| &\leq |x^{q+1} - xR_q(x)| + |xR_q(x) - G(x, R_q(x))| \\ &\leq M^{1/q} 6 \cdot 4^{-k}(2M)^q + 6(2M)^2 4^{-k} \leq 6 \cdot 4^{-k}(2M)^{q+1}, \end{aligned}$$

since $2M \geq M^{1/q} + 1$ for $M \geq 1$. This concludes (21).

Step 2: We may construct $R_q(x)$ with the following helper network

$$Q(x,y) := (x, G(x,y)) \in \mathbb{R}^2$$

with $\operatorname{depth}(Q) \simeq \operatorname{width}(Q) \simeq k$. Then, with $\boldsymbol{e}_2 := (0,1)$, there holds $R_2 = \boldsymbol{e}_2 \cdot Q(x,x)$ and

$$R_q(x) = \boldsymbol{e}_2 \cdot \underbrace{Q \circ \dots \circ Q}_{q-1-\text{times}}(x,x).$$

Lemma 73 shows $\operatorname{depth}(R_q) \simeq (q-1)k$ and $\operatorname{width}(R_q) \simeq k$.

Step 3: We construct the network $F_{j,\dots,d}$ for $1 \leq j \leq d$ by

$$F_{j,\dots,d}(x_j, \dots, x_d) \begin{cases} G(R_{q_j}(x_j), G(R_{q_{j+1}}(x_{j+1}), \dots G(R_{q_{d-1}}(x_{d-1}), R_{q_d}(x_d))\dots)) & \text{for } j < d \\ R_{q_d}(x_d) & \text{for } j = d. \end{cases}$$

We prove by induction that $|F_{j,\dots,d}(x_j,\dots,x_d)| \leq 2M$ for all $1 \leq j \leq d$ as well as

$$\mathrm{err}_j := |\prod_{i=j}^{d} x_i^{q_i} - F_{j,\dots,d}(x_j,\dots,x_d)| \leq 6 \cdot 4^{-k}(2M)^{q+2(d-j)q}. \tag{22}$$

For $j = d$, $F_d(x_d) = R_{q_d}(x_d)$ and (21) shows $|x_d^{q_d} - F_d(x_d)| \leq 6(2M)^{q_d}4^{-k}$. This implies $|F_d(x_d)| \leq |x_d|^{q_d} + 6(2M)^{q_d}4^{-k} \leq 2M$.

Assume that (22) holds for some $1 < j+1 \leq d$. The error estimates from Corollary 75 and (21) show

$$\begin{aligned}
\mathrm{err}_j &\leq |\prod_{i=j}^{d} x_i^{q_i} - x_j^{q_j} F_{j+1,\dots,d}(x_{j+1},\dots,x_d)| \\
&\quad + |x_j^{q_j} - R_{q_j}(x_j)||F_{j+1,\dots,d}(x_{j+1},\dots,x_d)| \\
&\quad + |R_{q_j}(x_j)F_{j+1,\dots,d}(x_{j+1},\dots,x_d) - G(R_{q_j}(x_j), F_{j+1,\dots,d}(x_{j+1},\dots,x_d))| \\
&\leq 2M\mathrm{err}_{j+1} + 2M6 \cdot 4^{-k}(2M)^{q_j} + 6(2M)^2 4^{-k} \\
&\leq 6 \cdot 4^{-k}(2M)^{q+(d-j)q} \underbrace{\left((2M)^{-2q} + (2M)^{q_j+1-q-2(d-j)q} + (2M)^{2-q-2(d-j)q}\right)}_{\leq 1}
\end{aligned}$$

since $q \geq 1$ and $2M \geq 3$. Moreover, there holds

$$|F_{j,\dots,d}(x_j,\dots,x_d)| \leq \prod_{i=j}^{d} |x_i|^{q_i} + 6 \cdot 4^{-k}(2M)^{(2d+1)q} \leq 2M.$$

This concludes the induction and proves (22).

Step 4: We may construct $F := F_{1,\dots,d}$ with the helper network

$$Q_j(x_1,\dots,x_d) := (x_1,\dots,x_{j-1}, G(x_j,x_{j+1}), x_{j+1},\dots,x_d) \quad 1 \leq j < d$$

with $\mathrm{depth}(Q_j) \simeq k$ and $\mathrm{width}(Q_j) \simeq k+d$ according to Lemma 73. There holds

$$F(x_1,\dots,x_d) = \boldsymbol{e}_1 Q_1 \circ \dots \circ Q_{d-2} \circ Q_{d-1}(R_{q_1}(x_1), R_{q_2}(x_2),\dots,R_{q_d}(x_d))$$

with $\boldsymbol{e}_1 := (1,0,\dots,0) \in \mathbb{R}^d$. From this, we obtain immediately that $\mathrm{depth}(F) \simeq (d+q)k$ and $\mathrm{width}(F) \simeq dk$. This concludes the proof. □

4.4. **Approximation of holomorphic functions.** This section is based on [8]. High dimensional functions with sufficient smoothness can be approximated efficiently by neural networks.

Lemma 77. *Let $F\colon \mathbb{C}^s \to \mathbb{C}$ denote a function that is holomorphic in each component and uniformly bounded on the domain $\Omega' := \prod_{i=1}^{s} B_{r_i}(0) \subset \mathbb{C}^s$ with $r_i > 0$. For any multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^s$ and the corresponding differentiation operator $\partial^{\boldsymbol{\alpha}} := \prod_{i=1}^{s} \partial_{\omega_i}^{\alpha_i}$, there holds*

$$|\partial^{\boldsymbol{\alpha}} F(\mathbf{0})| \leq \|F\|_{L^\infty(\Omega')} \prod_{i=1}^{s} \alpha_i! r_i^{-\alpha_i}.$$

Proof. Just as in the proof of Lemma 31, F satisfies the multidimensional analog of Cauchy's integral formula for all $\boldsymbol{\omega} \in \Omega'$: Choose n-distinct coordinates $\mathfrak{u} := \{d_1,\dots,d_n\} \subseteq \{1,\dots,s\}$, and for $\boldsymbol{z} \in \mathbb{R}^s$ define $(\boldsymbol{z};\mathfrak{u}) \in \mathbb{R}^s$ via

$$(\boldsymbol{z};\mathfrak{u})_i = \begin{cases} \omega_i & i \notin \{d_1,\dots,d_n\} \\ z_i & \text{else.} \end{cases}$$

Then, there holds for all $\widetilde{\boldsymbol{\omega}}$ sufficiently close to $\boldsymbol{\omega}$ that

$$F(\widetilde{\boldsymbol{\omega}}) = (2\pi i)^{-n} \int_{\partial B_{\varepsilon_1}(\omega_{d_1})} \cdots \int_{\partial B_{\varepsilon_n}(\omega_{d_n})} \frac{F(\boldsymbol{z};\mathfrak{u})}{(z_1 - \widetilde{\omega}_{d_1}) \dots (z_n - \widetilde{\omega}_{d_n})} \, dz_1 \dots dz_n,$$

where the parameters $\varepsilon_i > 0$, $i = 1, \dots, n$ are chosen sufficiently small such that the integration domains of the contour integrals above are contained in Ω'. Choosing $\varepsilon_j = r_j$ and $\boldsymbol{\omega} = \mathbf{0}$, we ensure that the contour integrals are contained in Ω'. Differentiation with respect to $\widetilde{\boldsymbol{\omega}}$

$$\partial^{\alpha}_{\widetilde{\omega}_{d_i}} \frac{1}{(z - \widetilde{\omega}_{d_i})} = \frac{\alpha!}{(z - \widetilde{\omega}_{d_i})^{1+\alpha}}$$

and setting $\widetilde{\boldsymbol{\omega}} = 0$ concludes

$$|\partial^{\boldsymbol{\alpha}} F(\mathbf{0})| \leq \|F\|_{L^\infty(\Omega')} \prod_{i=1}^{s} \alpha_i! r_i^{-\alpha_i}.$$

□

Lemma 78. *Let $(\varrho_i)_{i\in\mathbb{N}}$ denote a positive sequence and let $F\colon \mathbb{C}^s \to \mathbb{C}$ denote a function that is holomorphic in each component and uniformly bounded on the domain $\Omega' := \prod_{i=1}^{s} B_{1/2+\varrho_i}(0) \subset \mathbb{C}^s$. Let $0 < p < 1$ such that $\sum_{i=1}^{\infty} \varrho_i^{-p} < \infty$. Then, there exists a network F_n with $\|F - F_n\|_{L^\infty([-1/2,1/2]^s)} \leq C n^{1-1/p}$ such that $\operatorname{depth}(F_n) \simeq s \log(n)$ and $\operatorname{width}(F_n) \simeq n \log(n) + sn$*

Proof. For brevity of notation, we define $r_i := 1/2 + \varrho_i$ and $\rho_i := 1 + 2\varrho_i = 2r_i$. Since each holomorphic function has a convergent Taylor series, we obtain for $\boldsymbol{\omega} \in \Omega'$

$$F(\boldsymbol{\omega}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^s} \frac{\boldsymbol{\omega}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \partial^{\boldsymbol{\alpha}} F(\mathbf{0}),$$

where $\boldsymbol{\omega}^{\boldsymbol{\alpha}} := \prod_{i=1}^{s} \omega_i^{\alpha_i}$ and $\boldsymbol{\alpha}! := \prod_{i=1}^{s} \alpha_i!$. We order the $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots \in \mathbb{N}_0^s$ such that

$$\boldsymbol{\rho}^{-\boldsymbol{\alpha}_k} := \prod_{i=1}^{s} \rho_i^{-\alpha_{k,i}} \geq \boldsymbol{\rho}^{-\boldsymbol{\alpha}_{k+1}} := \prod_{i=1}^{s} \rho_i^{-\alpha_{k+1,i}}$$

for all $k \in \mathbb{N}$. By definition, if $\alpha_i \leq \alpha_k$, we also have $\boldsymbol{\rho}^{-\boldsymbol{\alpha}_i} \geq \boldsymbol{\rho}^{-\boldsymbol{\alpha}_k}$. Note that

$$\sum_{k=1}^{\infty} (\boldsymbol{\rho}^{-\boldsymbol{\alpha}_k})^p = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^s} \boldsymbol{\rho}^{-p\boldsymbol{\alpha}} = \prod_{i=1}^{s} \sum_{\alpha=0}^{\infty} \rho_i^{-p\alpha} = \prod_{i=1}^{s} \frac{1}{1 - \rho_i^{-p}}.$$

Taking the logarithm of the last term above, we see

$$\log\Big(\prod_{i=1}^{s} \frac{1}{1 - \rho_i^{-p}}\Big) = -\sum_{i=1}^{s} \log(1 - \rho_i^{-p}) \lesssim \sum_{i=1}^{s} \rho_i^{-p} < \infty$$

independently of s, where we used that there exists $\delta > 0$ with $\rho_i \geq 1 + \delta > 0$ for all $i \in \mathbb{N}$ in the second to last estimate. This shows that $(\boldsymbol{\rho}^{-\boldsymbol{\alpha}_k})_{k\in\mathbb{N}} \in \ell_p$. Just as in (9), we use this fact together with the fact that the sequence is decreasing to obtain

$$\sum_{k=n+1}^{\infty} \boldsymbol{\rho}^{-\boldsymbol{\alpha}_k} \lesssim n^{1-1/p}.$$

With this in mind, we define the approximation

$$F_n(\boldsymbol{\omega}) := \sum_{k=1}^{n} \frac{\boldsymbol{\omega}^{\boldsymbol{\alpha}_k}}{\boldsymbol{\alpha}_k!} \partial^{\boldsymbol{\alpha}_k} F(\mathbf{0}).$$

Recall that $\boldsymbol{\rho}^{\boldsymbol{\alpha}} = 2^{|\boldsymbol{\alpha}|}\boldsymbol{r}^{\boldsymbol{\alpha}}$. Lemma 77 shows for $\boldsymbol{\omega} \in [-1/2, 1/2]^s$ that

$$\begin{aligned}|F(\boldsymbol{\omega}) - F_n(\boldsymbol{\omega})| &\leq \sum_{k=n+1}^{\infty} \frac{\boldsymbol{\omega}^{\boldsymbol{\alpha}_k}}{\boldsymbol{\alpha}_k!}|\partial^{\boldsymbol{\alpha}_k} F(\mathbf{0})| \leq \|F\|_{L^\infty(\Omega')} \sum_{k=n+1}^{\infty} \frac{|\boldsymbol{\omega}^{\boldsymbol{\alpha}_k}|}{\boldsymbol{\alpha}_k!}\Big(\prod_{i=1}^{s} \alpha_{k,i}!\Big)\boldsymbol{r}^{-\boldsymbol{\alpha}_k} \\ &\leq \|F\|_{L^\infty(\Omega')} \sum_{k=n+1}^{\infty} \frac{|\boldsymbol{\omega}^{\boldsymbol{\alpha}_k}|2^{|\boldsymbol{\alpha}_k|}}{\boldsymbol{\alpha}_k!}\Big(\prod_{i=1}^{s} \alpha_{k,i}!\Big)\boldsymbol{\rho}^{-\boldsymbol{\alpha}_k} \\ &\leq \|F\|_{L^\infty(\Omega')} \sum_{k=n+1}^{\infty} \boldsymbol{\rho}^{-\boldsymbol{\alpha}_k} \lesssim n^{1-1/p}\|F\|_{L^\infty(\Omega')}.\end{aligned}$$

It remains to show that F_n can be approximated by a neural network. To the end, we use Corollary 76 to find a network G_k with

$$\|G_k - \boldsymbol{\omega}^{\boldsymbol{\alpha}_k}\|_{L^\infty([-1,1]^s)} \leq n^{-1/p}$$

such that $\operatorname{depth}(G_k) \simeq (s+q)|\log(n)|$ and $\operatorname{width}(G_k) \simeq \log(n)s$, where $q = \max \boldsymbol{\alpha}_k$. Note that $\max \boldsymbol{\alpha}_k \leq k$.

Lemma 73 shows that

$$\widetilde{F}_n := \sum_{k=1}^{n} \frac{G_k(\boldsymbol{\omega})}{|\boldsymbol{\alpha}_k|!}\partial^{\boldsymbol{\alpha}_k} F(\mathbf{0}).$$

is a network with $\operatorname{depth}(\widetilde{F}_n) \simeq (s+n)\log(n)$ and $\operatorname{width}(\widetilde{F}_n) \simeq n\log(n) + sn$. Finally, the error between F_n and $\widetilde{F}_n$ satisfies

$$|F_n - \widetilde{F}_n|(\boldsymbol{\omega}) \leq nn^{-1/p}.$$

This concludes the proof □

Remark 79. *Note that we can immediately apply the previous lemma to our random Poisson problem from Section 2. We proved in Lemma 29 and Lemma 32 that $F\colon \boldsymbol{\omega} \mapsto G(u(\boldsymbol{\omega}))$ is a holomorphic function that satisfies the assumptions of Lemma 78. This implies, that we can approximate $G(u(\omega))$ by a neural network without curse of dimensionality. Note that this problem is much harder, than just to compute $\mathbb{E}(G(u))$, as we approximate the whole distribution of $G(u)$, which encodes far more information than just its mean.*

Remark 80. *The method of proof in Lemma 78 uses the truncated Taylor expansion of the holomorphic function F to obtain a polynomial approximation. In the context of random processes F which are parametrized in $\boldsymbol{\omega}$, this is called the* Polynomial Chaos *(PC) approximation.*

4.5. **Approximation of solutions of PDEs.** This section is based on [6]. We already saw that neural networks are good at approximating smooth high-dimensional functions. However, even certain non-smooth functions can be approximated efficiently, as shown in the following. We consider the equation

$$-\frac{1}{2}\Delta u(x) + b(x)\cdot\nabla u(x) + c(x)u(x) = f(x) \quad \text{for } x \in \mathbb{R}^d \tag{23}$$

where we assume that c and b are continuous. Lets recall the Feynman-Kac formula

Theorem 81. *Let $u \in C^2(\mathbb{R}^d)$ denote a solution of (23) with $c(x) \geq c_0 > 0$ and let $X(t) \in \mathbb{R}^d$ denote a solution of the (d-dimensional) SDE*

$$dX(t) = -b(X(t))dt + dB(t),\ t \geq 0$$

with $X(0) = x \in \mathbb{R}^d$ almost surely. Then, there holds

$$u(x) = \mathbb{E}\Big(\int_0^\infty \exp\Big(- \int_0^t c(X(s))\,ds\Big) f(X(t))\,dt\Big).$$

As classical numerical tool to solve SDEs, we have the Euler-Maruyama method for equations of the form

$$dX = F(X(t),t)\,dt + G(X(t),t)\,dB(t), \quad X(0) = X_0.$$

Assume a sequence of time-steps $0 = t_0 < t_1 < \ldots < t_n = T$ and define the approximation via

$$X_{i+1} = X_i + F(X_i, t_i)\delta t_i + G(X_i, t_i)\delta B_i \tag{24}$$

with $X_i \approx X(t_i)$, $\delta t_i = t_{i+1} - t_i$, and $\delta B_i := B(t_{i+1}) - B(t_i) \sim N(0, \delta t_i)$.

We have the following classical strong convergence result.

Lemma 82. *Let F and G be Lipschitz continuous in both arguments and let $\sup_{0\le t\le T} \mathbb{E}|X(t)|^2 < \infty$. Then, the Euler-Maruyama scheme converges with strong order $1/2$, i.e.,*

$$\sqrt{\mathbb{E}|X(t_i) - X_i|^2} \le Ce^{CT}\big(\max_{j=1,\ldots,n} \delta t_j\big)^{1/2}$$

for $i = 1, \ldots, n$ and $C > 0$ which depends only on F, G, and T.

Lemma 83. *Assume that the coefficient functions $x \mapsto F(x, t_i)$ and $x \mapsto G(x, t_i)$ from (24) can be represented exactly by neural networks F_i and G_i on a sequence of time-steps $t_0 < t_1 < \ldots < t_n$ such that*

$$\max\{\dim(G_i), \dim(F_i)\} \le m \in \mathbb{N} \quad \text{for all } i = 0, \ldots, n.$$

Then, for almost all $\omega \in \Omega$, the Euler-Maruyama approximation $X_0 \mapsto (X_i(\omega))_{i=0,\ldots,n}$ can be represented by a sequence of neural networks $X_0 \mapsto R_{i,\omega}(X_0)$ with

$$\operatorname{depth}(R_{i,\omega}) \le (\max\{2, m\} + 2)i + 2 \quad \text{and} \quad \operatorname{width}(R_{i,\omega}) \le 2m + 2d.$$

Proof. Fix $\omega \in \Omega$. The construction is inductive. The map $X_0 \mapsto X_0$ can be represented by the identity network $R_0 := \mathrm{id}$ with two layers and width equal to $2d$. Assume that $X_i(\omega) = R_{i,\omega}(X_0)$. Then, we have with (24)

$$\begin{aligned} X_{i+1}(\omega) &= R_{i,\omega}(X_0) + F_i(R_{i,\omega}(X_0))\delta t_i + G_i(R_{i,\omega}(X_0))\delta B_i(\omega) \\ &= (\mathrm{id}_{\mathbb{R}^d\to\mathbb{R}^d} + F_i\delta t_i + G_i\delta B_i(\omega)) \circ R_{i,\omega}(X_0). \end{aligned}$$

We have with Lemma 73 that (note that $\delta B_i(\omega) \in \mathbb{R}$ is just a number)

$$\begin{aligned} &\operatorname{width}(\mathrm{id}_{\mathbb{R}^d\to\mathbb{R}^d} + F_i\delta t_i + G_i\delta B_i(\omega)) \le 2m + 2d \quad \text{and} \\ &\operatorname{depth}(\mathrm{id}_{\mathbb{R}^d\to\mathbb{R}^d} + F_i\delta t_i + G_i\delta B_i(\omega)) \le \max\{m, 2\} + 2. \end{aligned}$$

Hence, R_{i+1} exists and

$$\operatorname{depth}(R_{i+1}) \le \max\{m, 2\} + 2 + \operatorname{depth}(R_i). \quad \text{and} \quad \operatorname{width}(R_{i+1}) \le \max\{2m + 2d, \operatorname{width}(R_i)\}.$$

This concludes the proof. □

Since we can't possibly emulate functions on the whole space $\mathbb{R}^d$ with a neural network, we have to cut off the solution u at some point. The following result shows us how to do that. In the proof, we use a simple tail bound for the standard normal distribution, i.e., for $Z \sim \mathcal{N}(0, 1)$ and $x > 0$, there holds

$$\mathbb{P}(Z \ge x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-s^2/2}\,ds \le \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{s}{x} e^{-s^2/2}\,ds = \frac{e^{-x^2/2}}{\sqrt{2\pi}x}. \tag{25}$$

Lemma 84. *Let u denote the solution of (23) with constant $c(x) := c_0 > 0$ and $b \in L^\infty(\mathbb{R}^d)^d$, $f \in L^\infty(\mathbb{R}^d)$ such that f is supported on a ball $B_n(0) \subset \mathbb{R}^d$ with radius $n \in \mathbb{N}$. Then, there holds*

$$|u(x)| \leq Ce^{-\alpha\sqrt{|x|-n}} \quad \text{for all } |x| > n+1$$

and constants $C, \alpha > 0$ which depend only on c_0 and $\|f\|_{L^\infty(\mathbb{R}^d)}$, $\|b\|_{L^\infty(\mathbb{R}^d)}$.

Proof. In the present case, the Feynman-Kac formula simplifies to

$$u(x) = \int_0^\infty e^{-tc_0}\mathbb{E}(f(X(t)))\,dt = \int_0^T e^{-tc_0}\mathbb{E}(f(X(t)))\,dt + \int_T^\infty e^{-tc_0}\mathbb{E}(f(X(t)))\,dt$$

for all $T > 0$. The second term can be estimated by

$$\Big|\int_T^\infty e^{-tc_0}\mathbb{E}(f(X(t)))\,dt\Big| \leq c_0^{-1}e^{-Tc_0}\sup_{t\geq T}|\mathbb{E}(f(X(t))| \leq c_0^{-1}e^{-Tc_0}\|f\|_{L^\infty(\mathbb{R}^d)}.$$

By definition of X, we have

$$X(t) - X(0) = -\int_0^t b(X(t))\,dt + B(t).$$

Let $C > 0$ be a free parameter. There holds for all $0 \leq t \leq T$.

$$\begin{aligned}\mathbb{P}(|X(t) - X(0)| \geq (C+1)T\|b\|_{L^\infty(\mathbb{R}^d)}) &\leq \mathbb{P}(|B(t)|/\sqrt{t} \geq C\sqrt{T}\|b\|_{L^\infty(\mathbb{R}^d)})\\ &= \frac{\sqrt{2}}{\sqrt{\pi}}\int_{C\sqrt{T}\|b\|_{L^\infty(\mathbb{R}^d)}}^\infty \exp(-s^2/2)\,ds\\ &\lesssim \frac{1}{C\sqrt{T}\|b\|_{L^\infty(\mathbb{R}^d)}}\exp(-C^2T\|b\|_{L^\infty(\mathbb{R}^d)}^2/2),\end{aligned}$$

where we used $B(t)/\sqrt{t} \sim \mathcal{N}(0,1)$ and (25). This shows that for $|x| = |X(0)| \geq n + 1 + (C+1)T\|b\|_{L^\infty(\mathbb{R}^d)}$, we have

$$\begin{aligned}\mathbb{E}(|f(X(t))|) &\leq 0\mathbb{P}(|X(t) - X(0)| < (C+1)T\|b\|_{L^\infty(\mathbb{R}^d)})\\ &\quad + \|f\|_{L^\infty(\mathbb{R}^d)}\mathbb{P}(|X(t) - X(0)| \geq (C+1)T\|b\|_{L^\infty(\mathbb{R}^d)})\\ &\lesssim \frac{\|f\|_{L^\infty(\mathbb{R}^d)}}{C\sqrt{T}\|b\|_{L^\infty(\mathbb{R}^d)}}\exp(-C^2T\|b\|_{L^\infty(\mathbb{R}^d)}^2/2).\end{aligned}$$

Altogether, we obtain

$$\begin{aligned}|u(x)| &\leq c_0^{-1}e^{-Tc_0}\|f\|_{L^\infty(\mathbb{R}^d)} + T\sup_{0\leq t\leq T}\mathbb{E}(|f(X(t))|)\\ &\lesssim \|f\|_{L^\infty(\mathbb{R}^d)}\Big(c_0^{-1}e^{-Tc_0} + \frac{\sqrt{T}}{C}\exp(-C^2T\|b\|_{L^\infty(\mathbb{R}^d)}^2/2)\Big).\end{aligned}$$

Choosing $T = C$ and C as large as possible shows $C \simeq \sqrt{|x|-n}$ and we conclude the proof. □

Theorem 85. *Let u denote the solution of (23) with constant $c(x) := c_0 > 0$, $b(x) := b_0 \in \mathbb{R}^d$ and $f \in W^{1,\infty}(\mathbb{R}^d)$ such that f is supported on a ball around zero with radius $r > 0$. Assume that f can be represented exactly by a neural network on $\mathbb{R}^d$. Given $\varepsilon > 0$, there exists a network $U\colon \mathbb{R}^d \to \mathbb{R}$ and $n \in \mathbb{N}$ with*

$$\begin{aligned}\operatorname{depth}(U) &\leq C\Big(2 + \operatorname{depth}(f) + (2+d)\varepsilon^{-\alpha}\Big)\\ \operatorname{width}(U) &\leq C\max\{2d, \varepsilon^{-4}\max\{\operatorname{width}(f), 2d, 3\}\}\end{aligned}$$

such that $\|u(x) - U(x)\|_{L^2(B_n(0))} \leq \varepsilon$ *and* $|u(x)| \leq \varepsilon$ *for all* $|x| \geq n$. *The constants* $C, \alpha, n > 0$ *do not depend on* d *or* ε.

Proof. Lemma 84 shows that $|u(x)| \leq \varepsilon$ for all $|x| \geq n > 0$ for sufficiently large $n \in \mathbb{N}$ depending only on $\|f\|_{L^\infty(\mathbb{R}^d)}$, $\|b\|_{L^\infty(\mathbb{R}^d)}$, $r > 0$, and c_0. Just as in the proof of Lemma 84, we have

$$u(x) = \int_0^T e^{-tc_0}\mathbb{E}(f(X(t)))\,dt + \int_T^\infty e^{-tc_0}\mathbb{E}(f(X(t)))\,dt$$

for all $T > 0$ and

$$\Big|\int_T^\infty e^{-tc_0}\mathbb{E}(f(X(t)))\,dt\Big| \leq c_0^{-1}e^{-Tc_0}\|f\|_{L^\infty(\mathbb{R}^d)}.$$

We define

$$u_T(x) := \int_0^T e^{-tc_0}\mathbb{E}(f(X(t)))\,dt.$$

There holds

$$\mathbb{E}(f(X(t)) \leq \|f\|_{L^\infty(\mathbb{R}^d)}$$

and hence, this inspires the approximation

$$u_T(x) \approx U(x) := \frac{T}{MN}\sum_{j=1}^N\sum_{i=1}^M e^{-t_ic_0}f(R_{i,\omega_j}(x)) = TQ_M^t(Q_N^\omega((t,\omega)\mapsto e^{-tc_0}f(R_{i,\omega}(x)))), \tag{26}$$

where we use the Monte Carlo quadrature rules

$$\frac{1}{T}\int_0^T g(t,\omega)\,dt \approx Q_M^t(g)(\omega) := \frac{1}{M}\sum_{i=1}^M g(t_i,\omega),$$
$$\int_\Omega g(t,\omega)\,d\omega \approx Q_N^\omega(g)(t) := \frac{1}{N}\sum_{j=1}^N g(t,\omega_j)$$

for randomly chosen $t_i \in [0,T]$ and $\omega_j \in \Omega$. The R_{i,ω_j} are constructed with Lemma 83 on a partition $\widetilde{t}_k := k\tau t_i$ for some $\tau > 0$. Let $\mathbb{E}^\omega$ denote the expectation on Ω and $\mathbb{E}^t := \frac{1}{T}\int_0^T dt$ the expectation on $[0,T]$. Then, we may write

$$u_T(x) = T\mathbb{E}^t\mathbb{E}^\omega\Big((t,\omega)\to e^{-tc_0}f(X(t,\omega))\Big).$$

To estimate the error, we observe

$$\begin{aligned}|u_T(x) - U(x)| \leq{}& (T\mathbb{E}^t - Q_M^t)\mathbb{E}^\omega\Big((t,\omega)\mapsto e^{-tc_0}f(X(\omega,t,x))\Big)\\ &+ \frac{T}{M}\sum_{i=1}^M\mathbb{E}^\omega\Big(\omega\mapsto e^{-t_ic_0}\big(f(X(\omega,t_i,x)) - f(R_{i,\omega}(x))\big)\Big)\\ &+ (\mathbb{E}^\omega - Q_N^\omega)\frac{T}{M}\sum_{i=1}^M\big(\omega\mapsto e^{-t_ic_0}f(R_{i,\omega}(x))\big) =: E_1 + E_2 + E_3.\end{aligned}$$

We collect the random variables $\xi := (\omega_1,\dots,\omega_N,t_1,\dots,t_M) \in \Xi := \Omega^N\times[0,T]^M$ in one variable and endow the space Ξ with the corresponding product probability measures (note that $[0,T]$ has measure one). With the corresponding expectation $\mathbb{E}^\xi$ on Ξ, we

may write $\mathbb{E}^\xi = \mathbb{E}^{\boldsymbol{\omega}}\mathbb{E}^{\boldsymbol{t}}$, with $\mathbb{E}^{\boldsymbol{\omega}}$ and $\mathbb{E}^{\boldsymbol{t}}$ denoting the expectations on Ω^N and $[0,T]^M$, respectively. The standard Monte Carlo error estimate from Theorem 1 shows

$$\mathbb{E}^\xi(E_1^2) = \mathbb{E}^{\boldsymbol{\omega}}\mathbb{E}^{\boldsymbol{t}}(E_1^2) \leq \frac{T^2}{M}\|\mathbb{E}^\omega\Big((\cdot,\omega) \mapsto e^{-tc_0} f(X(\omega,t,x))\Big)\|_{L^2(\Omega)}^2 \leq \frac{T^2}{M}\|f\|_{L^\infty(\mathbb{R}^d)}^2 \quad (27)$$

as well as

$$\mathbb{E}^\xi(E_3^2) = \mathbb{E}^{\boldsymbol{t}}\mathbb{E}^{\boldsymbol{\omega}}(E_3^2) \leq \mathbb{E}^{\boldsymbol{t}}\Big(\frac{1}{N}\|\frac{T}{M}\sum_{i=1}^M \big(\omega \mapsto e^{-t_i c_0} f(R_{i,\omega}(x))\big)\|_{L^2(\Omega)}^2\Big) \leq \frac{T^2}{N}\|f\|_{L^\infty(\mathbb{R}^d)}^2. \quad (28)$$

Standard estimates show

$$\begin{aligned} |E_2| &\leq T \max_{i=1,\dots,M} \|f(X(\cdot,t_i,x)) - f(R_{i,\cdot}(x))\|_{L^2(\Omega)} \\ &\leq T\|f\|_{W^{1,\infty}(\mathbb{R}^d)} \max_{i=1,\dots,M} \|X(\cdot,t_i,x) - R_{i,\cdot}(x)\|_{L^2(\Omega)}. \end{aligned} \quad (29)$$

Note that for $x \in B_n(0)$, there holds $dX(x) = -b_0\,dt + dB$ and hence

$$\mathbb{E}|X(t,x)|^2 \lesssim |x|^2 + |\int_0^t b_0\,dt|^2 + \mathbb{E}|B(t)|^2 \leq n^2 + t|b_0| + t.$$

This shows $\sup_{0\leq t\leq T}\mathbb{E}|X(t,x)|^2 \leq n^2 + (1+|b_0|)T < \infty$ and, with Lemma 82, shows $\sqrt{\mathbb{E}^\omega|X(t_i,x) - R_i(x)|^2} \leq Ce^{CT}\sqrt{\tau}$. With (29), this implies

$$\sqrt{\mathbb{E}^\xi(E_2^2)} = \sqrt{\mathbb{E}^{\boldsymbol{t}}|E_2|^2} \leq C\|f\|_{W^{1,\infty}(\mathbb{R}^d)}e^{CT}\sqrt{\tau}. \quad (30)$$

The combination of (27), (28), and (30) shows

$$\sqrt{\mathbb{E}^\xi|u_T(x) - U(x)|^2} \lesssim \Big(\frac{T}{\sqrt{N}} + \frac{T}{\sqrt{M}}\Big)\|f\|_{L^\infty(\mathbb{R}^d)} + Ce^{CT}T\|f\|_{W^{1,\infty}(\mathbb{R}^d)}\sqrt{\tau}.$$

Choosing $N \simeq M \simeq \varepsilon^{-2}$, $T \simeq \log(\varepsilon)/c_0$, and $\tau \simeq \varepsilon^{2+2(C/c_0+1)}$, we obtain

$$\sqrt{\mathbb{E}^\xi|u(x) - U(x)|^2} \leq |u(x) - u_T(x)| + \sqrt{\mathbb{E}^\xi|u(x) - U(x)|^2} \leq \varepsilon.$$

This implies

$$\mathbb{E}^\xi\Big(\|u - U\|_{L^2(B_n(0))}^2\Big) = \int_{B_n(0)} \mathbb{E}^\xi|u(x) - U(x)|^2\,dx \leq |B_n(0)|\varepsilon^2.$$

Since we bound the expectation of the positive quantity $\|u - U\|_{L^2(B_n(0))}^2$ by $|B_n(0)|\varepsilon^2$, we know that there is at least one $\xi \in \Xi$ with $\|u - U\|_{L^2(B_n(0))}^2 \leq |B_n(0)|\varepsilon^2$. Note that $n \in \mathbb{N}$ depends only on u and hence on f. We may estimate the volume of the ball by

$$|B_n(0)| = \frac{(\sqrt{\pi}n)^d}{\Gamma(d/2+1)} \leq \frac{(\pi n^2)^{\lfloor d/2\rfloor+1}}{\lfloor d/2\rfloor!}.$$

Since $x^j/j! \to 0$ as $j \to \infty$ for all $x \in \mathbb{R}$, we can bound $|B_n(0)| \leq C$, where $C > 0$ depends on n but not on d.

It remains to show that $U(x) = U_\xi(x)$ can be represented by a neural network. By assumption, f is a neural network and $R_{i,\omega_j}(\cdot)$ are neural networks by Lemma 83. Lemma 73 shows that $f \circ R_{i,\omega_j}(\cdot)$ is a neural network and another application of Lemma 73 shows that the weighted sum $U(x)$ can be represented by a neural network with (note that $\operatorname{depth}(b_0) = 2$ and $\operatorname{width}(b_0) \leq d$)

$$\operatorname{depth}(U) \leq 2 + \operatorname{depth}(f) + \max_{\substack{i=1,\dots M\\ j=1,\dots N}} \operatorname{depth}(R_{i,\omega_j}) \leq 2 + \operatorname{depth}(f) + (2+2)T\tau^{-1} + 2$$

and

$$\begin{aligned}\operatorname{width}(U) &\leq \max\{2d, \sum_{i=1}^{M}\sum_{j=1}^{N} \max\{\operatorname{width}(f), \operatorname{width}(R_{i,\omega_j})\}\\ &\leq \max\{2d, \sum_{i=1}^{M}\sum_{j=1}^{N} \max\{\operatorname{width}(f), 2\dim(b_0) + 2d, 2\}\\ &\leq \max\{2d, NM \max\{\operatorname{width}(f), 2d + 2d, 2\}\}.\end{aligned}$$

This concludes the proof. □

Remark 86. *Note that the assumption that f is compactly supported and can exactly be represented by a neural network is restrictive. While it is straightforward to construct a not-trivial network f with compact support by*

$$f(x_1, \dots, x_d) := 1 - \min\{\sum_{i=1}^{d} |x_i|, 1\},$$

there might be some restrictions for more general right-hand sides. The proof of Theorem 85 can be improved by including multiplication with a cut-off function into the construction. This would remove the restriction of f being a neural network on the whole of $\mathbb{R}^d$ and replace it with f being a neural network on $B_n(0)$.

4.6. **Convergence of gradient descent on a two-layer network.** This section is inspired by [4]. Consider the two layer network $F\colon \mathbb{R}^d \to \mathbb{R}$ defined by

$$F(\boldsymbol{W}, \boldsymbol{x}) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi(\boldsymbol{w}_i \cdot \boldsymbol{x}),$$

with $\boldsymbol{w}_i$ denoting the rows of $\boldsymbol{W} \in \mathbb{R}^{m\times s}$. We consider the second layer as fixed and only train the first layer $\boldsymbol{W}$. The normalization $\frac{1}{\sqrt{m}}$ is only for convenience in the proofs below. We consider the standard quadratic loss function

$$\mathcal{L}(\boldsymbol{W}) := \frac{1}{2} \sum_{i=1}^{n} (F(\boldsymbol{W}, \boldsymbol{x}_i) - y_i)^2$$

for given data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and responses $y_1, \dots, y_n$. For simplicity, we assume $|\boldsymbol{x}_i| = 1$, $i = 1, \dots, n$. We compute

$$\partial_{\boldsymbol{w}_j} \mathcal{L}(\boldsymbol{W}) = \frac{1}{\sqrt{m}} \sum_{i=1}^{n} (F(\boldsymbol{W}, \boldsymbol{x}_i) - y_i) \boldsymbol{x}_i \mathbf{1}_{\boldsymbol{w}_j \cdot \boldsymbol{x}_i \geq 0}. \tag{31}$$

Define the vectors $\boldsymbol{y}_i \in \mathbb{R}^{sm}$ by

$$\boldsymbol{y}_i(\boldsymbol{W}) := (\boldsymbol{x}_i \mathbf{1}_{\boldsymbol{w}_1 \cdot \boldsymbol{x}_i \geq 0}, \dots, \boldsymbol{x}_i \mathbf{1}_{\boldsymbol{w}_m \cdot \boldsymbol{x}_i \geq 0})$$

in order to write

$$\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) = \frac{1}{\sqrt{m}} \sum_{i=1}^{n} (F(\boldsymbol{W}, \boldsymbol{x}_i) - y_i) \boldsymbol{y}_i(\boldsymbol{W}).$$

To simplify the proof, we have a number of assumptions that can be circumvented with more careful analysis. First, we assume that the data is normalized (which is a pretty common practice in applications), i.e., $|\boldsymbol{x}_i| = 1$. This allows allows us to avoid using a bias in the network. Moreover, we will assume that the initial guess $\boldsymbol{W}_0$ for the gradient descent (Algorithm 4) is drawn from a uniform distribution on the sphere in $\mathbb{R}^s$, i.e., each

row is drawn independently and we write $\boldsymbol{W}_0 \simeq U_s$. Note that such a sample can be generated by first drawing $\widetilde{\boldsymbol{W}}_0 \sim \mathcal{N}(0,1) \in \mathbb{R}^s$ and setting $\boldsymbol{W}_0 := \widetilde{\boldsymbol{W}}_0/|\widetilde{\boldsymbol{W}}_0|$. Finally, we assume that no two vectors $\boldsymbol{x}_i$ are parallel, i.e., $\boldsymbol{x}_i \notin \{\boldsymbol{x}_j, -\boldsymbol{x}_j\}$ for all $i \neq j$. In this section, we will prove the following result.

Theorem 87. *Let $\boldsymbol{W}_0 \sim U_d$ be the random initialization of the network F and let $\delta > 0$. For sufficiently large $m \in \mathbb{N}$, there exists $0 < \kappa < 1$ and a step-size $\alpha > 0$ such that with probability $1 - \delta$, Algorithm 4 applied to $\mathcal{L}(\boldsymbol{W})$ satisfies $\mathcal{L}(\boldsymbol{W}_\ell) \leq \kappa^\ell \mathcal{L}(\boldsymbol{W}_0)$ for $\ell \in \mathbb{N}$.*

Remark 88. *Theorem 87 states the following: As long as the network F is sufficiently wide, i.e., $m \in \mathbb{N}$ is large, there is a high probability such that gradient descent with initial guess $\boldsymbol{W}_0$ will find $\boldsymbol{W}$ with $\mathcal{L}(\boldsymbol{W})$ arbitrarily small.*

Note that this result is an extension of the universal approximation theorem (Theorem 70), which only provides existence of $\boldsymbol{W}$ such that $\mathcal{L}(\boldsymbol{W})$ is small. The result from this section additionally shows that $\boldsymbol{W}$ can be found by applying gradient descent.

While all the other assumptions on the data can be weakened by more careful analysis, and a similar proof works for deep networks, the assumption that $m \in \mathbb{N}$ is sufficiently large is essential.

4.6.1. *Random distribution on the unit sphere.* We first prove a couple of results on random vectors on the unit sphere. This is required since the initialization $\boldsymbol{W}_0$ of the gradient descent is chosen randomly.

Lemma 89. *Let $S_d := \{\boldsymbol{z} \in \mathbb{R}^d : |\boldsymbol{z}| = 1\}$ denote the unit sphere. For $\boldsymbol{x} \in \mathbb{R}^d$, let $C_\pm(\varepsilon, \boldsymbol{x}) := \{\boldsymbol{z} \in \mathbb{R}^d : |\boldsymbol{z}| = 1, \pm\boldsymbol{z} \cdot \boldsymbol{x} \geq \varepsilon\}$ denote spherical caps and let $E(\varepsilon, \boldsymbol{x}) := S_d \setminus (C_+(\varepsilon, \boldsymbol{x}) \cup C_-(\varepsilon, \boldsymbol{x}))$ denote the equator. Then, there holds (with $|\cdot|$ for surface area)*

$$|S_d| = \frac{2\pi^{d/2}}{\Gamma(d/2)} \quad \text{and} \quad |E(\varepsilon, \boldsymbol{x})| \simeq |S_{d-1}|\varepsilon.$$

Proof. Without loss of generality, we may assume $\boldsymbol{x} = (0, \dots, 0, 1) \in \mathbb{R}^d$. For $\boldsymbol{z} \in \mathbb{R}^d$, let $\boldsymbol{z}' := (\boldsymbol{z}_1, \dots, \boldsymbol{z}_{d-1}, 0) \in \mathbb{R}^d$ denote the first $d-1$ coordinates. Let $R_\varepsilon := S_{d-1} \times [-\varepsilon, \varepsilon]$ denote the cylinder with radius one and height 2ε. Define the map $\phi\colon R_\varepsilon \to \mathbb{R}^d$ via $\phi(\boldsymbol{z}) := \boldsymbol{z}/|\boldsymbol{z}| = \boldsymbol{z}/\sqrt{1 + \boldsymbol{z}_d^2}$ and compute

$$\partial_{\boldsymbol{z}_i} \phi_j(\boldsymbol{z}) = \begin{cases} \delta_{ij} & 1 \leq i \leq d-1, \\ \delta_{jd} \dfrac{\sqrt{1+\boldsymbol{z}_d^2} - \frac{\boldsymbol{z}_d^2}{\sqrt{1+\boldsymbol{z}_d^2}}}{1+\boldsymbol{z}_d^2} = \delta_{jd}(1+\boldsymbol{z}_d^2)^{-3/2} & i = d. \end{cases}$$

Hence, we have $\det D\phi = (1 + \boldsymbol{z}_d^2)^{-3/2}$. Moreover, there holds for all $\boldsymbol{z} \in E(\varepsilon, \boldsymbol{x})$ that

$$|\boldsymbol{z}'|^2 = 1 - \boldsymbol{z}_d^2 \geq 1 - \varepsilon^2$$

and therefore $\widetilde{\boldsymbol{z}} := \boldsymbol{z}/|\boldsymbol{z}'| \in R_{\varepsilon/\sqrt{1-\varepsilon^2}}$ as well as $\phi(\widetilde{\boldsymbol{z}}) = \boldsymbol{z}$. Hence, an integral transformation allows us to estimate

$$|E(\varepsilon, \boldsymbol{x})| = \int_{R_{\varepsilon/\sqrt{1-\varepsilon^2}}} \det(D\phi)\, d\widetilde{\boldsymbol{z}} \simeq |R_{\varepsilon/\sqrt{1-\varepsilon^2}}| \simeq \varepsilon |S_{d-1}|.$$

□

Lemma 90. *For any choice of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, there exists a signature $\sigma \in \{+,-\}^n$ such that*

$$\Big|\bigcap_{i=1}^{n} C_{\sigma_i}(0, \boldsymbol{x}_i)\Big| \geq \frac{|S_d|}{2\sum_{j=0}^{d-1}\binom{n-1}{j}},$$

where we define $\binom{n-1}{j} = 0$ for $j \geq n$.

Remark 91. *Note that this bound is much better than the naive bound obtained by the fact that there are at most 2^n different signatures.*

Proof. Given a set of hyperplanes $H_1, \dots, H_n \in \mathbb{R}^d$, their union $\bigcup_{i=1}^n H_i$ splits $\mathbb{R}^d$ into a number of open *cells* C_j, $j = 1, \dots, m(n,d) \in \mathbb{N}$. Obviously, there holds $m(n,1) = 2$ as a one dimensional hyperplane is just $\{0\}$ and $m(1,d) = 2$ since one hyperplane splits $\mathbb{R}^d$ into two cells.

Assume n hyperplanes $H_1, \dots, H_n$ with cells C_j and add another hyperplane H_{n+1}. A new cell is generated by splitting some C_j with H_{n+1}. In this case $C_j \cap H_{n+1}$ is a cell of H_{n+1} generated by $H_1, \dots, H_n$. Hence, we have

$$m(n+1,d) \leq m(n,d) + m(n,d-1).$$

Since also

$$\sum_{j=0}^{d-1}\binom{n}{j} = \sum_{j=0}^{d-1}\binom{n-1}{j} + \sum_{j=0}^{d-2}\binom{n-1}{j}$$

we see from induction that

$$m(n,d) \leq 2\sum_{j=0}^{d-1}\binom{n-1}{j}$$

(there even holds equality).

To estimate the minimal area of an intersection of the form $C := \bigcap_{i=1}^n C_{\sigma_i}(\boldsymbol{x}_i)$, define the family

$$\mathcal{C} := \Big\{\bigcap_{i=1}^{n} C_{\sigma_i}(\boldsymbol{x}_i) \,:\, (\sigma_i)_{i=1}^n \in \{-1,1\}^n\Big\}.$$

Note that each $C \in \mathcal{C}$ is the intersection of some cell generated by $\{H_i := \big\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{z}\cdot\boldsymbol{x}_i = 0\big\},\ i = 1, \dots, n\}$ with S_d. Hence the sets $C \in \mathcal{C}$ are disjoint and their number is bounded by $\#\mathcal{C} \leq m(n,d)$. Since $S_d = S_d \cap \bigcup_{C\in\mathcal{C}} C$, there exists at least one C with $|C| \geq |S_d|/m(n,d)$. This concludes the proof. □

Lemma 92. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ with $|\boldsymbol{x}_i| = 1$, $i = 1, \dots, n$ such that no two $\boldsymbol{x}_i$ are parallel. Then, there exists $\varepsilon > 0$ and $\boldsymbol{w}_\pm \in S_d$ such that for all $\widetilde{\boldsymbol{w}}_\pm \in \mathbb{R}^d$ with $|\boldsymbol{w}_\pm - \widetilde{\boldsymbol{w}}_\pm| \leq \varepsilon$, there holds*

$$\pm\widetilde{\boldsymbol{w}}_\pm \cdot \boldsymbol{x}_1 > 0 \quad \textit{and} \quad \operatorname{sign}(\widetilde{\boldsymbol{w}}_+ \cdot \boldsymbol{x}_j) = \operatorname{sign}(\widetilde{\boldsymbol{w}}_- \cdot \boldsymbol{x}_j) \quad \textit{for all } j = 2, \dots, n. \tag{32}$$

For $\boldsymbol{w}_\pm \sim U_d \in \mathbb{R}^d$, the probability p of $\boldsymbol{w}_\pm$ satisfying (32) is bounded from below by

$$p \geq c\frac{\varepsilon}{\sum_{j=0}^{d-2}\binom{n-1}{j}},$$

where $c > 0$ is independent of d, n, and the $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. The constant $\varepsilon > 0$ satisfies $\varepsilon \simeq \min_{1\leq i\neq j\leq n} |\boldsymbol{x}_i - \boldsymbol{x}_j|^2$.

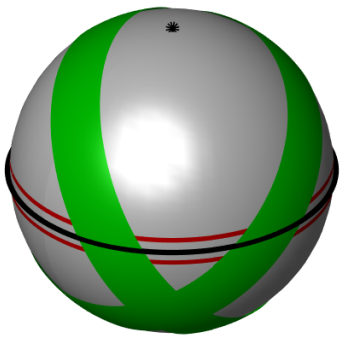

FIGURE 10. The sphere depicts S_d and the north-pole is $\boldsymbol{x}_1$. The black equator depicts S_{d-1} and the green areas symbolize the union of the equators $E(\varepsilon, \boldsymbol{x}_i)$, $i = 2, \dots, n$. The vectors $\boldsymbol{w}_\pm$ must be close to S_{d-1} in order for $\pm \boldsymbol{w}_\pm \cdot \boldsymbol{x}_1 > 0$. However, they can't lie in the green areas, since $\boldsymbol{w}_\pm \cdot \boldsymbol{x}_j$ does not change sign for $j = 2, \dots, n$. In the proof of Theorem 92, we first show that the intersection of S_{d-1} with the complement of the green areas has positive $d-1$-dimensional surface measure (this is V). Then, we define the red areas above and below S_{d-1} (those are V_+ and V_-) which are sufficiently close to V.

Proof. Consider $H := \{\boldsymbol{z} \in \mathbb{R}^d \,:\, \boldsymbol{z} \cdot \boldsymbol{x}_1 = 0\}$. Without loss of generality, we may assume $H = \mathbb{R}^{d-1} \times \{0\}$ and $\boldsymbol{x}_1 = \boldsymbol{e}_d$. We redefine $S_{d-1} := S_{d-1} \times \{0\}$ as a subset of $\mathbb{R}^{d-1} \times \{0\}$. See Figure 10 for a sketch of the proof strategy.

For $\boldsymbol{x}_j \in S_d$, let $\boldsymbol{x}'_j \in \mathbb{R}^{d-1}$ denote the first $d-1$ dimensions and $x_{j,d}$ the last, i.e., $\boldsymbol{x}_j = (\boldsymbol{x}'_j, x_{j,d})$. Since no two $\boldsymbol{x}_i$ are parallel, there exists $\varepsilon_0 > 0$ such that $|\boldsymbol{x}_1 \pm \boldsymbol{x}_j| \geq \varepsilon_0$ for $j > 1$. With $|\boldsymbol{x}'_j|^2 + |x_{j,d}|^2 = 1$, there holds

$$\varepsilon_0^2 \leq |\boldsymbol{x}'_j|^2 + |x_{j,d} \pm 1|^2 = 2 \pm 2x_{j,d}.$$

This implies $|x_{j,d}| \leq 1 - \varepsilon_0^2/2$ and therefore $|\boldsymbol{x}'_j| \geq \varepsilon_0^2/2$. Recall the equator $E(\varepsilon, \boldsymbol{x}_i)$ from Lemma 89 and note that

$$E(\varepsilon, \boldsymbol{x}_j) \cap S_{d-1} = \{\boldsymbol{z} \in S_{d-1} \,:\, |\boldsymbol{z} \cdot \boldsymbol{x}'_j| \leq \varepsilon\} = E_{d-1}\Big(\frac{\varepsilon}{|\boldsymbol{x}'_j|}, \frac{\boldsymbol{x}'_j}{|\boldsymbol{x}'_j|}\Big),$$

where E_{d-1} denotes the equator of S_{d-1}. Lemma 89 shows consequently for the surface measure on S_{d-1} that

$$|\bigcup_{\substack{j=1 \\ i \neq j}}^{n} E(\varepsilon, \boldsymbol{x}_j) \cap S_{d-1}| \leq c_0 n |S_{d-2}| \varepsilon / \varepsilon_0^2$$

for some constant $c_0 > 0$. With Lemma 90 applied to $\mathbb{R}^{d-1}$, we find some signature $\sigma \in \{+,-\}^n$, such that

$$V := S_{d-1} \cap \bigcap_{j=2}^{n} C_{\sigma_j}(0, \boldsymbol{x}_j) \setminus \bigcup_{j=2}^{n} E(\varepsilon, \boldsymbol{x}_j)$$

which satisfies

$$|V| \geq \frac{|S_{d-1}|}{2\sum_{j=0}^{d-2}\binom{n-1}{j}} - c_0 n |S_{d-2}| \frac{\varepsilon}{\varepsilon_0^2} = \frac{|S_{d-1}|}{4\sum_{j=0}^{d-2}\binom{n-1}{j}}$$

for $\varepsilon = \frac{\varepsilon_0^2 |S_{d-1}|}{n c_0 |S_{d-2}|} \frac{1}{4\sum_{j=0}^{d-2}\binom{n-1}{j}}$.

Let $\boldsymbol{z} \in V$ and $\widetilde{\boldsymbol{z}} \in S_d$ with $|\boldsymbol{z} - \widetilde{\boldsymbol{z}}| < \varepsilon$. Then, there holds for $j > 1$

$$|\widetilde{\boldsymbol{z}} \cdot \boldsymbol{x}_j| > |\boldsymbol{z} \cdot \boldsymbol{x}_j| - \varepsilon \geq 0. \tag{33}$$

This inspires the definition

$$V_\pm := \big\{ \boldsymbol{z} \in S_d \,:\, \operatorname{dist}(V, \boldsymbol{z}) < \varepsilon/2,\, \pm z_d \geq \varepsilon/4 \big\}.$$

Note that (33) implies for $\boldsymbol{w} \in V_+$ and $\widetilde{\boldsymbol{w}} \in \mathbb{R}^d$ with $|\boldsymbol{w} - \widetilde{\boldsymbol{w}}| \leq \varepsilon/2$ and $j > 1$ that

$$|\widetilde{\boldsymbol{w}} \cdot \boldsymbol{x}_j| > 0 \quad \text{and therefore} \quad \operatorname{sign}(\widetilde{\boldsymbol{w}} \cdot \boldsymbol{x}_j) = \sigma_j.$$

Moreover, there holds

$$\widetilde{\boldsymbol{w}} \cdot \boldsymbol{x}_1 = \boldsymbol{z}_d \geq \varepsilon/4.$$

The analogous estimates hold for V_- and hence all $\boldsymbol{w}_\pm \in V_\pm$ satisfy (32). Similarly to Lemma 89, we may estimate $|V_\pm| \gtrsim |V|\varepsilon/4$.

It remains to calculate the probability of randomly picking $\boldsymbol{w} \sim U_d \in V_+$ (analogously for V_-). Since $\boldsymbol{w} \sim \mathcal{N}(0,1)$, the normalization $\boldsymbol{w}/|\boldsymbol{w}|$ is distributed uniformly on the sphere. Therefore, the probability is given by $|V_+|/|S_d|$. We conclude the proof with the formula for $|S_d|$ by

$$\frac{|S_d|}{|S_{d-1}|} = \frac{\sqrt{\pi}\Gamma((d-1)/2)}{\Gamma(d/2)} \leq \sqrt{\pi}.$$

□

4.6.2. *The gradient of $\mathcal{L}(\boldsymbol{W})$.* The following results will allow us to estimate the gradient of the loss function.

Lemma 93. *Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in S_d$ such that no two $\boldsymbol{x}_i$ are parallel, let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_m \sim U_d \in \mathbb{R}^d$ be chosen randomly and let $\tilde{\boldsymbol{w}}_1, \dots, \tilde{\boldsymbol{w}}_m \in \mathbb{R}^d$ such that*

$$|\tilde{\boldsymbol{w}}_i - \boldsymbol{w}_i| < \varepsilon \quad \textit{for all } i = 1, \dots, m,$$

with $\varepsilon > 0$ from Lemma 92. Define the matrix $X \in \mathbb{R}^{n \times m}$ via $X_{ij} := \mathbf{1}_{\tilde{\boldsymbol{w}}_j \cdot \boldsymbol{x}_i \geq 0}$ and assume that no two $\boldsymbol{x}_i$ are parallel. Given $\delta > 0$, there exists $m \in \mathbb{N}$ such that the matrix X has full rank with probability $1 - \delta$. The constant m grows polynomially in n and logarithmically in δ.

Proof. For each $\boldsymbol{x}_i$, we apply Lemma 92 (with $\boldsymbol{x}_1 = \boldsymbol{x}_i$ in the notation of Lemma 92) and obtain $\boldsymbol{w}_{i,\pm}$ with the properties (32). For $|\widetilde{\boldsymbol{w}}_{j,\pm} - \boldsymbol{w}_{j,\pm}| \leq \varepsilon$, this implies immediately

$$\mathbf{1}_{\widetilde{\boldsymbol{w}}_{j,+} \cdot \boldsymbol{x}_i \geq 0} = \mathbf{1}_{\widetilde{\boldsymbol{w}}_{j,-} \cdot \boldsymbol{x}_i \geq 0} \quad \text{for all } i \neq j$$

and

$$\mathbf{1}_{\widetilde{\boldsymbol{w}}_{i,+} \cdot \boldsymbol{x}_i \geq 0} \neq \mathbf{1}_{\widetilde{\boldsymbol{w}}_{i,-} \cdot \boldsymbol{x}_i \geq 0}.$$

If $\boldsymbol{w}_1 = \boldsymbol{w}_{1,+}$, $\boldsymbol{w}_2 = \boldsymbol{w}_{1,-}$, $\boldsymbol{w}_3 = \boldsymbol{w}_{2,+}$ and so on, this shows that the rows of X are linearly independent.

Lemma 92 gives a lower bound on the probability $0 < p < 1$ of randomly choosing a vector $\boldsymbol{w}_i$ with the above properties. Hence, given m random vectors $\boldsymbol{w}$, the probability of finding at least n vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_n$ with the above properties is given by the binomial distribution and hence

$$P := \sum_{j=n}^{m} \binom{m}{j} p^j (1-p)^{m-j}.$$

Clearly, the probability tends to one if $m \to \infty$. A tail bound (Hoeffding's inequality) on the binomial distribution reveals

$$P = \sum_{j=n}^{m} \binom{m}{j} p^j (1-p)^{m-j} \geq 1 - \exp\Big(-2m\big(p - \frac{n}{m}\big)^2\Big). \tag{34}$$

With the lower bound on p from Lemma 92, we conclude that

$$m \geq \max\{2n \frac{\sum_{j=0}^{d-2} \binom{n-1}{j}}{c\varepsilon}, |\log(\delta)|\}$$

is sufficient for $P \geq 1 - \delta$. This concludes the proof. □

In the following, we will consider the weight matrices $\boldsymbol{W}$ also as vectors and hence $|\boldsymbol{W}| := \sqrt{\sum_{i,j} \boldsymbol{W}_{ij}^2}$ denotes the Frobenius norm.

Lemma 94. *Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in S_d$ such that no two $\boldsymbol{x}_i$ are parallel. Let $m \in \mathbb{N}$ and $\varepsilon > 0$ from Lemma 93. For $m' \in \mathbb{N}$, let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{m'} \sim U_d \in \mathbb{R}^d$ and define*

$$\boldsymbol{W} := \begin{pmatrix} \boldsymbol{w}_1^T \\ \vdots \\ \boldsymbol{w}_{m'}^T \end{pmatrix} \in \mathbb{R}^{m' \times d}.$$

Let $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{m' \times d}$ with $|\boldsymbol{W} - \widetilde{\boldsymbol{W}}| \leq C$.

If $m' \gtrsim m$ (the hidden constant depends only on d), the matrix $Y \in \mathbb{R}^{n \times n}$, $Y_{jk} := \boldsymbol{y}_j(\widetilde{\boldsymbol{W}}) \cdot \boldsymbol{y}_k(\widetilde{\boldsymbol{W}})$ with $\boldsymbol{y}_j(\widetilde{\boldsymbol{W}}))$ from (31) is symmetric positive definite with probability $1 - \delta$.

Proof. Since $|\boldsymbol{W} - \widetilde{\boldsymbol{W}}|^2 = \sum_{j=1}^{m'} |\boldsymbol{w}_j - \widetilde{\boldsymbol{w}}_j|^2 \leq C^2$, there exist at least $m'/2$ indices $1 \leq j_k \leq m'$ such that

$$|\boldsymbol{w}_{j_k} - \widetilde{\boldsymbol{w}}_{j_k}|^2 \leq 2C^2/m'.$$

With $m' \geq 2C^2/\varepsilon^2$, me may restrict the $\boldsymbol{y}_i := \boldsymbol{y}_i(\widetilde{\boldsymbol{W}})|_{j_1,\dots,j_{m'}}$ to the indices $j_1, \dots, j_{m'}$ and prove linear independence of this subsystem which satisfies $|\boldsymbol{w}_j - \widetilde{\boldsymbol{w}}_j| \leq \varepsilon$ for all $j = j_k$, $k = 1, \dots, m'$. Assume that $\sum_{i=1}^{n} \alpha_i \boldsymbol{y}_i = 0$ for some $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$. Then, there holds with the matrix $X \in \mathbb{R}^{n \times m'}$ from Lemma 93 that

$$0 = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{\widetilde{\boldsymbol{w}}_k \cdot \boldsymbol{x}_i \geq 0} \boldsymbol{x}_i = \boldsymbol{\alpha}^T X_{:,k} \tag{35}$$

for all $k = 1, \dots, m'$. This implies $\boldsymbol{\alpha}^T X = 0$, and since X has full rank n with probability $1 - \delta$, we have $\boldsymbol{\alpha} = 0$. Hence, Y is regular and symmetric and there holds

$$Y\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} = \Big| \sum_{i=1}^{n} \alpha_i \boldsymbol{y}(\widetilde{\boldsymbol{W}}) \Big|^2 > 0.$$

This concludes the proof. □

Corollary 95. *Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in S_d$ such that no two $\boldsymbol{x}_i$ are parallel. Let $\boldsymbol{W} \sim U_d$ with $|\boldsymbol{W} - \widetilde{\boldsymbol{W}}| \leq \widetilde{C}$. For each $m' \in \mathbb{N}$ and $\delta > 0$, there exist $m \geq m'$ such that with probability $1 - \delta$ over $\boldsymbol{W} \sim U_d$, there holds*

$$\gamma m |\boldsymbol{\alpha}|^2 \leq Y\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} \leq \Gamma_{\widetilde{C}} + m\Gamma |\boldsymbol{\alpha}|^2$$

for all $\boldsymbol{\alpha} \in \mathbb{R}^n$, where Y is the matrix from Lemma 94. The constants $\gamma, \Gamma > 0$ do not depend on $\widetilde{C}$ ($\Gamma_{\widetilde{C}}$ does depend on $\widetilde{C}$) and all constants are independent of m.

Proof. Let $m_C \in \mathbb{N}$ be minimal such that the statement of Lemma 94 is true for some $C \geq 1$. Observe that at most $m_{\widetilde{C}}\widetilde{C}^2$ indices $1 \leq j_k \leq m$ satisfy $|\boldsymbol{w}_{j_k} - \widetilde{\boldsymbol{w}}_{j_k}|^2 \geq 1/m_{\widetilde{C}}$. Without loss of generality, assume that those are the first $n_0 := m_{\widetilde{C}}\widetilde{C}^2$ indices. We set $m = n_0 + km_1$ for some $k \in \mathbb{N}$ and split the remaining indices into chunks of size m_1, i.e, $j = n_0 + rm_1 + 1, \dots, n_0 + (r+1)m_1$ for $0 \leq r < k$. We may apply Lemma 94 to the vectors $\boldsymbol{w}_j$, $j = n_0 + rm_1 + 1, \dots, n_0 + (r+1)m_1$ and obtain corresponding matrices $Y_{r+1} \in \mathbb{R}^{n \times n}$ with Y_0 denoting the matrix corresponding to the first n_0 indices.

Finally, let Y denote the matrix corresponding to all indices $j = 1, \dots, m$. Note that $Y = \sum_{r=0}^{k-1} Y_r$ (because $\boldsymbol{y}_i \cdot \boldsymbol{y}_k = \boldsymbol{x}_i \cdot \boldsymbol{x}_k \sum_{j=1}^m \mathbf{1}_{\widetilde{\boldsymbol{w}}_j \cdot \boldsymbol{x}_i} \mathbf{1}_{\widetilde{\boldsymbol{w}}_j \cdot \boldsymbol{x}_k}$). For all $r > 0$, we can apply Lemma 94 with $C \leq 1$, since

$$\sqrt{\sum_{j=n_0+rm_1+1}^{n_0+(r+1)m_1} |\boldsymbol{w}_{j_k} - \widetilde{\boldsymbol{w}}_{j_k}|^2} \leq m_1/m_{\widetilde{C}} \leq 1.$$

Moreover, with m_1 and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ fixed, there are only finitely many possible matrices Y_r, $r > 0$. Hence, there holds $\max_{1 \leq r \leq k} \|Y_r\|_2 \leq C_{\mathrm{mat}}$ independently of the choice of $\boldsymbol{W}$ and $\widetilde{\boldsymbol{W}}$. Moreover, Lemma 94 shows that each Y_r is positive definite with probability $1 - \delta$. Again, the tail bound (25) for the binomial distribution shows that the probability P to have $0 < k' := \lfloor k(1-\delta)/2 \rfloor < k$ positive definite matrices $Y_{j_1}, \dots, Y_{j_{k'}}$ among $Y_1, \dots, Y_k$ is bounded by

$$P \geq 1 - \exp\Big(- 2k(1 - \delta - \frac{k'}{k})^2 \Big) = 1 - \exp(-(1-\delta)^2 k/2).$$

The minimal ellipticity constant among the positive definite Y_r is bounded since there are only finitely many different matrices, i.e.

$$\min_{1 \leq i \leq k'} \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} Y_{j_i}\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} \geq c_{\mathrm{mat}} |\boldsymbol{\alpha}|^2.$$

Hence, we have with probability $1 - \delta'$ for $k \simeq \log(\delta')$ that

$$\|Y\|_2 \leq \|Y_0\|_2 + \sum_{r=1}^k \|Y_k\|_2 \lesssim \|Y_0\|_2 + k \quad \text{and} \quad Y\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} = Y_0\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} + \sum_{r=1}^k Y_r\boldsymbol{\alpha} \cdot \boldsymbol{\alpha} \gtrsim k|\boldsymbol{\alpha}|^2$$

and the hidden constants do not depend on $\widetilde{C} > 0$ (Note that the norm of Y_0 does depend on $\widetilde{C}$). This concludes the proof. □

Lemma 96. *Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in S_d$ such that no two $\boldsymbol{x}_i$ are parallel, choose $\delta > 0$. Let $\boldsymbol{W} \sim U_d \in \mathbb{R}^{m \times d}$ and assume $|\boldsymbol{W} - \widetilde{\boldsymbol{W}}| \leq C$. There exists $m \in \mathbb{N}$ and constants $0 < q < Q < \infty$ which depend only on the data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, $\delta > 0$, and d (but not on C) such that*

$$q\mathcal{L}(\widetilde{\boldsymbol{W}}) \leq |\nabla_{\boldsymbol{W}} \mathcal{L}(\widetilde{\boldsymbol{W}})|^2 \leq Q\mathcal{L}(\widetilde{\boldsymbol{W}})$$

with probability at least $1 - \delta$ over the initialization $\boldsymbol{W}$.

Proof. We obtain from (31)

$$|\nabla_{\boldsymbol{W}}\mathcal{L}(\widetilde{\boldsymbol{W}})|^2 = \frac{1}{m}\Big|\sum_{i=1}^{n}(F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})\Big|^2.$$

With the matrix Y from Corollary 95, this can be written as

$$|\nabla_{\boldsymbol{W}}\mathcal{L}(\widetilde{\boldsymbol{W}})|^2 = \frac{1}{m}Y\boldsymbol{\alpha}\cdot\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}\in\mathbb{R}^n$ with $\alpha_i := (F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)$. With probability $1-\delta$, Corollary 95 shows

$$\frac{1}{m}\Big|\sum_{i=1}^{n}(F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})\Big|^2 \leq \frac{\Gamma_C+m\Gamma}{m}\sum_{i=1}^{n}(F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)^2$$

as well as

$$\frac{1}{m}\Big|\sum_{i=1}^{n}(F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})\Big|^2 \geq \gamma\sum_{i=1}^{n}(F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)^2$$

We also have

$$\sum_{i=1}^{n}(F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)-y_i)^2 = 2\mathcal{L}(\widetilde{\boldsymbol{W}}).$$

Choosing m sufficiently large, we force $\Gamma_C/m\leq 1$ and conclude the proof. □

Lemma 97. *Given $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\in S_d$ such that no two $\boldsymbol{x}_i$ are parallel. Let $\boldsymbol{W}\sim U_d\in\mathbb{R}^{m\times d}$ and $\widetilde{\boldsymbol{W}}\in\mathbb{R}^{m\times d}$ with $|\boldsymbol{W}-\widetilde{\boldsymbol{W}}|\leq C$. Then, given $\delta>0$, there exists $\varepsilon_0>0$ such that for all $0<\varepsilon\leq\varepsilon_0$, there holds with probability $1-\delta$ that*

$$|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W})-\nabla_{\boldsymbol{W}}\mathcal{L}(\widetilde{\boldsymbol{W}})|\leq Qn|\boldsymbol{W}||\boldsymbol{W}-\widetilde{\boldsymbol{W}}|+\varepsilon\sqrt{\mathcal{L}(\boldsymbol{W})}.$$

as long as $m\gtrsim C^2\varepsilon^{-2}n$ with $Q>0$ being independent of n, m, C and ε.

Proof. Since $|\boldsymbol{W}-\widetilde{\boldsymbol{W}}|^2=\sum_{j=1}^{m}|\boldsymbol{w}_j-\widetilde{\boldsymbol{w}}_j|^2\leq C^2$, there are at most C^2/ε^2 indices j_k with $|\boldsymbol{w}_{j_k}-\widetilde{\boldsymbol{w}}_{j_k}|\geq\varepsilon$. Recall from Lemma 89 that

$$\Big|\bigcup_{i=1}^{n}E(\varepsilon,\boldsymbol{x}_i)\Big|\to 0$$

as $\varepsilon\to\infty$. Hence, we find $\varepsilon>0$ such that with probability $1-\delta$, there holds $\mathbf{1}_{\boldsymbol{w}_k\cdot\boldsymbol{x}_i\geq 0}=\mathbf{1}_{\widetilde{\boldsymbol{w}}_k\cdot\boldsymbol{x}_i\geq 0}$ for all $1\leq k\leq m$ and $1\leq i\leq n$ with $i\neq j_k$. This implies

$$|\boldsymbol{y}_i(\boldsymbol{W})-\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})|\leq C^2/\varepsilon^2\quad\text{for all } i=1,\dots,n.$$

Hence, we have

$$\begin{aligned}|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}) - \nabla_{\boldsymbol{W}}\mathcal{L}(\widetilde{\boldsymbol{W}})| &\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|F(\boldsymbol{W},\boldsymbol{x}_i)-y_i||\boldsymbol{y}_i(\boldsymbol{W})-\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})| \\ &\quad + \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|F(\boldsymbol{W},\boldsymbol{x}_i)-F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)||\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})| \\ &\leq \frac{1}{\sqrt{m}}\Big(\sum_{i=1}^{n}|F(\boldsymbol{W},\boldsymbol{x}_i)-y_i|^2\Big)^{1/2}\Big(\sum_{i=1}^{n}|\boldsymbol{y}_i(\boldsymbol{W})-\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})|^2\Big)^{1/2} \\ &\quad + \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|F(\boldsymbol{W},\boldsymbol{x}_i)-F(\widetilde{\boldsymbol{W}},\boldsymbol{x}_i)||\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})| \\ &\lesssim m^{-1/2}\sqrt{\mathcal{L}(\boldsymbol{W})}\frac{C^2}{\varepsilon^2}\sqrt{n} + m^{-1/2}n|\boldsymbol{W}||\boldsymbol{W}-\widetilde{\boldsymbol{W}}|\max_{1\leq i\leq n}|\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})|.\end{aligned}$$

by using the Lipschitz continuity of F (exercise) in the last estimate. Choosing $m \in \mathbb{N}$ sufficiently large and the fact $|\boldsymbol{y}_i(\widetilde{\boldsymbol{W}})| = \sqrt{\sum_{j=1}^{m}\mathbf{1}_{\boldsymbol{w}_j\cdot\boldsymbol{x}_i\geq 0}|\boldsymbol{x}_i|^2} \leq \sqrt{m}$ conclude the proof. □

4.6.3. *Proof of Theorem 87.* With all the preliminary results in the previous sections, the proof of the main result is rather short.

Proof. Recall q, Q from Lemma 96 and assume that the random initialization $\boldsymbol{W}_0 \sim U_s$ is such that the estimates from Lemma 96 and Lemma 97 hold (for sufficiently large $m \in \mathbb{N}$, the probability of this is $1-\delta$). Let $\boldsymbol{W}_\ell$ denote the iterations of Algorithm 4. As long as $|\boldsymbol{W}_\ell - \boldsymbol{W}_0| \leq C$ (with C from Lemma 96 and Lemma 97), Lemma 96 and Lemma 97 show that the assumptions of Lemma 68 are satisfied and hence prove

$$Q^{-1}|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_\ell)|^2 \leq \mathcal{L}(\boldsymbol{W}_\ell) \leq \kappa^\ell \mathcal{L}(\boldsymbol{W}_0).$$

Note that the constants q, Q, ε from Lemma 96&97 are independent of C (only m grows with C). Hence, also the constants α and κ from Lemma 68 are independent of C. By definition of Algorithm 4, there holds $|\boldsymbol{W}_{\ell+1} - \boldsymbol{W}_\ell| \leq \alpha|\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_\ell)|$, and we have

$$|\boldsymbol{W}_\ell - \boldsymbol{W}_0| \leq \alpha\sqrt{Q}\mathcal{L}(\boldsymbol{W}_0)^{1/2}\sum_{j=0}^{\infty}\kappa^{\ell/2}.$$

Thus, setting $C := \sqrt{Q}\mathcal{L}(\boldsymbol{W}_0)^{1/2}\sum_{j=0}^{\infty}\kappa^{\ell/2} < \infty$ concludes the proof. □

Remark 98. *Going through the proof, we notice that the main idea of is the following: Gradient descent works as long as the gradient $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_\ell)$ is sufficiently large. For large $m \in \mathbb{N}$, we show that the difference between $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_0)$ and $\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_\ell)$ is small, i.e., most of the weights do not change much during the application of gradient descent. Essentially, if the the initial gradient is large enough (which happens with high probability), all the other gradients will be large enough too.*

5. High dimensional approximation: Sparse grids

To illustrate the idea, we first look at standard tensor interpolation: For a given set of intervals $\mathcal{T}$, let $\mathcal{Q}^1(\mathcal{T})$ denote the continuous functions which are affine on each interval

in $\mathcal{T}$. Let $I_\ell\colon C([0,1]) \to \mathcal{Q}^1(\{[k2^{-\ell},(k+1)2^{-\ell}] : k=0,\dots,2^\ell-1\})$ denote the nodal interpolation operator in 1D, i.e.,

$$I_\ell v(t_k) = v(t_k) \quad \text{for all } t_k = k2^{-\ell},\ k=0,\dots,2^\ell.$$

We denote with I_ℓ^x that the interpolation operator is applied in dimension x. The approximation on the d-dimensional tensor mesh

$$\mathcal{T}_\ell^\otimes := \{ \prod_{i=1}^d [k_i 2^{-\ell},(k_i+1)2^{-\ell}] : k_1,\dots,k_d \in \{0,\dots,2^\ell-1\}\}$$

is given for $v \in C^0([0,1]^d)$ by

$$(I_\ell^\otimes v)(\boldsymbol{x}) := I_\ell^{x_1}(I_\ell^{x_2}\dots(I_\ell^{x_d}v)\dots)(x_1,\dots,x_d) \in \mathcal{Q}^1(\mathcal{T}_\ell^\otimes),$$

where

$$\mathcal{Q}^1(\mathcal{T}_\ell^\otimes) := \{v \in C^0([0,1]^d) : \forall 1 \le i \le d,\ (x_i \mapsto v(x))|_T \text{ is a polynomial of degree } \le 1\}.$$

Similarly to the proof of the approximation theorem for the standard nodal interpolation operator, one can show

$$\|v - I_\ell^\otimes v\|_{L^\infty([0,1]^d)} \le C2^{-\ell}\|v\|_{C^1([0,1]^d)}$$

for $v \in C^1([0,1]^d)$. As we see, the computation of $I_\ell^\otimes v$ requires the evaluation of $2^{d\ell}$ points in $[0,1]^d$ and hence is impractical for many purposes (if $d=100$, and $\ell=1$, we would need 2^{100} points).

The sparse grid idea is as follows: With the definition $I_{-1}=0$, we may rewrite

$$\begin{aligned}
(I_\ell^\otimes v)(\boldsymbol{x}) &= \sum_{\ell_1=0}^{\ell}(I_{\ell_1}^{x_1} - I_{\ell_1-1}^{x_1})(I_\ell^{x_2}\dots(I_\ell^{x_d}v)\dots)(x_1,\dots,x_d)\\
&= \sum_{\ell_1=0}^{\ell}\sum_{\ell_2=0}^{\ell}(I_{\ell_1}^{x_1} - I_{\ell_1-1}^{x_1})(I_{\ell_2}^{x_2} - I_{\ell_2-1}^{x_2})(I_\ell^{x_3}\dots(I_\ell^{x_d}v)\dots)(x_1,\dots,x_d)\\
&= \sum_{\boldsymbol{\ell}=(\ell_1,\dots,\ell_d)\in\{0,\dots,\ell\}^d} \underbrace{(I_{\ell_1}^{x_1} - I_{\ell_1-1}^{x_1})(I_{\ell_2}^{x_2} - I_{\ell_2-1}^{x_2})\dots(I_{\ell_d}^{x_d} - I_{\ell_d-1}^{x_d})}_{=:\Delta_{\boldsymbol{\ell}}}(v)(\boldsymbol{x}).
\end{aligned}$$

Lemma 99. *For a subset $\mathfrak{u} \subseteq \{1,\dots,d\}$ let $\partial_{\boldsymbol{x}_\mathfrak{u}} := \prod_{i\in\mathfrak{u}}\partial_{x_i}$ denote the partial derivatives in all directions in $\mathfrak{u}$. For sufficiently smooth $v \in C^0([0,1]^d)$, there holds*

$$\|\Delta_{\boldsymbol{\ell}} v\|_{H^1([0,1]^d)} \le 4^d 2^{-|\boldsymbol{\ell}|}\|\partial_{\boldsymbol{x}_\mathfrak{u}} v\|_{H^1([0,1]^d)},$$

where $|\boldsymbol{\ell}| := \ell_1 + \ldots + \ell_d$ and $\mathfrak{u} \subseteq \{1,\dots,d\}$ contains each dimension i with $\ell_i > 0$.

Proof. Let $x_0 = (x_{0,1},\dots,x_{0,d}) \in [0,1]^d$ and $i \in \{1,\dots,d\}$. Choose $k \in \mathbb{N}$ such that $|k2^{-\ell} - x_{0,i}|$ is minimal. Without loss of generality, we assume $x_{0,i} \ge k2^{-\ell}$ (the other case works analogously). Rolle's theorem implies that there exists $\xi \in (k2^{-\ell},(k+1)2^{-\ell})$ with

$$\partial_{x_i}(1 - I_\ell^{x_i})v(x_{0,1},\dots,x_{0,i-1},\xi,x_{0,i+1},\dots,x_{0,d}) = 0.$$

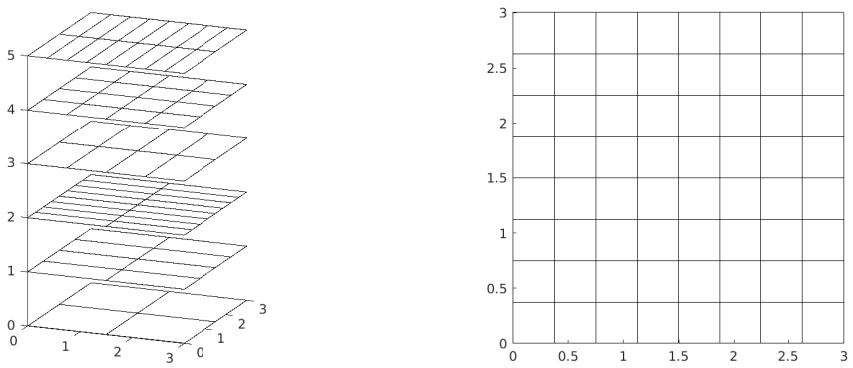

FIGURE 11. The different sparse grid contributions on the left stacked on top of each other combine to the full grid on the right. The interpolation operator $I_{\boldsymbol{\ell}} = I_{\ell_1}^{x_1} I_{\ell_2}^{x_2}$ corresponds to one of the grids on the left-hand side (e.g., grid number 1 for $\boldsymbol{\ell} = (1,2)$ or grid number 5 for $\boldsymbol{\ell} = (3,1)$).

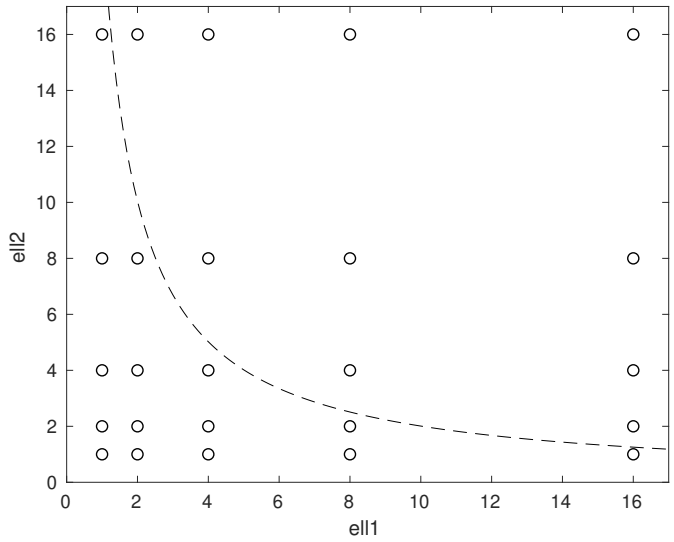

FIGURE 12. The circles represent the number of degrees of freedom of $I_{\boldsymbol{\ell}}$ in each coordinate direction. The sparse grid approach only uses interpolation operators which correspond to circles below the dashed line. This shape is the upper right quadrant of the so-called *hyperbolic cross*.

With this, there holds

$$\begin{aligned}
(1 - I_\ell^{x_i}) v(x_0) &= \int_{k2^{-\ell}}^{x_{0,i}} \partial_{x_i} (1 - I_\ell^{x_i}) v(x_{0,1}, \dots, x_{0,i-1}, z, x_{0,i+1}, \dots, x_{0,d})\, dz \\
&= \int_{k2^{-\ell}}^{x_{0,i}} \int_{\xi}^{z} \partial_{x_i}^2 v(x_{0,1}, \dots, x_{0,i-1}, t, x_{0,i+1}, \dots, x_{0,d})\, dt\, dz \\
&\le |(k+1)2^{-\ell} - k2^{-\ell}|^{3/2} \|\partial_{x_i}^2 v(x_{0,1}, \dots, x_{0,i-1}, \cdot, x_{0,i+1}, \dots, x_{0,d})\|_{L^2([k2^{-\ell}, (k+1)2^{-\ell}])} \\
&\le 2^{-\ell} \|\partial_{x_i}^2 v(x_{0,1}, \dots, x_{0,i-1}, \cdot, x_{0,i+1}, \dots, x_{0,d})\|_{L^2([k2^{-\ell}, (k+1)2^{-\ell}])}.
\end{aligned}$$

We define

$$\Omega_k := \otimes_{j=1}^{i-1} [0,1] \times [k2^{-\ell}, (k+1)2^{-\ell}] \times \otimes_{j=i+1}^{d} [0,1].$$

This results in

$$\begin{aligned}\|(1-I_\ell^{x_i})v\|_{L^2(\Omega_k)}^2 &\le 2^{-3\ell}\int_{\Omega_k}\|\partial_{x_i}^2 v(x_{0,1},\dots,x_{0,i-1},\cdot,x_{0,i+1},\dots,x_{0,d}\|_{L^2(\Omega_k)}^2\,dx_0\\ &\le 2^{-4\ell}\|\partial_{x_i}^2 v\|_{L^2(\Omega_k)}^2.\end{aligned}$$

Since $[0,1]^d = \dot{\bigcup}_{k=0}^{2^\ell-1}\Omega_k$, we obtain

$$\|(1-I_\ell^{x_i})v\|_{L^2([0,1]^d)} \le 2^{-2\ell}\|\partial_{x_i}^2 v\|_{L^2([0,1]^d)}.$$

Analogously, we show

$$\|\nabla(1-I_\ell^{x_i})v\|_{L^2([0,1]^d)} \le 2^{-\ell}\|\partial_{x_i} v\|_{H^1([0,1]^d)}.$$

The triangle inequality concludes

$$\begin{aligned}\|(I_\ell^{x_i}-I_{\ell-1}^{x_i})v\|_{H^1([0,1]^d)} &\le \|(1-I_{\ell-1}^{x_i})v\|_{H^1([0,1]^d)} + \|(1-I_\ell^{x_i})v\|_{H^1([0,1]^d)}\\ &\le 2^{-(\ell-1)}\|\partial_{x_i} v\|_{H^1([0,1]^d)} + 2^{-\ell}\|\partial_{x_i} v\|_{H^1([0,1]^d)}\\ &\le 2^{-\ell+2}\|\partial_{x_i}^2 v\|_{L^2([0,1]^d)}.\end{aligned}$$

Assume that $\ell_i > 0$ for all $1 \le i \le d$. Iteration of this result in all dimension shows

$$\begin{aligned}\|\nabla(I_{\ell_1}^{x_1}-I_{\ell_1-1}^{x_1})\dots(I_{\ell_d}^{x_d}-I_{\ell_d-1}^{x_d})v\|_{L^2([0,1]^d)} &\le 2^{-\ell_1+2}\|\partial_{x_1}^2\nabla_{x_2,\dots,x_d}(I_{\ell_2}^{x_2}-I_{\ell_2-1}^{x_2})\dots(I_{\ell_d}^{x_d}-I_{\ell_d-1}^{x_d})v\|_{L^2([0,1]^d)}\\ &= 2^{-\ell_1+2}\|\nabla_{x_2,\dots,x_d}(I_{\ell_2}^{x_2}-I_{\ell_2-1}^{x_2})\dots(I_{\ell_d}^{x_d}-I_{\ell_d-1}^{x_d})\partial_{x_1}^2 v\|_{L^2([0,1]^d)}\\ &\dots\\ &\le 4^d 2^{-\ell_1-\dots-\ell_d}\|\partial_{x_1}^2\partial_{x_2}^2\dots\partial_{x_d}^2 v\|_{L^2([0,1]^d)}.\end{aligned}$$

The proof for the L^2-norm works analogously. If some of the ℓ_i are zero, we just skip those dimensions in the proof and obtain the stated result. □

With the last result to obtain an error of $2^{-\ell}$, we may ignore all $\Delta_{\boldsymbol{\ell}}$ with $|\boldsymbol{\ell}| > \ell$. This leads to the sparse grid interpolation operator I_ℓ^d defined by

$$I_\ell^d v := \sum_{\substack{\boldsymbol{\ell}\in\{0,\dots,\ell\}^d\\ |\boldsymbol{\ell}|\le\ell}} \Delta_{\boldsymbol{\ell}} v. \tag{36}$$

This truncation is illustrated in Figures 11–12. To analyze the error, we need the following nice combinatorial identity.

Lemma 100. *There holds*

$$\#\big\{\boldsymbol{\ell}\in\mathbb{N}_0^d \,:\, |\boldsymbol{\ell}| = j\big\} = \binom{j+d-1}{d-1}.$$

Proof. There are many proofs of this identity. A nice one goes like this: Imagine the index $\boldsymbol{\ell}\in\mathbb{N}_0^d$ as

$$\underbrace{1\dots1}_{\ell_1}\,|\,\underbrace{1\dots1}_{\ell_2}\,|\,\dots\,\dots\,|\,\underbrace{1\dots1}_{\ell_d}$$

This line contains the $|\ell|+d-1$ symbols $z\in\{1,|\}$. Exactly $d-1$ of the symbols z must satisfy $z = |$. Hence there are $\binom{j+d-1}{d-1}$ possibilities. □

Theorem 101. *The sparse grid interpolation error satisfies*

$$\|(1-I_\ell^d)v\|_{H^1([0,1]^d)} \le C4^d(\ell+d)^{d-1}2^{-\ell}|v|_{H^2_{\mathrm{mix}}([0,1]^d)},$$

where

$$|v|_{H^2_{\mathrm{mix}}([0,1]^d)} := \max_{\mathfrak{u}\subseteq\{1,\dots,d\}} \|\partial^2_{\boldsymbol{x}_\mathfrak{u}} v\|_{L^2([0,1]^d)}.$$

Proof. Given $v \in H^2_{\mathrm{mix}}([0,1]^d)$ we may formally write

$$v = \sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d} \Delta_{\boldsymbol{\ell}} v.$$

As shown in Lemma 99, we have

$$\|\Delta_{\boldsymbol{\ell}} v\|_{H^1([0,1]^d)} \leq 4^d 2^{-|\boldsymbol{\ell}|} \|v\|_{H^2_{\mathrm{mix}}([0,1]^d)}.$$

This implies that the series above converges absolutely and hence we may write the approximation error as

$$v - I^d_\ell v = \sum_{\substack{\boldsymbol{\ell}\in\mathbb{N}_0^d \\ |\boldsymbol{\ell}|>\ell}} \Delta_{\boldsymbol{\ell}} v.$$

Altogether, we have

$$\|v - I^d_\ell v\|_{H^1([0,1]^d)} \leq \sum_{\substack{\boldsymbol{\ell}\in\mathbb{N}_0^d \\ |\boldsymbol{\ell}|>\ell}} \|\Delta_{\boldsymbol{\ell}} v\|_{H^1([0,1]^d)} \leq 4^d \|v\|_{H^2_{\mathrm{mix}}([0,1]^d)} \sum_{\substack{\boldsymbol{\ell}\in\mathbb{N}_0^d \\ |\boldsymbol{\ell}|>\ell}} 2^{-|\boldsymbol{\ell}|}.$$

The sum can be rewritten as

$$\sum_{\substack{\boldsymbol{\ell}\in\mathbb{N}_0^d \\ |\boldsymbol{\ell}|>\ell}} 2^{-|\boldsymbol{\ell}|} = \sum_{j=\ell+1}^{\infty} 2^{-j} \sum_{\substack{\boldsymbol{\ell}\in\mathbb{N}_0^d \\ |\boldsymbol{\ell}|=j}} 1 = \sum_{j=\ell+1}^{\infty} 2^{-j} \binom{j+d-1}{d-1},$$

where we used Lemma 100 for the last identity. There holds for $x \in (0,1)$

$$\begin{aligned}\sum_{j=\ell+1}^{\infty} x^j \binom{j+d-1}{d-1} &= \partial_x^{d-1} \sum_{j=\ell+1}^{\infty} x^{j+d-1}/(d-1)! = \partial_x^{d-1} \frac{x^{\ell+d}}{1-x}/(d-1)! \\ &= \sum_{k=0}^{d-1} \binom{d-1}{k} \partial_x^k x^{\ell+d} \partial_x^{d-1-k} (1-x)^{-1}/(d-1)!\end{aligned}$$

since the series converges absolutely. There holds

$$\begin{aligned}&\binom{d-1}{k} \partial_x^k x^{\ell+d} \partial_x^{d-1-k} (1-x)^{-1}/(d-1)! \\ &= \frac{(d-1)(d-2)\cdots(d-k)}{k!(d-1)!} \big((\ell+d)\cdots(\ell+d-k+1)\big)(d-1-k)! \frac{x^{\ell+d-k}}{(1-x)^{d-k}} \\ &= \frac{1}{k!}\big((\ell+d)\cdots(\ell+d-k+1)\big) \frac{x^{\ell+d-k}}{(1-x)^{d-k}} \leq \frac{(\ell+d)^{d-1}}{k!} \frac{x^{\ell+d-k}}{(1-x)^{d-k}}.\end{aligned}$$

Inserting $x = 1/2$, we end up with

$$\sum_{j=\ell+1}^{\infty} 2^{-j} \binom{j+d-1}{d-1} \lesssim (\ell+d)^{d-1} 2^{-\ell}.$$

This concludes the proof. □

The representation in (36) is not really good for implementation due to cancelation effects and the requirement to constantly transform coefficient vectors between different meshes. A better variant is provided by the inclusion-exclusion formula which is an interesting combinatorial fact in it self.

Lemma 102. *For $d \in \mathbb{N}$ and $r \leq d$, the binomial coefficient satisfies the identity*

$$\sum_{q=0}^{r} (-1)^q \binom{d}{q} = (-1)^r \binom{d-1}{r}.$$

Proof. The proof works by induction. For $r = 0$, there holds $\binom{d}{0} = \binom{d-1}{0} = 1$. Assume the statement holds for $r < d$. Then, we have

$$\sum_{q=0}^{r+1} (-1)^q \binom{d}{q} = (-1)^{r+1} \binom{d}{r+1} + \sum_{q=0}^{r} (-1)^q \binom{d}{q} = (-1)^{r+1} \left(\binom{d}{r+1} - \binom{d-1}{r} \right).$$

The well-known identity

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$$

with $n = d - 1$ and $k = r$ concludes the proof. □

Lemma 103. *With $I_{\boldsymbol{\ell}} := I_{\ell_1}^{x_1} I_{\ell_2}^{x_2} \dots I_{\ell_d}^{x_d}$ for $\boldsymbol{\ell} \in \mathbb{N}_0^d$, there holds*

$$I_\ell^d = \sum_{k=0}^{d} (-1)^k \binom{d-1}{k} \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| = \ell - k}} I_{\boldsymbol{\ell}}.$$

Proof. We rewrite (36) by

$$I_\ell^d = \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| \leq \ell}} \Delta_{\boldsymbol{\ell}} = \sum_{\substack{\boldsymbol{\ell}' \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}'| \leq \ell}} \alpha_{\boldsymbol{\ell}'} I_{\boldsymbol{\ell}'} \tag{37}$$

for some $\alpha_{\boldsymbol{\ell}} \in \mathbb{R}$. Given the definition

$$\Delta_{\boldsymbol{\ell}} = (I_{\ell_1}^{x_1} - I_{\ell_1 - 1}^{x_1})(I_{\ell_2}^{x_2} - I_{\ell_2 - 1}^{x_2}) \dots (I_{\ell_d}^{x_d} - I_{\ell_d - 1}^{x_d})$$

we note that a particular $I_{\boldsymbol{\ell}'}$ appears in (37) if and only if there exists $\boldsymbol{\ell} \in \mathbb{N}_0^d$ with

$$|\boldsymbol{\ell}| \leq \ell \quad \text{and} \quad \ell_i' \leq \ell_i \leq \ell_i' + 1 \quad \text{for all } i = 1, \dots, d. \tag{38}$$

Moreover, the sign of that $I_{\boldsymbol{\ell}'}$ is determined by the parity (odd or even) of the number of dimensions i with $\ell_i = \ell_i' + 1$. For $q = 0, \dots, d$, let

$$P_q(\boldsymbol{\ell}') := \left\{ \boldsymbol{\ell} \in \mathbb{N}_0^d \,:\, \boldsymbol{\ell} \text{ satisfies (38) and } \ell_{i_k} = \ell_{i_k}' + 1,\ k = 1, \dots, q \right\}.$$

Then, we observe $P_q(\boldsymbol{\ell}') = \emptyset$ if $|\boldsymbol{\ell}'| > |\boldsymbol{\ell}| - q$. Moreover, since for each choice of q indices i_k we have an element of $P_q(\boldsymbol{\ell}')$, there holds

$$\# P_q(\boldsymbol{\ell}') = \binom{d}{q}.$$

This implies

$$\alpha_{\boldsymbol{\ell}'} = \sum_{q=0}^{d} (-1)^q \# P_q(\boldsymbol{\ell}') = \sum_{q=0}^{\ell - |\boldsymbol{\ell}'|} (-1)^q \# P_q(\boldsymbol{\ell}') = \sum_{q=0}^{\ell - |\boldsymbol{\ell}'|} (-1)^q \binom{d}{q}.$$

Lemma 102 shows for $r = \ell - |\boldsymbol{\ell}'| \leq d$

$$\sum_{q=0}^{\ell-|\boldsymbol{\ell}'|} (-1)^q \binom{d}{q} = (-1)^{\ell-|\ell'|} \binom{d-1}{\ell-|\ell'|}.$$

Altogether, we see with $k = \ell - |\ell'|$

$$I_\ell^d = \sum_{k=0}^{d} (-1)^k \binom{d-1}{k} \sum_{\substack{\boldsymbol{\ell}' \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}'| = \ell - k}} I_{\boldsymbol{\ell}'}.$$

This concludes the proof. □

Lemma 104. *The number of evaluations of v required for the computation of $I_\ell^d v$ is less than*

$$d\binom{\ell+d-1}{d-1} 2^\ell \leq d(\ell+d)^{d-1} 2^\ell.$$

Proof. We use the representation from Lemma 103. Each $I_{\boldsymbol{\ell}} v$ requires $2^{|\boldsymbol{\ell}|}$ evaluations of v for computation. Lemma 100 concludes the proof. □

The last result together with Theorem 101 shows the following: A sparse grid of size $h > 0$ (this means $2^{-\ell} = h$) allows an interpolation error of

$$\|(1 - I_\ell^d) v\|_{H^1([0,1]^d)} \lesssim (1 + |\log(h)|)^\alpha h$$

for some exponent $\alpha \in \mathbb{N}$ with a cost of computation of $I_\ell^d v$ less than

$$\mathcal{O}\big((1 + |\log(h)|)^\alpha h^{-1}\big).$$

This means that the error estimate with respect to cost reads

$$\|(1 - I_\ell^d) v\|_{H^1([0,1]^d)} \lesssim \mathrm{cost}^{-1}$$

(up to logarithmic factors). The convergence rate is independent of the dimension.

Instead of the sparse interpolation operator, we may also consider the sparse Galerkin projection. Define the (quad-)mesh

$$\mathcal{T}_{\boldsymbol{\ell}}^{\otimes} := \Big\{ \prod_{i=1}^{d} [k_i 2^{-\ell_i}, (k_i+1) 2^{-\ell_i}] \,:\, k_i \in \{0, \dots, 2^{\ell_i} - 1\},\, i = 1, \dots, d \Big\}$$

for $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d) \in \mathbb{N}_0^d$. Note that we don't have a triangle mesh any more. However, the abstract theory just used the fact that

$$\mathcal{X}_\ell = \bigoplus_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| \leq \ell}} \mathcal{Q}^1(\mathcal{T}_{\boldsymbol{\ell}}^{\otimes})$$

is a closed subspace of $H^1([0,1]^d)$. Hence, we may apply all the results of the previous sections.

Theorem 105. *We consider*

$$-\Delta u = f \text{ in } [0,1]^d,$$
$$u = 0 \text{ on } \partial[0,1]^d.$$

Assume that $u \in H^2_{\mathrm{mix}}([0,1]^d)$ and let $u_\ell \in \mathcal{X}_\ell$ denote the unique Galerkin approximation. Then, there holds

$$\|u - u_\ell\|_{H^1([0,1]^d)} \leq C 4^d (\ell+d)^{d-1} 2^{-\ell} \|u\|_{H^2_{\mathrm{mix}}([0,1]^d)}.$$

Proof. Note that $I_\ell u \in \mathcal{Q}^1(\mathcal{T}_\ell^\otimes)$ by definition. This implies that $I_\ell^d u \in \mathcal{X}_\ell$. Thus, the Céa Lemma and Theorem 101 show the statement. □

Analogously to the proof of Lemma 104, we obtain that

$$\dim \mathcal{X}_\ell \lesssim d(\ell+d)^{d-1}2^\ell.$$

There are many examples of high-dimensional PDEs in practical applications such as finance, physics, and chemistry. One notable example is the Schrödinger eigenvalue problem: Given $n \in \mathbb{N}$ electrons and $m \in \mathbb{N}$ nuclei, the goal is to to find the wave function $\psi\colon \mathbb{R}^{3n} \to \mathbb{C}$ which gives a probability density of the position $x_i \in \mathbb{R}^3$ of the i-th electron. The wave function is a solution of the problem

$$-\frac{1}{2}\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{3}\partial_{x_i}^2\psi(x_1,\dots,x_n)}_{\text{Laplace in every dimension } x_i} + \Big(\underbrace{-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Z_j}{|x_i-R_j|^2}}_{\text{force between electrons and nuclei}} + \underbrace{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{1}{|x_i-x_j|^2}}_{\text{force between electrons}} \Big)\psi(x_1,\dots,x_n)$$
$$= E\psi(x_1,\dots,x_n).$$

The position of the nuclei of the atoms is given by $R_j \in \mathbb{R}^3$ and Z_j is the charge of the j-th nucleus. Finally, $E \in \mathbb{C}$ is the eigenvalue of the wave-function ψ. The first part of the operator (the Laplacian) is often abbreviated with T and the remaining part with V. This allows us to write the equation as

$$(T+V)\psi = E\psi.$$

We do not yet know how to solve eigenvalue problems, however in the simplified setting

$$(T+V)\psi = f$$

for some right-hand side f and $Z_j < 0$ for all $j = 1,\dots,m$, we can derive a weak formulation analogously to the previous chapters. (Note that a negative charge is not physical for a nucleous, however, for $Z_j > 0$ one needs Fredholm theory not covered in this lecture to show well-posedness of the weak form.) This results in a problem with $d = 3n$ and hence standard FEM is out of the question even for a moderate number of electrons.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[3] Josef Dick, Frances Y. Kuo, and Ian H. Sloan. High-dimensional integration: The quasi-monte carlo way. *Acta Numerica*, 22:133–288, 2013.

[4] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv:1810.02054*, 2018.

[5] Adrian Ebert, Peter Kritzer, and Dirk Nuyens. Constructing qmc finite element methods for elliptic pdes with random coefficients by a reduced cbc construction. In Bruno Tuffin and Pierre L'Ecuyer, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 183–205, Cham, 2020. Springer International Publishing.

[6] P. Grohs and L. Herrmann. Deep neural network approximation for high-dimensional elliptic pdes with boundary conditions. *arXiv:2007.05384*, 2020.

[7] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[8] Christoph Schwab and Jakob Zech. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl. (Singap.)*, 17(1):19–55, 2019.

[9] L. Ridgway Scott and Shangyou Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.

[10] Onno van Gaans. Probability measures on metric spaces. Technical report, Delft University, 2003.
[11] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103 – 114, 2017.