Found Comput Math (2016) 16:1555–1605
DOI 10.1007/s10208-016-9327-7

FOUNDATIONS OF
COMPUTATIONAL
MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Multi-index Stochastic Collocation Convergence Rates for Random PDEs with Parametric Regularity

Abdul-Lateef Haji-Ali[1] · **Fabio Nobile**[2] ·
Lorenzo Tamellini[3,4] · **Raúl Tempone**[1]

Received: 3 November 2015 / Revised: 8 July 2016 / Accepted: 18 July 2016 /
Published online: 26 August 2016

Abstract We analyze the recent Multi-index Stochastic Collocation (MISC) method for computing statistics of the solution of a partial differential equation (PDE) with random data, where the random coefficient is parametrized by means of a countable sequence of terms in a suitable expansion. MISC is a combination technique based on mixed differences of spatial approximations and quadratures over the space of random data, and naturally, the error analysis uses the joint regularity of the solution with respect to both the variables in the physical domain and parametric variables. In MISC, the number of problem solutions performed at each discretization level is not determined by balancing the spatial and stochastic components of the error, but rather by suitably extending the knapsack-problem approach employed in the con-

Communicated by Albert Cohen.

✉ Raúl Tempone
raul.tempone@kaust.edu.sa

Abdul-Lateef Haji-Ali
abdullateef.hajiali@kaust.edu.sa

Fabio Nobile
fabio.nobile@epfl.ch

Lorenzo Tamellini
tamellini@imati.cnr.it

[1] CEMSE, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[2] MATHICSE-CSQI, Ecole Polytechnique Fédérale de Lausanne, Station 8, CH-1015 Lausanne, Switzerland

[3] Dipartimento di Matematica "F. Casorati", Università di Pavia, Via Ferrata 5, 27100 Pavia, Italy

[4] CNR-IMATI, Via Ferrata 1, 27100 Pavia, Italy

struction of the quasi-optimal sparse-grids and Multi-index Monte Carlo methods, i.e., we use a greedy optimization procedure to select the most effective mixed differences to include in the MISC estimator. We apply our theoretical estimates to a linear elliptic PDE in which the log-diffusion coefficient is modeled as a random field, with a covariance similar to a Matérn model, whose realizations have spatial regularity determined by a scalar parameter. We conduct a complexity analysis based on a summability argument showing algebraic rates of convergence with respect to the overall computational work. The rate of convergence depends on the smoothness parameter, the physical dimensionality and the efficiency of the linear solver. Numerical experiments show the effectiveness of MISC in this infinite dimensional setting compared with the Multi-index Monte Carlo method and compare the convergence rate against the rates predicted in our theoretical analysis.

Keywords Multi-level · Multi-index Stochastic Collocation · Infinite dimensional integration · Elliptic partial differential equations with random coefficients · Finite element method · Uncertainty quantification · Random partial differential equations · Multivariate approximation · Sparse grids · Stochastic Collocation methods · Multi-level methods · Combination technique

Mathematics Subject Classification 41A10 (approx by polynomials) · 65C20 (models, numerical methods) · 65N30 (Finite elements) · 65N05 (Finite differences)

1 Introduction

In this work, we analyze and apply the recent MISC method [22] to the approximation of quantities of interest (outputs) from the solutions of linear elliptic partial differential equations (PDEs) with random coefficients. Such equations arise in many applications in which the coefficients of the PDE are described in terms of random variables/fields due either to a lack of knowledge of the system or to its inherent non-predictability. We focus on the weak approximation of the solution of the following linear elliptic $\boldsymbol{y}$-parametric problem:

$$\begin{cases} -\operatorname{div}(a(\boldsymbol{x},\boldsymbol{y})\,\nabla u(\boldsymbol{x},\boldsymbol{y})) = \varsigma(\boldsymbol{x}) & \text{in } \mathcal{B} \\ u(\boldsymbol{x},\boldsymbol{y}) = 0 & \text{on } \partial\mathcal{B}. \end{cases} \tag{1}$$

Here, $\mathcal{B} \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ denotes the "physical domain," and the operators div and ∇ act with respect to the physical variable, $\boldsymbol{x} \in \mathcal{B}$, only. We assume that $\mathcal{B}$ has a tensor structure, i.e., $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2 \times \cdots \times \mathcal{B}_D$, with $D \in \mathbb{N}$, $\mathcal{B}_i \subset \mathbb{R}^{d_i}$ and $\sum_{i=1}^{D} d_i = d$, see, e.g., Fig. 1a, b. This assumption simplifies the analysis detailed in the following, although MISC can be applied to more general domains, such as

- domains obtained by mapping from a reference tensor domain as in Fig. 1c, by suitably extending the approaces in [17,28];

- non-tensor domains that can be immersed in a tensor bounding box, $\mathcal{B} \subset \hat{\mathcal{B}} = \hat{\mathcal{B}}_1 \times \hat{\mathcal{B}}_2 \times \cdots \times \hat{\mathcal{B}}_D$, as in Fig. 1d, whose mesh is obtained as a tensor product of meshes on each component, $\hat{\mathcal{B}}_i$, of the bounding box;
- domains that admit a structured mesh, i.e., with a regular connectivity, whose level of refinement in each "direction" can be set independently, as in Fig. 1e;
- domains that can be decomposed in patches satisfying any of the conditions above (observe that the meshes on each patch need not be conforming).

The parameter $\boldsymbol{y} = \{y_j\}_{j\geq 1}$ in (1) is a random sequence whose components are independent and uniformly distributed random variables. More precisely, each y_j has support in $[-1, 1]$ with measure $\frac{\mathrm{d}\lambda}{2}$, where $\mathrm{d}\lambda$ is the standard Lebesgue measure. We further define $\Gamma = \times_{j\geq 1}[-1, 1]$ (hereafter referred to as the "stochastic domain" or the "parameter space"), with the cylindrical probability measure $\mathrm{d}\mu = \times_{j\geq 1} \frac{\mathrm{d}\lambda}{2}$, (see, e.g., [5, Chapter 3, Section 5]).

The right-hand side of (1), namely the deterministic function ς, does not play a central role in this work, and it is assumed to be a smooth function of class $C_0^\infty(\overline{\mathcal{B}})$, where $\overline{\mathcal{B}}$ denotes the closure of $\mathcal{B}$. This regularity requirement can be relaxed, but we keep it to ease the presentation, since our main goal in this work is to track the effect of the regularity of the coefficient a in (1) on the MISC convergence rate. Here, we focus on the following family of diffusion coefficients:

$$a(\boldsymbol{x}, \boldsymbol{y}) = e^{\kappa(\boldsymbol{x},\boldsymbol{y})}, \text{ with } \kappa(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j\geq 1} \psi_j(\boldsymbol{x}) y_j, \tag{2}$$

where $\{\psi_j\}_{j\geq 1}$ is a sequence of functions $\psi_j \in C^t(\overline{\mathcal{B}})$ for $t \geq 0$ such that $\left\|\psi_j\right\|_{L^\infty(\mathcal{B})} \to 0$ as $j \to \infty$. Hereafter, without loss of generality, we assume that the sequence $\{\left\|\psi_j\right\|_{L^\infty(\mathcal{B})}\}_{j\geq 1}$ is ordered in decreasing order. Thanks to a straightforward application of the Lax–Milgram lemma, the well-posedness of (1) in the classical Sobolev space, $V = H_0^1(\mathcal{B})$, is guaranteed almost surely (a.s.) in Γ if two functions, $a_{\min}, a_{\max} : \Gamma \to \mathbb{R}$, exist such that

$$0 < a_{\min}(\boldsymbol{y}) \leq a(\boldsymbol{x}, \boldsymbol{y}) \leq a_{\max}(\boldsymbol{y}) < \infty, \quad \forall \boldsymbol{x} \in \mathcal{B}, \quad \text{a.s. in } \Gamma. \tag{3}$$

Moreover, the equation is well posed in the Bochner space, $L^q(\Gamma; V)$, for some $q \geq 1$,[1] (see [1,8] and the following discussion), provided that sufficiently high moments of the functions $1/a_{\min}$ and $a_{\max}$ are bounded. The goal of our computation is the approximation of an expected value,

$$\mathbb{E}[F] = \mathbb{E}[\Theta(u)] \in \mathbb{R},$$

where Θ is a deterministic bounded and linear functional, and $F(\boldsymbol{y}) = \Theta(u(\cdot, \boldsymbol{y}))$ is a real-valued random variable, $F : \Gamma \to \mathbb{R}$. To this end, we utilize the Multi-index

[1] Recall that, given $q \geq 1$, $L^q(\Gamma; V) = \left\{v : \Gamma \to V \text{ strongly measurable, such that } \int_\Gamma \|u\|_V^q \, \mathrm{d}\mu < \infty\right\}$.

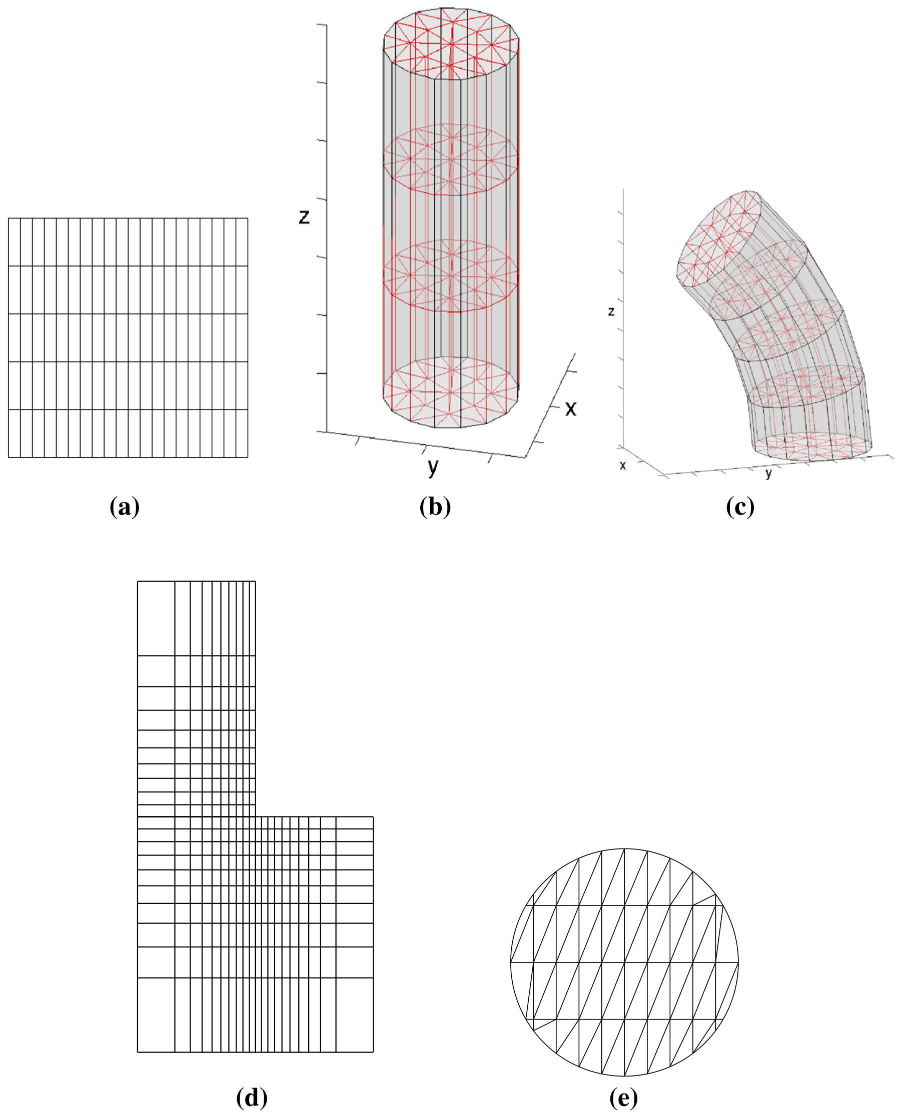


Fig. 1 Examples of physical domains on which MISC can be applied: **a** and **b** are within the framework of this work, while treating (**c**) requires the introduction of a mapping from (**b**). MISC can also be formulated in non-tensor domains as in (**d**) and (**e**), but extending the analysis of the present work to this case is less straightforward and out of the scope of this work

Stochastic Collocation (MISC) method, which we have introduced in a general setting in a previous work [22].

In MISC, we consider a decomposition in terms of tensorized univariate details (i.e., a tensorized hierarchical decomposition), for both the discrete space in which (1) is solved for a fixed value of $\boldsymbol{y} \in \Gamma$ and for the quadrature operator used to approximate the expected value of F, relying on the well-established theory of sparse-grid approximation of PDEs on the one hand [6,7,21,26,41] and of sparse-grid quadrature on the other hand [1,6,15,34,35,40]. We use tensor products of such univariate details, obtaining combined deterministic-stochastic, first-order mixed differences to build the MISC estimator of $\mathbb{E}[F]$ by selecting the most effective mixed differences with an optimization approach inspired by the literature on the knapsack problem (see, e.g., [30]). The knapsack approach also was used in [33] to obtain the so-called quasi-optimal sparse grids for PDEs with stochastic coefficients and in [6,20] in the context of sparse-grid resolution of high-dimensional PDEs.

The resulting method can be seen as an extension of the sparse-grid combination technique for PDEs with stochastic coefficients, as well as a fully sparse, non-randomized version of the Multi-level Monte Carlo method [2,9,16,27]. In particular, MISC differs from other works in the literature that attempt to optimally combine spatial and stochastic resolution levels [4,25,29,37,38] in two aspects. First, MISC uses combined deterministic-stochastic, first-order differences, which allows us to exploit not only the regularity of the solution with respect to the spatial variables and the stochastic parameters, but also the mixed deterministic-stochastic regularity whenever available. Second, the MISC estimator is built upon an optimization procedure, whereas the above-mentioned works try to balance the error contributions arising from the deterministic and stochastic components of the method without taking into account the corresponding costs. Finally, MISC can also be seen as a sparse-grid quadrature version of the Multi-index Monte Carlo method that was proposed and analyzed in [23].

In [22], MISC was introduced in a general setting and we restricted the analysis to the case of problems of type (1) depending on a *finite* number of random variables, $\boldsymbol{y} \in \Gamma \subset \mathbb{R}^N$, with $N < \infty$. Here, we provide a complexity analysis of MISC in the more challenging case in which the diffusion coefficient a depends on a countable sequence of random variables, $\{y_j\}_{j\geq 1}$. Furthermore, we aim at tracking the dependence of the MISC converge rate on the smoothness of the realizations of a. This new framework requires that the tools used to prove the complexity of the method be changed: while in [22] we used a "direct counting" argument, i.e., we derived a complexity estimate by explicitly summing the work and the error contributions associated with each mixed difference included in the MISC estimator, here instead we base our proof on a summability argument and on suitable interpolation estimates in mixed regularity spaces. We mention that in [14] an infinite dimensional analysis based on a direct counting argument was recently carried out in the case of hyperbolic cross-type index sets that might arise when quasi-optimizing the work contribution of sparse grid stochastic collocation without spatial discretization.

The rest of this work is organized as follows. Section 2 introduces suitable assumptions and a class of random diffusion coefficients that we consider throughout the work; functional analysis results that are needed for the subsequent analysis of the MISC method are also provided. The MISC method is reviewed in Sect. 3. A complexity analysis of MISC with an infinite number of random variables is carried out in Sect. 4, where we provide a general convergence theorem. In Sect. 5, we discuss the application of MISC to the specific class of diffusion coefficients that we consider here and track the dependence of the convergence rate on the regularity of the diffusion coefficient. Section 6 presents some numerical tests to verify the convergence analysis conducted in the previous section. Finally, Sect. 7 provides some conclusions and final remarks. In the Appendix, we include some technical results on the summability and regularity properties of certain random fields written in terms of their series expansion.

In the following, $\mathbb{N}$ denotes the set of integer numbers including zero, while $\mathbb{N}_+$ denotes the set of positive integer numbers excluding zero. We refer to sequences in $\mathbb{N}^{\mathbb{N}_+}$ and $\mathbb{N}_+^{\mathbb{N}_+}$ as "multi-indices." Moreover, we often use a vector notation for sequences, i.e., we formally treat sequences as vectors in $\mathbb{N}^{\mathbb{N}_+}$ (or $\mathbb{R}^{\mathbb{N}_+}$) and mark them with bold type. We employ the following notation, with the understanding that $N < \infty$ for actual vectors and $N = \infty$ for sequences:

- $\mathbf{1}$ denotes a vector in $\mathbb{N}^N$ whose components are all equal to one;
- $\boldsymbol{e}_\ell^N$ denotes the ℓ-th canonical vector in $\mathbb{R}^N$, i.e., $\left(\boldsymbol{e}_\ell^N\right)_i = 1$ if $\ell = i$ and zero otherwise; however, for the sake of readability, we often omit the superscript N whenever it is obvious from context. For instance, if $\boldsymbol{v} \in \mathbb{R}^N$, we write $\boldsymbol{v} - \boldsymbol{e}_1$ instead of $\boldsymbol{v} - \boldsymbol{e}_1^N$;
- given $\boldsymbol{v} \in \mathbb{R}^N$, $|\boldsymbol{v}| = \sum_{i=1}^N v_i$, $|\boldsymbol{v}|_0$ denotes the number of nonzero components of $\boldsymbol{v}$, $\max(\boldsymbol{v}) = \max_{i=1,\dots N} v_i$ and $\min(\boldsymbol{v}) = \min_{i=1,\dots N} v_i$;
- $\mathfrak{L}_+$ denotes the set of sequences with positive components with only finitely many elements larger than 1, i.e., $\mathfrak{L}_+ = \{\boldsymbol{p} \in \mathbb{N}_+^{\mathbb{N}_+} : |\boldsymbol{p} - \mathbf{1}|_0 < \infty\}$;
- given $\boldsymbol{v} \in \mathbb{R}^N$ and $f : \mathbb{R} \to \mathbb{R}$, $f(\boldsymbol{v})$ denotes the vector obtained by applying f to each component of $\boldsymbol{v}$, $f(\boldsymbol{v}) = [f(v_1), f(v_2), \dots, f(v_N)] \in \mathbb{R}^N$;
- given $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^N$, the inequality $\boldsymbol{v} > \boldsymbol{w}$ holds true if and only if $v_i > w_i$ $\forall i = 1, \dots, N$;
- given $\boldsymbol{v} \in \mathbb{R}^D$ and $\boldsymbol{w} \in \mathbb{R}^N$, we denote their concatenation by $[\boldsymbol{v}, \boldsymbol{w}] = (v_1, \dots, v_D, w_1, \dots, w_N) \in \mathbb{R}^{D+N}$;
- given a set with finite cardinality, $\mathcal{G} \subset \mathbb{N}_+$, we define the set $\mathbb{N}^{\mathcal{G}} = \{\boldsymbol{z} \in \mathbb{N}^{\mathbb{N}_+} : z_j = 0,\ \forall j \notin \mathcal{G}\}$. We similarly define $\mathbb{R}^{\mathcal{G}}$ and $\mathbb{C}^{\mathcal{G}}$.

2 Functional Setting

Even though condition (3) ensures a.s. well-posedness of (1) in V, we need to make sure that realizations of u a.s. belong to more regular spaces to prove a convergence rate result for MISC. More specifically, due to the classic spatial sparse-grid approximation theory, we need certain conditions on the mixed derivatives of u with respect to the physical coordinates. To this end, we introduce suitable functional spaces (tensor products of fractional Sobolev spaces, see also [19]) and then a "shift regularity" assumption (see Assumption A1), i.e., we assume that the regularity of the realizations of u is induced "in a natural way" by the regularity of a, of the forcing, ς, and of the smoothness of the physical domain, $\mathcal{B}$. In other words, we rule out "pathological"/"ad hoc" examples in which u is very regular despite that the data are not, e.g., when the forcing is chosen such that $u \in C^q$ for $q > 2$ even in the presence of a domain with corners. First, recall the definition of a fractional Sobolev space for $l_i \in \mathbb{R}_+ \setminus \mathbb{N}_+$ and $\mathcal{B}_i \subset \mathbb{N}^{d_i}$:

$$H^{l_i}(\mathcal{B}_i) = \left\{ u \in H^{\lfloor l_i \rfloor}(\mathcal{B}_i) : \sup_{\boldsymbol{\alpha} \in \mathbb{N}^{d_i}, |\boldsymbol{\alpha}| = \lfloor l_i \rfloor} \int_{\mathcal{B}_i} \int_{\mathcal{B}_i} \frac{|D^{\boldsymbol{\alpha}} u(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} u(\boldsymbol{x}')|^2}{|\boldsymbol{x} - \boldsymbol{x}'|^{d_i + 2(l_i - \lfloor l_i \rfloor)}} d\boldsymbol{x} d\boldsymbol{x}' < \infty \right\},$$

extending the definition of a standard Sobolev space $H^{l_i}(\mathcal{B}_i)$ for l_i integer. The tensorized fractional Sobolev space can then be defined as

$$H^{\boldsymbol{l}}(\mathcal{B}) = H^{l_1}(\mathcal{B}_1) \otimes \cdots \otimes H^{l_D}(\mathcal{B}_D)$$

for $\boldsymbol{l} = (l_i)_{i=1}^D \in \mathbb{R}_+^D$.[2] Finally, the mixed fractional Sobolev spaces, that we will need for our analysis, can be defined for each $\boldsymbol{q} \in \mathbb{R}_+^D$ as

$$\mathcal{H}^{\mathbf{1}+\boldsymbol{q}}(\mathcal{B}) = \bigcap_{j=1}^{D} H^{\boldsymbol{e}_j + \boldsymbol{q}}(\mathcal{B}).$$

Observe that, while mixed fractional spaces, $\mathcal{H}^{\mathbf{1}+\boldsymbol{q}}(\mathcal{B})$, are the proper setting for the forthcoming analysis, we will, for ease of presentation, not look for the most general mixed space in which the solution lives. Instead, we will be content with deducing mixed regularity from inclusions in standard Sobolev spaces, to the point that Assumption A1 will be written in terms of standard Sobolev spaces. For this, we observe that $\mathcal{H}^{\mathbf{1}}(\mathcal{B}) = H^1(\mathcal{B})$ holds and in general we have the following inclusion result between standard and mixed fractional Sobolev spaces:

$$u \in H^{1+r}(\mathcal{B}) \Rightarrow u \in \mathcal{H}^{\mathbf{1}+r\boldsymbol{q}}(\mathcal{B}) \quad \text{for } r \in (0, \infty) \quad \text{and} \quad 0 < |\boldsymbol{q}| \leq 1. \tag{4}$$

Before stating precisely the shift-regularity assumption on u, we need some more notation and setup. First, observe that this assumption needs to be stated in the complex domain, for reasons that will be made clear later. We therefore extend the diffusion coefficient from $a(\cdot, \boldsymbol{y})$ with $\boldsymbol{y} \in \Gamma$ to $a(\cdot, \boldsymbol{z})$ with $\boldsymbol{z} \in \mathbb{C}^{\mathbb{N}_+}$, so that the corresponding solution of (1), $u(\cdot, \boldsymbol{z})$, becomes a $H_0^1(\mathcal{B})$ function taking values in $\mathbb{C}$, i.e., $u(\cdot, \boldsymbol{z}) \in H_0^1(\mathcal{B}, \mathbb{C})$. Since the complex-valued version of problem (1) is well posed as long as there exists δ such that $\mathfrak{Re}\,[a(\boldsymbol{x}, \boldsymbol{z})] > \delta > 0$ for almost every (a.e.) $\boldsymbol{x} \in \mathcal{B}$, and since our approximation method will cover the countable set of parameters $\boldsymbol{z} \in \mathbb{C}^{\mathbb{N}_+}$ by multiple subsets of finite cardinality, we define the following region in $\mathbb{C}^{\mathcal{G}}$ for a set of finite cardinality, $\mathcal{G} \subset \mathbb{N}_+$:

$$\Sigma_{\mathcal{G},\delta} = \left\{ \boldsymbol{z} \in \mathbb{C}^{\mathcal{G}} : \mathfrak{Re}\,[a(\boldsymbol{x}, \boldsymbol{z})] \geq \delta > 0 \text{ for a.e. } \boldsymbol{x} \in \mathcal{B} \right\}. \tag{5}$$

[2] We recall that $H^{\boldsymbol{l}}(\mathcal{B})$ is the completion of formal sums $v = \sum_{k=1}^{K} v_{1,k} v_{2,k} \cdots v_{D,k}$ with $v_{i,k} \in H^{l_i}(\mathcal{B}_i)$ with respect to the norm induced by the inner product

$$(v, w)_{H^{\boldsymbol{l}}(\mathcal{B})} = \sum_{k,i} (v_{1,k}, w_{1,i})_{H^{l_1}(\mathcal{B}_1)} (v_{2,k}, w_{2,i})_{H^{l_2}(\mathcal{B}_2)} \cdots (v_{D,k}, w_{D,i})_{H^{l_D}(\mathcal{B}_D)}.$$

We are now ready to state the assumption on the link between the regularity of the coefficient, a, and the regularity of solution, u.

Assumption A1 (Shift assumption) For a given $\mathcal{B}$, let $\psi_j \in C^t(\overline{\mathcal{B}})$ (cf. eq. (2)) and $\varsigma \in C_0^{\infty}(\overline{\mathcal{B}})$. We assume that there exists r such that $1 < r < t$ and that, for any finite set $\mathcal{G} \subset \mathbb{N}_+$ and any $\boldsymbol{z} \in \Sigma_{\mathcal{G},\delta}$, the three following conditions hold:

1. $u(\cdot, \boldsymbol{z}) \in H^{1+r}(\mathcal{B}, \mathbb{C}) \cap H_0^1(\mathcal{B}, \mathbb{C})$;
2. $\frac{Du(\cdot,z_j)}{Dz_j} \in H^{1+r}(\mathcal{B}, \mathbb{C}), \quad \forall j \in \mathcal{G}$, where $\frac{Du(\cdot,z_j)}{Dz_j}$ denotes the partial complex derivative of u;
3. $\|u(\cdot, \boldsymbol{z})\|_{H^{1+s}(\mathcal{B},\mathbb{C})} \leq C(\delta, s, \varsigma, \mathcal{B}) \, \|a(\cdot, \boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})}$, with $C(\delta, s, \varsigma, \mathcal{B}) \to \infty$ for $\delta \to 0$, for every $s = 1, \dots, \lfloor r \rfloor$.

In the following, we will need to ensure that $\|u(\cdot, \boldsymbol{z})\|_{H^{1+s}(\mathcal{B},\mathbb{C})}$, for $s > 0$, is uniformely bounded for all $\boldsymbol{z}$ in certain subregions of the complex plane. Note that this is a stronger condition than what is stated in the previous assumption, where we only assumed pointwise control on the norms of u (i.e., we gave a bound that depends on $\boldsymbol{z}$). In particular, we show at the end of this section that the possibility of having such a uniform bound depends on certain summability properties of the diffusion coefficient. Toward this end, we state the following assumption, which also guarantees the well-posedness of Problem (1).

Assumption A2 (Summability of the diffusion coefficient) For every $s = 0, 1, \dots, s_{\max} \leq r$, define the sequences $\boldsymbol{b}_s = \{b_{s,j}\}_{j \geq 1}$ where

$$b_{s,j} = \max_{\boldsymbol{s} \in \mathbb{N}^d : |\boldsymbol{s}| \leq s} \left\| D^{\boldsymbol{s}} \psi_j \right\|_{L^{\infty}(\mathcal{B})}, \quad j \geq 1. \tag{6}$$

We assume that an increasing sequence $\{p_s\}_{s=0}^{s_{\max}}$ exists such that $0 < p_0 \leq \cdots \leq p_{s_{\max}} < \frac{1}{2}$ and $\boldsymbol{b}_s \in \ell^{p_s}$, i.e.,

$$\|\boldsymbol{b}_s\|_{\ell^{p_s}}^{p_s} = \sum_{j \geq 1} b_{s,j}^{p_s} < \infty. \tag{7}$$

We observe that with the above assumption, $b_{s,j} \to 0^+$ as $j \to \infty$ and $0 \leq b_{0,j} \leq b_{s,j}$ for every $s = 0, 1, \dots, s_{\max}$. Moreover, given Assumption A2, we have that $\boldsymbol{b}_0 \in \ell^1$, which, together with the fact that $y_j \in [-1, 1]$ for $j \geq 1$, guarantees that condition (3) holds and therefore that (1) is well posed in V a.s. in Γ. Incidentally, we observe that the conditions in Assumption A2 are sufficient but not necessary for condition (3) to hold: Indeed, one would only need $\boldsymbol{b}_s \in \ell^2$ for some integer $s \geq 1$, see Lemma 15 (and Corollary 16 for a specific example) in the Appendix.

As suggested above, the fact that, for a fixed s, the sequence $\boldsymbol{b}_s$ is p_s-summable plays a central role in this work. Indeed, if $\boldsymbol{b}_s$ is p_s-summable, we show that $\|u(\cdot, \boldsymbol{z})\|_{H^{1+s}(\mathcal{B},\mathbb{C})}$ is uniformely bounded with respect to $\boldsymbol{z}$ in a region of the complex plane whose size is proportional to $\|\boldsymbol{b}_s\|_{\ell^{p_s}}^{p_s}$. We use this fact to show convergence of the MISC method, with the convergence rate dictated by both p_0 and p_s. In Theorem 10, we detail how to optimally choose the value of s in the range $0, 1, \dots, s_{\max}$, which is the main result of this work. Restricting the range of values of s by $p_s < \frac{1}{2}$ is not crucial; we could relax this to $p_s < 1$. However, we follow this more stringent assumption because it considerably simplifies some technical steps in the following discussion without affecting the main part of the proof, as we make clear below (see Remark 4). What is important is that $s_{\max}$ might be strictly smaller than $\lfloor r \rfloor$ (i.e., it could happen that $\boldsymbol{b}_r$ is p_r-summable but with $p_r > \frac{1}{2}$, or not summable at all); in this case, the line of proof we propose does not fully exploit the regularity of the solution, u.

Example 1 In the numerical section of this work, we consider either $\mathcal{B} = [0, 1]$, i.e., $d = D = d_1 = 1$, or $\mathcal{B} = [0, 1]^3$, i.e., $d = D = 3$, $d_i = 1$ and $\mathcal{B}_i = [0, 1]$ for $i = 1, 2, 3$. In both cases, we consider the following form for $\kappa(\boldsymbol{x}, \boldsymbol{y})$:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} A_{\boldsymbol{k}} \sum_{\boldsymbol{\ell}\in\{0,1\}^d} y_{\boldsymbol{k},\boldsymbol{\ell}} \prod_{i=1}^{d} \left(\cos\left(\pi k_i x_i\right)\right)^{\ell_i} \left(\sin\left(\pi k_i x_i\right)\right)^{1-\ell_i}. \tag{8}$$

Observe that it is possible to write κ in the form of (2) using a bijective mapping from $\{y_{\boldsymbol{k},\boldsymbol{\ell}}\}_{\boldsymbol{k}\in\mathbb{N}^d,\boldsymbol{\ell}\in\{0,1\}^d}$ to $\{y_j\}_{j\geq 1}$. We also choose the following values for the $A_{\boldsymbol{k}}$ coefficients:

$$A_{\boldsymbol{k}}= \left(\sqrt{3}\right) 2^{\frac{|\boldsymbol{k}|_0}{2}} (1 + |\boldsymbol{k}|^2)^{-\frac{\nu+d/2}{2}}, \tag{9}$$

for some $\nu > 0$. We observe that ν is a parameter dictating the $\boldsymbol{x}$-regularity of the realizations of κ, hence of a. Moreover, the parameters ν and d govern the p_s-summability of the sequence $\boldsymbol{b}_s$ for any s and, as a consequence, the overall convergence of the MISC method, as discussed earlier. Section 5 analyzes the summability properties of the series (8).

We conclude this preliminary section by making the shape of the above-mentioned regions in the complex plane more precise and showing how their sizes depend on the summability properties of a. In particular, we will exploit the fact that for any finite set $\mathcal{G} \subset \mathbb{N}_+$, for every $s = 0, 1, \ldots, s_{\max}$ and for any $\boldsymbol{z} \in \mathbb{C}^{\mathcal{G}}$ we have $\kappa(\cdot, \boldsymbol{z}) \in C^s(\overline{\mathcal{B}}, \mathbb{C})$, $\|\kappa(\cdot, \boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})} \leq \sum_{j\in\mathcal{G}} |z_j| b_{s,j}$ and infer, from the multivariate Faà di Bruno formula (see "Appendix 1" and [13]), that $a(\cdot, \boldsymbol{z}) \in C^s(\overline{\mathcal{B}}, \mathbb{C})$ as well, with the estimate

$$\|a(\cdot, \boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})} \leq \frac{s!}{(\log 2)^s} \|a(\cdot, \boldsymbol{z})\|_{C^0(\overline{\mathcal{B}},\mathbb{C})} \left(1 + \|\kappa(\cdot, \boldsymbol{z})\|_{C^s(\overline{\mathcal{B}})}\right)^s, \quad \forall \boldsymbol{z} \in \mathbb{C}^{\mathcal{G}}. \tag{10}$$

Next, for a given $\zeta > 1$, let $\mathcal{E}_\zeta$ denote the polyellipse in the complex plane

$$\mathcal{E}_\zeta = \left\{ z \in \mathbb{C} : \ \mathfrak{Re}\,[z] \leq \frac{\zeta + \zeta^{-1}}{2} \cos\vartheta, \ \ \mathfrak{Im}\,[z] \leq \frac{\zeta - \zeta^{-1}}{2} \sin\vartheta, \ \ \vartheta \in [0, 2\pi) \right\}.$$

For any sequence $\boldsymbol{\zeta} = \{\zeta_j\}_{j\geq 1}$ with $\zeta_j > 1$ for every $j \geq 1$ and for any finite set, $\mathcal{G} \subset \mathbb{N}_+$, we introduce the Bernstein polyellipse:

$$\mathcal{E}_{\boldsymbol{\zeta}}^{\mathcal{G}} = \{\boldsymbol{z} \in \mathbb{C}^{\mathcal{G}} : z_j \in \mathcal{E}_{\zeta_j} \text{ for all } j \in \mathcal{G}\}. \tag{11}$$

Lemma 1 (Holomorphic complex continuation of u in $H^1_0(\mathcal{B};\mathbb{C})$ in a Bernstein polyellipse) *Consider the sequence $\boldsymbol{b}_0$ defined in* (6)*. For any $\delta > 0$, let $E_\delta > 2$ be such that*

$$\frac{\pi}{E_\delta} = -\left\|\boldsymbol{b}_0\right\|_{\ell^1} - \log\delta + \log\cos\left(\frac{\pi}{E_\delta}\right),$$

and consider the sequence $\boldsymbol{\zeta}_0 = \{\zeta_{0,j}\}_{j\geq 1}$, with

$$\zeta_{0,j} = \tau_{0,j} + \sqrt{\tau_{0,j}^2 + 1} > 1 \tag{12}$$

$$\tau_{0,j} = \frac{\pi}{E_\delta}\frac{(b_{0,j})^{p_0-1}}{\|\boldsymbol{b}_0\|_{\ell^{p_0}}^{p_0}}, \tag{13}$$

with p_0 as in (7)*. Then, for any finite set $\mathcal{G} \subset \mathbb{N}_+$, the solution, u, admits a holomorphic complex continuation, $u : \mathbb{C}^{\mathcal{G}} \to H^1_0(\mathcal{B},\mathbb{C})$, in the Bernstein polyellipse, $\mathcal{E}^{\mathcal{G}}_{\boldsymbol{\zeta}_0} \subset \Sigma_{\mathcal{G},\delta}$, with*

$$\sup_{z\in\mathcal{E}^{\mathcal{G}}_{\boldsymbol{\zeta}_0}} \|u(\cdot,\boldsymbol{z})\|_{H^1(\mathcal{B})} \leq C_{0,u} = \frac{\|\varsigma\|_{H^{-1}(\mathcal{B})}}{\delta} < \infty,$$

with $C_{0,u}$ independent of $\mathcal{G}$.

Proof It is well known in the literature that $u : \mathbb{C}^{\mathcal{G}} \to H^1_0(\mathcal{B},\mathbb{C})$ is holomorphic in the region $\Sigma_{\mathcal{G},\delta}$ defined in (5) (see, e.g., [1]). To compute the parameters $\{\zeta_j\}_{j\in\mathcal{G}}$ of a Bernstein polyellipse contained in $\Sigma_{\mathcal{G},\delta}$, we rewrite $a(\boldsymbol{x},\boldsymbol{z})$ as

$$\begin{aligned}
a(\boldsymbol{x},\boldsymbol{z}) &= \exp\left(\sum_{j\in\mathcal{G}} z_j\psi_j(\boldsymbol{x})\right)\\
&= \exp\left(\sum_{j\in\mathcal{G}} \mathfrak{Re}\left[z_j\right]\psi_j(\boldsymbol{x})\right)\exp\left(\sum_{j\in\mathcal{G}} i\,\mathfrak{Im}\left[z_j\right]\psi_j(\boldsymbol{x})\right)\\
&= \exp\left(\sum_{j\in\mathcal{G}} \mathfrak{Re}\left[z_j\right]\psi_j(\boldsymbol{x})\right)\left[\cos\left(\sum_{j\in\mathcal{G}} \mathfrak{Im}\left[z_j\right]\psi_j(\boldsymbol{x})\right)\right.\\
&\quad\left. + i\sin\left(\sum_{j\in\mathcal{G}} \mathfrak{Im}\left[z_j\right]\psi_j(\boldsymbol{x})\right)\right],
\end{aligned}$$

so that $\Sigma_{\mathcal{G},\delta}$ can be rewritten as

$$\Sigma_{\mathcal{G},\delta} = \left\{ \boldsymbol{z} \in \mathbb{C}^{\mathcal{G}} : \exp\left(\sum_{j\in\mathcal{G}} \mathfrak{Re}\left[z_j\right]\psi_j(\boldsymbol{x}) \right) \cos\left(\sum_{j\in\mathcal{G}} \mathfrak{Im}\left[z_j\right]\psi_j(\boldsymbol{x}) \right) \right.$$
$$\left. \geq \delta \;\; \text{for a.e. } \boldsymbol{x} \in \mathcal{B} \right\}.$$

Now, for some $E > 2$ that we choose in the following, the two following conditions on $\boldsymbol{z}$ imply that $\boldsymbol{z} \in \Sigma_{\mathcal{G},\delta}$:

$$\begin{cases} \cos\left(\sum_{j\in\mathcal{G}} |\mathfrak{Im}\left[z_j\right]| \, b_{0,j} \right) \geq \cos\left(\dfrac{\pi}{E}\right) \\ \exp\left(-\sum_{j\in\mathcal{G}} |\mathfrak{Re}\left[z_j\right]| \, b_{0,j} \right) \geq \dfrac{\delta}{\cos\left(\frac{\pi}{E}\right)}; \end{cases}$$

equivalently, we write

$$\begin{cases} \sum_{j\in\mathcal{G}} |\mathfrak{Im}\left[z_j\right]| \, b_{0,j} \leq \dfrac{\pi}{E} \\ \sum_{j\in\mathcal{G}} |\mathfrak{Re}\left[z_j\right]| \, b_{0,j} \leq -\log\delta + \log\cos\left(\dfrac{\pi}{E}\right). \end{cases}$$

For a fixed value of E, the equations above define a second region, $\Omega_{\mathcal{G},\delta}$, included in $\Sigma_{\mathcal{G},\delta}$. In turn, the previous conditions are verified if the following conditions, which define a hyper-rectangular region, $\mathcal{R}_\delta \subset \Omega_{\mathcal{G},\delta}$, are verified:

$$\begin{cases} |\mathfrak{Im}\left[z_j\right]| \leq \tau_{0,j} = \dfrac{\pi (b_{0,j})^{p_0-1}}{E \, \|\boldsymbol{b}_0\|_{\ell^{p_0}}^{p_0}}, \\ |\mathfrak{Re}\left[z_j\right]| \leq 1 + w_{0,j}, \quad \text{with } w_{0,j} = \dfrac{(b_{0,j})^{p_0-1}}{\|\boldsymbol{b}_0\|_{\ell^{p_0}}^{p_0}} \left(-\|\boldsymbol{b}_0\|_{\ell^1} - \log\delta + \log\cos\left(\dfrac{\pi}{E}\right) \right), \end{cases}$$

provided that δ and E are such that the quantity $-\|\boldsymbol{b}_0\|_{\ell^1} - \log\delta + \log\cos\left(\frac{\pi}{E}\right)$ remains positive. Observe that for sufficiently small $\delta > 0$ such E exists, since $f(E) = \log\cos\left(\frac{\pi}{E}\right)$ is a monotonically increasing function, with $f(E) \to -\infty$ for $E \to 2$ and $f(E) \to 0$ for $E \to \infty$, and $-\log\delta$ is positive for sufficiently small δ. In particular, for any $\delta > 0$, we choose $E = E_\delta$ such that $w_{0,j} = \tau_{0,j}$, which leads to

$$\frac{\pi}{E_\delta} = -\|\boldsymbol{b}_0\|_{\ell^1} - \log\delta + \log\cos\left(\frac{\pi}{E_\delta}\right).$$

We observe that with this choice, $\tau_{0,j}$ (and hence $w_{0,j}$) actually does not depend on $\mathcal{G}$, therefore we can define the sequence $\boldsymbol{\tau}_0 = \{\tau_{0,j}\}_{j\geq 1}$.

We are now in the position to compute the Bernstein polyellipses that touch the boundary of $\mathcal{R}_\delta$ on the real and imaginary axes. For the real axis, we have to enforce

$$\frac{\zeta_{j,\text{real}} + \zeta_{j,\text{real}}^{-1}}{2} = 1 + \tau_{0,j} \Rightarrow \zeta_{j,\text{real}} = 1 + \tau_{0,j} + \sqrt{(1+\tau_{0,j})^2 - 1},$$

while for the imaginary axis we have to enforce

$$\frac{\zeta_{j,\text{imag}} - \zeta_{j,\text{imag}}^{-1}}{2} = \tau_{0,j} \Rightarrow \zeta_{j,\text{imag}} = \tau_{0,j} + \sqrt{\tau_{0,j}^2 + 1}.$$

The proof is concluded by observing that $\zeta_{j,\text{imag}} \leq \zeta_{j,\text{real}}$, i.e., the only polyellipse entirely contained in $\mathcal{R}_\delta$, and hence in $\Sigma_{\mathcal{G},\delta}$, is the one touching $\mathcal{R}_\delta$ on the imaginary axis, which also implies that the bound $\sup_{z\in\mathcal{E}_{\zeta}^{\mathcal{G}}} \|u(\cdot,z)\|_{H^1(\mathcal{B})} \leq C_{0,u} = \frac{\|\varsigma\|_{H^{-1}(\mathcal{B})}}{\delta} < \infty$ holds independently of $\mathcal{G}$. □

Lemma 2 (Holomorphic complex continuation of u in $H^{1+s}(\mathcal{B};\mathbb{C})$ in a Bernstein polyellipse) *For a given $s = 1, 2, \dots, s_{\max}$, let $\boldsymbol{\zeta}_s = \{\zeta_{s,j}\}_{j\geq 1}$, with*

$$\zeta_{s,j} = \tau_{s,j} + \sqrt{\tau_{s,j}^2 + 1} > 1, \tag{14}$$

$$\tau_{s,j} = \frac{\pi (b_{s,j})^{p_s - 1}}{E_\delta \|\boldsymbol{b}_s\|_{\ell^{p_s}}^{p_s}}, \tag{15}$$

with $\boldsymbol{b}_s$ as in (6), p_s as in (7), and E_δ as in Lemma 1. For any finite set $\mathcal{G} \subset \mathbb{N}_+$, $u : \mathbb{C}^{\mathcal{G}} \to H^{1+s}(\mathcal{B},\mathbb{C})$ is holomorphic in the Bernstein polyellipse $\mathcal{E}_{\boldsymbol{\zeta}_s}^{\mathcal{G}} \subset \Sigma_{\mathcal{G},\tilde{\delta}}$, with

$$\sup_{z\in\mathcal{E}_{\zeta_s}} \|u(\cdot,z)\|_{H^{1+s}(\mathcal{B})} \leq C_{s,u} = C(\tilde{\delta}, s, \varsigma, \mathcal{B}) M < \infty, \tag{16}$$

where $M = \frac{s!}{(\log 2)^s} e^{K\frac{\pi}{E_\delta}} \left(1 + K\frac{\pi}{E_\delta}\right)^s$, $K = \left(2 + \frac{1}{\min_{j\geq 1}\tau_{s,j}^2}\right)^{1/2}$, $\tilde{\delta} = e^{-K\frac{\pi}{E_\delta}}$, *$C(\tilde{\delta}, s, \varsigma, \mathcal{B})$ as in Assumption A1, and $C_{s,u}$ independent of $\mathcal{G}$.*

Proof From Assumption A1, $u : \mathbb{C}^{\mathcal{G}} \to H^{1+s}(\mathcal{B},\mathbb{C})$ is complex differentiable for every z in $\Sigma_{\mathcal{G},\varepsilon}$ for any $\varepsilon > 0$. It is therefore holomorphic in $\Sigma_{\mathcal{G},\varepsilon}$. Similarly to the previous lemma, we look for a region in which we have an a-priori bound on the $H^{1+s}(\mathcal{B},\mathbb{C})$ norm of u uniformly on z. Again from Assumption A1, we have that this is true in the region

$$\Xi_{\mathcal{G},\varepsilon}(M) = \{z \in \mathbb{C}^{\mathcal{G}} : \|a(\cdot,z)\|_{C^s(\overline{\mathcal{B}})} \leq M\} \cap \Sigma_{\mathcal{G},\varepsilon} \quad \text{for any } \varepsilon > 0.$$

However, contrary to the previous lemma, in this proof, we do not derive the expression of a polyellipse contained in $\Xi_{\mathcal{G},\varepsilon}(M)$, but content ourselves with verifying that the polyellipses, $\mathcal{E}_{\mathcal{G},\boldsymbol{\zeta}_s}$, proposed in the statement of the lemma (that we have obtained simply by replacing $b_{0,j}$ with $b_{s,j}$ in (13)) satisfy the requirement, i.e., $\mathcal{E}_{\mathcal{G},\boldsymbol{\zeta}_s} \subset \Xi_{\mathcal{G},\tilde{\delta}}(M)$, for every finite set, $\mathcal{G} \subset \mathbb{N}_+$, and for a certain $\tilde{\delta}$ that we specify later to control the coercivity of the problem. To this end, let us consider the univariate polyellipse $\mathcal{E}_{\zeta_{s,j}}$. We first prove that this polyellipse is contained in the following complex rectangle:

$$\mathcal{R}_j = \{z \in \mathbb{C} : |\mathfrak{Re}\,[z]| \leq \sqrt{1+\tau_{s,j}^2},\ \ |\mathfrak{Im}\,[z]\,| \leq \tau_{s,j}\}.$$

The bound on the imaginary part of z is a consequence of the choice of the polyellipse in (14), similarly to what was done in Lemma 1. For the real part, we compute the point z_0 where the polyellipse intersects the real axis by equating

$$\begin{aligned} z_0 &= \frac{\zeta_{s,j} + \frac{1}{\zeta_{s,j}}}{2} = \frac{\zeta_{s,j}^2 + 1}{2\zeta_{s,j}} = \frac{\tau_{s,j}^2 + 1 + \tau_{s,j}\sqrt{\tau_{s,j}^2+1}}{\tau_{s,j} + \sqrt{\tau_{s,j}^2+1}} \\ &= \left(\tau_{s,j}^2 + 1 + \tau_{s,j}\sqrt{\tau_{s,j}^2+1}\right)\left(\sqrt{\tau_{s,j}^2+1} - \tau_{s,j}\right) = \sqrt{1+\tau_{s,j}^2}. \end{aligned}$$

Furthermore, we observe that $|z| \leq \sqrt{1+2\tau_{s,j}^2} \leq K\tau_{s,j}$ for every $z \in \mathcal{R}_j$ and some $K > 0$; for instance, we could look for the smallest $\tau_{s,j}$, say τ_{s,j^*}, choose K accordingly, i.e., such that $(K^2-2)\tau_{s,j^*}^2 \geq 1$, and obtain the value in the statement of the lemma. Next, according to (10) and Assumption A2,

$$\begin{aligned} \|a(\cdot,\boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})} &\leq \tfrac{s!}{(\log 2)^s}\, \|a(\cdot,\boldsymbol{z})\|_{C^0(\overline{\mathcal{B}},\mathbb{C})} \left(1 + \|\kappa(\cdot,\boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})}\right)^s \\ &\leq \quad \tfrac{s!}{(\log 2)^s} e^{\sum_{j\in\mathcal{G}} b_{0,j}|z_j|} \left(1 + \textstyle\sum_{j\in\mathcal{G}} b_{s,j}|z_j|\right)^s \end{aligned}$$

holds. We finish the proof by observing that for every, $\boldsymbol{z} \in \mathcal{E}_{\mathcal{G},\boldsymbol{\zeta}_s}$, we have

$$\sum_{j\in\mathcal{G}} b_{0,j}|z_j| \leq \sum_{j\in\mathcal{G}} b_{s,j}|z_j| \leq K\sum_{j\in\mathcal{G}} b_{s,j}\tau_{s,j} = K\frac{\pi}{E_\delta}\sum_{j\in\mathcal{G}} b_{s,j}\frac{(b_{s,j})^{p_s-1}}{\|\boldsymbol{b}_s\|_{\ell^{p_s}}^{p_s}} \leq K\frac{\pi}{E_\delta},$$

which gives uniform control of the norm of $\|a(\cdot,\boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})}$ within $\mathcal{E}_{\boldsymbol{\zeta}_s}^{\mathcal{G}}$ as required. More precisely, we have

$$\|a(\cdot,\boldsymbol{z})\|_{C^s(\overline{\mathcal{B}},\mathbb{C})} \leq M = \frac{s!}{(\log 2)^s} e^{K\frac{\pi}{E_\delta}} \left(1 + K\frac{\pi}{E_\delta}\right)^s, \quad \forall \boldsymbol{z} \in \mathcal{E}_{\mathcal{G},\boldsymbol{\zeta}_s},$$

which together with Assumption A1 gives the desired bound on $\|u(\cdot,\boldsymbol{z})\|_{H^{1+s}(\mathcal{B})}$ in (16) and

$$\mathfrak{Re}\,[a(\boldsymbol{x},\boldsymbol{z})] \geq e^{-K\frac{\pi}{E_\delta}} =: \tilde{\delta} > 0.$$

□

The following result from [22,33] is also needed. Since this result is concerned with the finite-dimensional case, i.e., $\mathcal{G} = \{1, 2, \ldots, N\}$ and $\boldsymbol{\zeta} \in \mathbb{R}^N$, we write, for ease of notation, $\mathcal{E}_{\boldsymbol{\zeta}}$ instead of $\mathcal{E}_{\boldsymbol{\zeta}}^{\mathcal{G}}$, i.e., $\mathcal{E}_{\boldsymbol{\zeta}} = \{\boldsymbol{z} \in \mathbb{C}^N : z_j \in \mathcal{E}_{\zeta_j} \text{ for } j = 1, 2, \ldots, N\}$.

Lemma 3 (Chebyshev expansion of a holomorphic function) *Given* $q_j \in \mathbb{N}$, *let* ϕ_{q_j} *be the family of Chebyshev polynomials of the first kind on* $[-1, 1]$, *i.e.,* $|\phi_{q_j}(y)| \leq 1$ *for all* $y \in [-1, 1]$, *and, for* $N \in \mathbb{N}_+$ *and any* $\boldsymbol{p} \in \mathbb{N}^N$, *let* $\Phi_{\boldsymbol{p}}(\boldsymbol{y}) = \prod_{j=1}^N \phi_{p_j}(y_j)$. *If* $f : [-1, 1]^N \to \mathbb{R}$ *admits a holomorphic complex extension in a Bernstein polyellipse,* $\mathcal{E}_{\boldsymbol{\zeta}}$, *for some* $\boldsymbol{\zeta} \in (1, \infty)^N$ *and if there exists* $0 < C_f < \infty$ *such that* $\sup_{z \in \mathcal{E}_{\boldsymbol{\zeta}}} |f(z)| \leq C_f$, *then* f *admits the following Chebyshev expansion:*

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{p} \in \mathbb{N}^N} f_{\boldsymbol{p}} \Phi_{\boldsymbol{p}}(\boldsymbol{y}),$$

$$f_{\boldsymbol{p}} = \frac{1}{\int_{[-1,1]^N} \Phi_{\boldsymbol{p}}^2(\boldsymbol{y}) \varrho_C(\boldsymbol{y}) d\boldsymbol{y}} \int_{[-1,1]^N} f(\boldsymbol{y}) \Phi_{\boldsymbol{p}}(\boldsymbol{y}) \varrho_C(\boldsymbol{y}) d\boldsymbol{y},$$

$$\varrho_C(\boldsymbol{y}) = \prod_{j=1}^N \left(\sqrt{1 - y_j^2}\right)^{-1},$$

which converges uniformely in $\mathcal{E}_{\boldsymbol{\zeta}}$. *Moreover the following bound on the coefficients* $f_{\boldsymbol{p}}$ *holds:*

$$|f_{\boldsymbol{p}}| \leq \sup_{z \in \mathcal{E}_{\boldsymbol{\zeta}}} |f(z)| 2^{|\boldsymbol{p}|_0} \prod_{j=1}^N \zeta_j^{-p_j},$$

where $|\boldsymbol{p}|_0$ *denotes the number of nonzero elements of* $\boldsymbol{p}$.

3 The Multi-index Stochastic Collocation Method

In this section, we introduce approximations of $\mathbb{E}[F]$ along the deterministic and stochastic dimensions and their decomposition in terms of tensorizations of univariate difference operators. We then recall the so-called mixed difference operators and the construction of the MISC estimator that was first introduced in [22] in a general setting.

3.1 Approximation Along the Deterministic and Stochastic Variables

Tensorized deterministic solver. Let $\{\mathbb{T}_i\}_{i=1}^D$ be the triangulations/meshes of each of the subdomains $\{\mathcal{B}_i\}_{i=1}^D$ composing the domain $\mathcal{B}$; denote by $\{h_i\}_{i=1}^D$ the mesh size on each mesh $\mathbb{T}_i$; and let $\bigotimes_{i=1}^D \mathbb{T}_i$ be the mesh for $\mathcal{B}$. Then, consider a numerical method for the approximation of the solution of (1) for a fixed value of the random variables, $\boldsymbol{y}$, based on such a mesh, e.g., finite differences, finite volumes, tensorized finite elements, or h-refined splines, such as those used in the isogeometric context. The values of h_i are actually given as functions of a positive integer value, $\alpha \geq 1$, referred to as a "deterministic discretization level", i.e., $h_i = h_i(\alpha)$. Observe that, in a more general setting, the mesh size needs not be a constant value over the subdomain $\mathcal{B}_i$ and could be for instance the result of a grading function intended to refine subregions of $\mathcal{B}_i$ as in Fig. 1d (see also [24, Remark 2.2] for further comments on locally refined meshes in the context of Multi-Level Monte Carlo methods). In this work, we restrict ourselves to constant h for ease of presentation. Given a multi-index, $\boldsymbol{\alpha} \in \mathbb{N}_+^D$, we denote by $u^{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})$ the approximation of u obtained by setting $h_i = h_i(\alpha_i)$ and use notation $F^{\boldsymbol{\alpha}}(\boldsymbol{y}) = \Theta[u^{\boldsymbol{\alpha}}(\cdot, \boldsymbol{y})]$. More specifically, in the following we will consider

$$h_i = h_{0,i} 2^{-\alpha_i}, \quad \text{for } i = 1, \ldots, D \tag{17}$$

and a method obtained by tensorizing piecewise multi-linear finite element spaces on each mesh, $\{\mathbb{T}_i\}_{i=1}^D$, discretizing each $\{\mathcal{B}_i\}_{i=1}^D$.

As already mentioned in the previous section, MISC could also be applied to more general domains, such as those discussed in Fig. 1, as long as some kind of “tensor structure” can be induced from the shape of the domain to the solver of the deterministic problem and the vector $\boldsymbol{\alpha}$ determines the refinement level of each component of such a tensor structure. The reason why we need such tensor structure will be made clear when we introduce the classic sparse-grids approach to solve the problem. For non-tensorial domains, we can always set $D = 1$ and consider an unstructured mesh for the whole domain, $\mathcal{B}$, having only one discretization level $\alpha \in \mathbb{N}_+$. In this way, we give up the sparse-grid approach on the deterministic part of the problem and obtain a variant of the Multi-Level Stochastic Collocation method discussed in [37,38], yet with a different algorithm for combining spatial and stochastic discretizations. See Remark 1 stated next and [22] for additional discussion on this aspect.

It would be straightforward to extend this setting to discretization methods based on degree elevation rather than on mesh refinement, such as spectral methods, p-refined finite elements or p- and k-refined splines. However, here we limit ourselves to the setting defined above. It would also be possible to include time-dependent problems in this framework, but in this case we might need to take care of possible constraints on discretization parameters, such as CFL conditions; a broader generalization could also include “non-physical” parameters such as tolerances for numerical solvers. Finally, more general problems, e.g., those depending on random variables with probability distributions other than uniform distributions or with uncertain boundary conditions and/or forcing terms could also be addressed with suitable modifications of the MISC methodology.

Tensorized quadrature formulae for expected value approximation. Similarly to what was presented for the deterministic problem, we base our approximation of the expected value of $F^{\boldsymbol{\alpha}}(\boldsymbol{y})$ on a tensorization of quadrature formulae over the stochastic domain, Γ.

Let $C^0([-1, 1])$ be the set of real-valued continuous functions over $[-1, 1]$, $\beta \in \mathbb{N}_+$ be referred to as a “stochastic discretization level”, and $m : \mathbb{N} \to \mathbb{N}$ be a strictly increasing function with $m(0) = 0$ and $m(1) = 1$, that we call a “level-to-nodes function”. At level β, we consider a set of $m(\beta)$ distinct quadrature points in $[-1, 1]$, $\mathcal{P}^{m(\beta)} = \{y_\beta^1, y_\beta^2, \ldots, y_\beta^{m(\beta)}\} \subset [-1, 1]$, and a set of quadrature weights, $\mathcal{W}^{m(\beta)} = \{\varpi_\beta^1, \varpi_\beta^2, \ldots, \varpi_\beta^{m(\beta)}\}$. We then define the quadrature operator as

$$Q^{m(\beta)} : C^0([-1,1]) \to \mathbb{R}, \quad Q^{m(\beta)}[f] = \sum_{j=1}^{m(\beta)} f(y_\beta^j)\varpi_\beta^j. \tag{18}$$

The quadrature weights are selected such that $Q^{m(\beta)}[y^k] = \int_{-1}^{1} \frac{y^k}{2} dy, \quad \forall k = 0, 1, \ldots, m(\beta) - 1$. The quadrature points are chosen to optimize the convergence properties of the quadrature error (the specific choice of quadrature points is discussed later in this section). In particular, for symmetry reasons, we define the trivial operator $Q^1[f] = f(0), \ \forall f \in C^0([-1,1])$.

Defining a quadrature operator over Γ is more delicate, since Γ is defined as a countable tensor product of intervals. To this end, we follow [36] and define, for any finitely supported multi-index $\boldsymbol{\beta} \in \mathfrak{L}_+$,

$$\mathcal{Q}^{m(\boldsymbol{\beta})} : C^0(\Gamma) \to \mathbb{R}, \quad \mathcal{Q}^{m(\boldsymbol{\beta})} = \bigotimes_{j \geq 1} Q^{m(\beta_j)},$$

where the j-th quadrature operator is understood to act only on the j-th variable, and the tensor product is well defined since it is composed of finitely many non-trivial factors (see [36] again). In practice, the value of $\mathcal{Q}^{m(\boldsymbol{\beta})}[f]$ can be obtained by considering the tensor grid $\mathcal{T}^{m(\boldsymbol{\beta})} = \times_{j \geq 1} \mathcal{P}^{m(\beta_j)}$ with cardinality $\#\mathcal{T}^{m(\boldsymbol{\beta})} = \prod_{j \geq 1} m(\beta_j)$ and computing

$$\mathcal{Q}^{m(\boldsymbol{\beta})}[f] = \sum_{j=1}^{\#\mathcal{T}^{m(\boldsymbol{\beta})}} f(\widehat{\boldsymbol{y}}_j) w_j,$$

where $\widehat{\boldsymbol{y}}_j \in \mathcal{T}^{m(\boldsymbol{\beta})}$ and w_j are (infinite) products of weights of the univariate quadrature rules. Notice that having $m(1) = 1$ is essential in this construction so that the cardinality of $\mathcal{T}^{m(\boldsymbol{\beta})}$ is finite for any $\boldsymbol{\beta} \in \mathfrak{L}_+$ and $\varpi_{\beta_j}^1 = 1$ whenever $\beta_j = 1$. All weights, w_j, are thus bounded.

Coming back to the choice of the univariate quadrature points, it is recommended, for optimal performance, that they are chosen according to the underlying measure, $\mathrm{d}\lambda/2$. Moreover, since we aim at a hierarchical decomposition of the operator, $\mathcal{Q}^{m(\boldsymbol{\beta})}$, it is useful (although not necessary, see e.g., [33]) that the nodes be *nested* collocation points, i.e., $\mathcal{P}^{m(\beta)} \subset \mathcal{P}^{m(\beta+1)}$ for any $\beta \geq 1$. Thus, we consider Clenshaw–Curtis points that are defined as

$$y_\beta^j = \cos\left(\frac{(j-1)\pi}{m(\beta)-1}\right), \quad 1 \leq j \leq m(\beta). \tag{19}$$

Clenshaw–Curtis points are nested provided that the level-to-nodes function is defined as

$$m(0) = 0, \ m(1) = 1, \ m(\beta) = 2^{\beta-1} + 1. \tag{20}$$

We close this section by mentioning that another family of nested points for uniform measures available in the literature is the Leja points, whose performance is equivalent to that of Clenshaw–Curtis points for quadrature purposes. See, e.g., [10,31,32,36] and references therein for definitions and comparison.

3.2 Construction of the MISC Estimator

It is straightforward to see that a direct approximation, $\mathbb{E}[F] \approx \mathcal{Q}^{m(\boldsymbol{\beta})}[F^{\boldsymbol{\alpha}}]$, is not a viable option in practical cases, due to the well-known "curse of dimensionality" effect. In [22], we proposed to use MISC as a computational strategy to combine spatial and stochastic discretizations in such a way as to obtain an effective approximation scheme for $\mathbb{E}[F]$.

MISC is based on a classic sparsification approach in which approximations like $\mathcal{Q}^{m(\boldsymbol{\beta})}[F^{\boldsymbol{\alpha}}]$ are decomposed in a hierarchy of operators. Only the most important of these operators are retained in the approximation. In more detail, let us denote for brevity $\mathcal{Q}^{m(\boldsymbol{\beta})}[F^{\boldsymbol{\alpha}}] = F_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and introduce the first-order difference operators for the deterministic and stochastic discretization operators, denoted, respectively, by Δ_i^{det} with $1 \le i \le D$ and Δ_j^{stoc} with $j \ge 1$:

$$\Delta_i^{\text{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \begin{cases} F_{\boldsymbol{\alpha},\boldsymbol{\beta}} - F_{\boldsymbol{\alpha}-\boldsymbol{e}_i,\boldsymbol{\beta}}, & \text{if } \alpha_i > 1, \\ F_{\boldsymbol{\alpha},\boldsymbol{\beta}} & \text{if } \alpha_i = 1, \end{cases}$$

$$\Delta_j^{\text{stoc}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \begin{cases} F_{\boldsymbol{\alpha},\boldsymbol{\beta}} - F_{\boldsymbol{\alpha},\boldsymbol{\beta}-\boldsymbol{e}_j}, & \text{if } \beta_j > 1, \\ F_{\boldsymbol{\alpha},\boldsymbol{\beta}} & \text{if } \beta_j = 1. \end{cases}$$

As a second step, we define the so-called mixed difference operators,

$$\begin{aligned} \boldsymbol{\Delta}^{\text{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] &= \bigotimes_{i=1}^{D} \Delta_i^{\text{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \Delta_1^{\text{det}}\left[\Delta_2^{\text{det}}\left[\cdots \Delta_D^{\text{det}}\left[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}\right]\right]\right] \\ &= \sum_{\boldsymbol{i}\in\{0,1\}^D} (-1)^{|\boldsymbol{i}|} F_{\boldsymbol{\alpha}-\boldsymbol{i},\boldsymbol{\beta}}, \end{aligned} \tag{21}$$

$$\boldsymbol{\Delta}^{\text{stoc}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \bigotimes_{j\ge 1} \Delta_j^{\text{stoc}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \sum_{\boldsymbol{j}\in\{0,1\}^{\mathbb{N}_+}} (-1)^{|\boldsymbol{j}|} F_{\boldsymbol{\alpha},\boldsymbol{\beta}-\boldsymbol{j}}, \tag{22}$$

with the convention that $F_{\boldsymbol{v},\boldsymbol{w}} = 0$ whenever a component of $\boldsymbol{v}$ or $\boldsymbol{w}$ is zero. Notice that, since $\boldsymbol{\beta}$ has finitely many components larger than 1, the sum on the right-hand side of (22) contains only a finite number of terms. Finally, letting

$$\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \boldsymbol{\Delta}^{\text{stoc}}[\boldsymbol{\Delta}^{\text{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]], \tag{23}$$

we define the Multi-index Stochastic Collocation (MISC) estimator of $\mathbb{E}[F]$ as

$$\mathcal{M}_{\mathcal{I}}[F] = \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathcal{I}} \boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}] = \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathcal{I}} c_{\boldsymbol{\alpha},\boldsymbol{\beta}} F_{\boldsymbol{\alpha},\boldsymbol{\beta}}, \quad c_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \sum_{\substack{[\boldsymbol{i},\boldsymbol{j}]\in\{0,1\}^{D+\mathbb{N}} \\ [\boldsymbol{\alpha}+\boldsymbol{i},\boldsymbol{\beta}+\boldsymbol{j}]\in\mathcal{I}}} (-1)^{\|[\boldsymbol{i},\boldsymbol{j}]\|_0}, \tag{24}$$

where $\mathcal{I} \subset \mathbb{N}_+^D \times \mathfrak{L}_+$. The second form of the MISC estimator is known as the "combination technique", since it expresses the MISC approximation as a linear combination of a number of tensor approximations, $F_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, and might be useful for the practical implementation of the method; we observe in particular that many of its coefficients, $c_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, are zero.

The effectiveness of MISC crucially depends on the choice of the index set, $\mathcal{I}$. Given the hierarchical structure of MISC, a natural requirement is that $\mathcal{I}$ should be downward-closed, i.e.,

$$\forall\, [\boldsymbol{\alpha}, \boldsymbol{\beta}] \in \mathcal{I}, \quad \begin{cases} [\boldsymbol{\alpha} - \boldsymbol{e}_i, \boldsymbol{\beta}] \in \mathcal{I} \text{ for all } 1 \le i \le D \text{ such that } \alpha_i > 1, \\ [\boldsymbol{\alpha}, \boldsymbol{\beta} - \boldsymbol{e}_j] \in \mathcal{I} \text{ for all } j \ge 1 \text{ such that } \beta_j > 1 \end{cases}$$

(see also [6,33,39]). In addition to this general constraint, in [22] we have detailed a procedure to derive an efficient set, $\mathcal{I}$, based on an optimization technique inspired by the Dantzig algorithm for the approximate solution of the knapsack problem (see [30]). In the following, we briefly summarize this procedure and refer to [22] as well as to [3,6,33] for a thorough discussion on the similarities between this procedure and the Dantzig algorithm.

The first step of our optimized construction consists of introducing the "work contribution," $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, and "error contribution", $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, for each operator, $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$. The work contribution measures the computational cost (measured, e.g., as a function of the total number of degrees of freedom, or in terms of computational time) required to add $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$ to $\mathcal{M}_{\mathcal{I}}[F]$, i.e.,

$$\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \text{Work}\left[\mathcal{M}_{\mathcal{I}\cup\{[\boldsymbol{\alpha},\boldsymbol{\beta}]\}}\right] - \text{Work}\left[\mathcal{M}_{\mathcal{I}}\right] = \text{Work}\left[\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right]. \tag{25}$$

Similarly, the error contribution measures how much the error, $|\mathbb{E}[F]-\mathcal{M}_{\mathcal{I}}[F]|$, would decrease if the operator $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$ were added to $\mathcal{M}_{\mathcal{I}}[F]$,

$$\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \left|\mathcal{M}_{\mathcal{I}\cup\{[\boldsymbol{\alpha},\boldsymbol{\beta}]\}}[F] - \mathcal{M}_{\mathcal{I}}[F]\right| = \left|\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right|. \tag{26}$$

We observe that the following decompositions of the total work and error of the MISC estimator hold:

$$\text{Work}\left[\mathcal{M}_{\mathcal{I}}\right] = \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathcal{I}} \Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}, \tag{27}$$

$$|\mathbb{E}[F] - \mathcal{M}_{\mathcal{I}}[F]| = \left|\sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\notin\mathcal{I}} \boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right| \le \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\notin\mathcal{I}} \left|\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right| = \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\notin\mathcal{I}} \Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}. \tag{28}$$

Although it would be tempting to define $\mathcal{I}$ as the set of couples $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ with the largest error contribution, this choice could be far from optimal in terms of computational cost. As suggested in the literature on the knapsack problem (see [30]), the benefit-to-cost ratio should be taken into account in the decision (see also [3,6,20,22,33]). More precisely, we propose to build the MISC estimator by first assessing the so-called profit of each operator $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$, i.e., the quantity

$$P_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \frac{\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}}{\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}}.$$

Then, we build an index set for the MISC estimator:

$$\mathcal{I} = \mathcal{I}(\epsilon) = \left\{ [\boldsymbol{\alpha}, \boldsymbol{\beta}] \in \mathbb{N}_+^D \times \mathfrak{L}_+ \; : \; P_{\boldsymbol{\alpha},\boldsymbol{\beta}} \geq \epsilon \right\}, \tag{29}$$

for a suitable $\epsilon > 0$. We observe that the obtained set is not necessarily downward-closed; we have to enforce this condition a posteriori. Obviously, $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ are not, in general, at our disposal. In practice, we base the construction of the MISC estimator on a-priori bounds for such quantities. More precisely, we derive a-priori ansatzes for these bounds from theoretical considerations and then fit the constants appearing in the ansatzes with some auxiliary computations. We refer to the entire strategy as a priori/a posteriori.

Remark 1 We remark that the general form of the MISC estimator (24) is quite broad and includes other related methods (i.e., methods that combine different spatial and stochastic discretization levels to optimize the computational effort) available in the literature, such as the "Multi-Level Stochastic Collocation" method [37,38] and the "Sparse Composite Collocation" method [4]; see also [25]. The main novelty of the MISC estimator (24)-(29) with respect to such methods is the profit-oriented selection of difference operators. Another difference from [37,38] is the fact that difference operators in our approach are introduced in both the spatial and stochastic domains. See also [4] for a similar construction, in which no optimization is performed. More details on the comparison between the above-mentioned methods and MISC can be found in [22].

4 Error Analysis of the MISC Method

In this section, we state and prove a convergence theorem for the profit-based MISC estimator based on the multi-index set (29). The theorem is based on a result from the previous work [33], which was proved in the context of sparse-grid approximation of Hilbert-space-valued functions. Since the sparse grid and the MISC constructions are identical, this theorem can be used verbatim here. In particular, it links the summability of the profits to the convergence rate of methods such as MISC and Sparse Grids Stochastic Collocation, i.e., based on a sum of difference operators. To use this result, we have to assess the summability properties of the profits. We thus introduce suitable estimates of the error and work contributions, $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, respectively. In particular, the estimate of $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ depends on the spatial regularity of the solution, on the convergence rate of the method used to solve the deterministic problems, and on the summability property of the Chebyshev expansion of the solution over the parameter space.

Theorem 4 (Convergence of the profit-based MISC estimator, see [33]) *If the profits, $P_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, satisfy the weighted summability condition*

$$\left(\sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathbb{N}_+^D\times\mathfrak{L}_+} P_{\boldsymbol{\alpha},\boldsymbol{\beta}}^p \Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}} \right)^{1/p} = C_P(p) < \infty$$

for some $0 < p \le 1$*, then*

$$\left|\mathbb{E}[F] - \mathcal{M}_{\mathcal{I}}[F]\right| \le C_P(p)\mathrm{Work}[\mathcal{M}_{\mathcal{I}}]^{1-1/p},$$

where $\mathrm{Work}[\mathcal{M}_{\mathcal{I}}]$ *is given by* (27).

We begin by introducing an estimate for the size of the work contribution, $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. To this end, let $\Delta W_{\boldsymbol{\alpha}}^{\mathrm{det}} = \mathrm{Work}\left[\boldsymbol{\Delta}^{\mathrm{det}}[F^{\boldsymbol{\alpha}}]\right]$, i.e., let it be the cost of computing $\boldsymbol{\Delta}^{\mathrm{det}}[F^{\boldsymbol{\alpha}}]$ according to (21).

Assumption A3 (*Spatial work contribution*) There exist $\gamma_i \in [1,\infty)$ for $i = 1,\ldots,D$ and $C_W > 0$ such that

$$\Delta W_{\boldsymbol{\alpha}}^{\mathrm{det}} \le C_W 2^{\sum_{i=1}^D \gamma_i d_i \alpha_i}, \tag{30}$$

where $2^{\sum_{i=1}^D d_i\alpha_i}$ is proportional to the number of degress of freedom in the mesh on level $\boldsymbol{\alpha}$, cf. equation (17), and γ_i are related to the used deterministic solver and to the sparsity structure of the linear system, which might be different on each $\mathcal{B}_i$ depending on the chosen discretization.

Lemma 5 (Total work contribution) *When using Clenshaw–Curtis points for the discretization over the parameter space, the work contribution, $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, of each difference operator, $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$, can be decomposed as*

$$\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}} \le C_W 2^{\sum_{i=1}^D \gamma_i d_i \alpha_i + |\boldsymbol{\beta}-\mathbf{1}|},$$

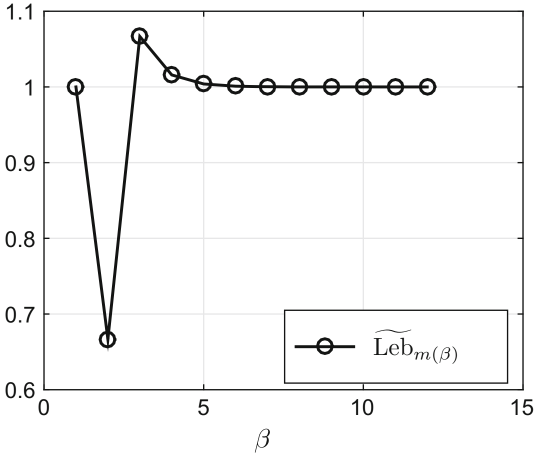
with γ_i and C_W as in Assumption A3.

Proof Combining (25) and (23), we have

$$\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \mathrm{Work}\left[\boldsymbol{\Delta}^{\mathrm{stoc}}[\boldsymbol{\Delta}^{\mathrm{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]]\right] = \mathrm{Work}\left[\boldsymbol{\Delta}^{\mathrm{stoc}}[\boldsymbol{\Delta}^{\mathrm{det}}[\mathcal{Q}^{m(\boldsymbol{\beta})}[F^{\boldsymbol{\alpha}}(\cdot)]]]\right].$$

Since the Clenshaw–Curtis points are nested, computing $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ (i.e., adding $[\boldsymbol{\alpha},\boldsymbol{\beta}]$ to the set $\mathcal{I}$ that defines the current MISC estimator) amounts to evaluating $F^{\boldsymbol{\alpha}}(\boldsymbol{y})$ in the set of "new" points added to the estimator by $\boldsymbol{\Delta}^{\mathrm{stoc}}[\cdot]$, i.e., $\times_{j:\beta_j>1}\left\{\mathcal{P}^{m(\beta_j)} \setminus \mathcal{P}^{m(\beta_j-1)}\right\}$, whose cardinality is $\prod_{j\ge1}(m(\beta_j) - m(\beta_j-1))$. The proof is then concluded by observing that the definition of $m(\beta)$ in (20) immediately gives $m(\beta_j) - m(\beta_j-1) \le 2^{\beta_j-1}$ and recalling Assumption A3. □

Fig. 2 Numerical evaluation of $\widetilde{\mathrm{Leb}}_{m(\beta)}$ for Clenshaw–Curtis points



Remark 2 We observe that the exponent $\boldsymbol{\beta} - \mathbf{1}$ guarantees that the directions along which no quadrature is actually performed (i.e., $\beta_j = 1$ for any $j \geq 1$) do not contribute to the total work.

Next, we prove a sequence of lemmas that allow us to conclude that an analogous estimate holds for the error contribution as well, i.e., that $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ can be bounded as a product of a term related to the spatial discretization and a term related to the approximation over the parameter space. To this end, we need to introduce the quantity

$$\mathrm{Leb}_{m(\beta)} = \sup_{f \in C^0([-1,1]), \|f\|_{L^\infty(-1,1)}=1} \left| Q^{m(\beta)}[f] - Q^{m(\beta-1)}[f] \right| \quad \forall \beta \in \mathbb{N}_+,$$

where $Q^{m(\beta)}[\cdot]$ are the univariate quadrature operators introduced in (18), and observe that $\mathrm{Leb}_1 = 1$. Next, let $\mathbb{L} = \max_{\beta \geq 1} \mathrm{Leb}_{m(\beta)}$, and note that $\mathbb{L} \leq 2$ since $Q^{m(\beta)}$ has positive weights. Moreover, a much smaller bound on $\mathbb{L}$ can be obtained for Clenshaw–Curtis points. Indeed, since Clenshaw–Curtis points are nested, we can also bound $\mathrm{Leb}_{m(\beta)} \leq \widetilde{\mathrm{Leb}}_{m(\beta)}$ with

$$\widetilde{\mathrm{Leb}}_{m(\beta)} = \sum_{y_\beta^j \in \mathcal{P}^{m(\beta)} \cap \mathcal{P}^{m(\beta-1)}} \left| \varpi_\beta^j - \varpi_{\beta-1}^j \right| + \sum_{y_j \in \mathcal{P}^{m(\beta)} \setminus \mathcal{P}^{m(\beta-1)}} \left| \varpi_\beta^j \right|,$$

and it can be verified numerically that $\widetilde{\mathrm{Leb}}_{m(\beta)}$ is bounded, attains its maximum for $\beta = 3$ and converges to 1 for $\beta \to \infty$, see Fig. 2. Therefore, we have $\mathbb{L} \leq \widetilde{\mathrm{Leb}}_{m(3)} \approx 1.067$.

Lemma 6 (Stochastic error contribution) *Let $f : \Gamma \to \mathbb{R}$ and $\boldsymbol{\beta} \in \mathfrak{L}_+$, and assume that the quadrature operator, $\mathcal{Q}^{m(\boldsymbol{\beta})}$, is built with Clenshaw–Curtis abscissae. If there exists a sequence, $\boldsymbol{\rho} = \{\rho_j\}_{j \geq 1}$, $\rho_j > 1$ for all j, such that*

1. $\sum_{j \geq 1} \frac{1}{\rho_j} < \infty$,

2. *there exists* $0 < C_f < \infty$ *such that for any finite set,* $\mathcal{G}' \subset \mathbb{N}_+$ *with* $\#\mathcal{G}' < \infty$, *the restriction of* f *on* $([-1,1])^{\mathcal{G}'}$ *admits a holomorphic complex extension in a Bernstein polyellipse,* $\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}'}$ *with* $\sup_{z \in \mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}'}} |f(z)| \leq C_f$.

Then, the set

$$\mathcal{J} = \left\{ j \geq 1 : \rho_j \leq 2(\mathbb{L})^{\frac{1}{3}} \right\},$$

has finite cardinality, i.e., $\#\mathcal{J} < \infty$ *and*

$$\begin{aligned}\left| \boldsymbol{\Delta}^{\mathrm{stoc}}[\mathcal{Q}^{m(\boldsymbol{\beta})} f] \right| &\leq C_{SE}(\boldsymbol{\rho}) \sup_{z \in \mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(z)| e^{-\sum_{j \geq 1} g_j m(\beta_j - 1)} \\ &\leq C_{SE}(\boldsymbol{\rho}) C_f e^{-\sum_{j \geq 1} g(\rho_j) m(\beta_j - 1)},\end{aligned}$$

holds, where $\mathcal{G}$ *is the support of* $\boldsymbol{\beta} - \mathbf{1}$,

$$0 < g(\rho_j) = \begin{cases} \log \rho_j, & \text{for } j \in \mathcal{J}, \\ \log \rho_j - \log 2 - \frac{1}{3} \log(\mathbb{L}), & \text{otherwise}, \end{cases}$$

and $C_{SE}(\boldsymbol{\rho}) < \infty$ *is independent of* $\boldsymbol{\beta}$.

Proof Let $\mathcal{G}$ be the support of $\boldsymbol{\beta} - \mathbf{1}$ with cardinality $\#\mathcal{G} < \infty$, $\boldsymbol{k} \in \mathbb{N}_+^{\mathcal{G}}$, and let $\Phi_{\mathcal{G},\boldsymbol{k}}$ denote the Chebyshev polynomials of the first kind with degree k_j along y_j for $j \geq 1$. We observe that $\Phi_{\mathcal{G},\boldsymbol{k}}$ are equivalent to the $\#\mathcal{G}$-variate Chebyshev polynomials over $[-1,1]^{\#\mathcal{G}}$ thanks to the product structure of the multivariate Chebyshev polynomials and to the fact that $\phi_0(y) = 1$. Next, consider the holomorphic extension of $f : \mathbb{C}^{\mathcal{G}} \to \mathbb{C}$, and its Chebyshev expansion over $\Phi_{\mathcal{G},\boldsymbol{k}}$ introduced in Lemma 3. Then

$$\begin{aligned}|\boldsymbol{\Delta}^{\mathrm{stoc}}[\mathcal{Q}^{m(\boldsymbol{\beta})} f]| &= \left| \boldsymbol{\Delta}^{\mathrm{stoc}} \left[\mathcal{Q}^{m(\boldsymbol{\beta})} \left[\sum_{\boldsymbol{k} \in \mathbb{N}_+^{\mathcal{G}}} f_{\boldsymbol{k}} \Phi_{\mathcal{G},\boldsymbol{k}} \right] \right] \right| \\ &= \left| \sum_{\boldsymbol{k} \in \mathbb{N}_+^{\mathcal{G}}} f_{\boldsymbol{k}} \boldsymbol{\Delta}^{\mathrm{stoc}} \left[\mathcal{Q}^{m(\boldsymbol{\beta})} \left[\Phi_{\mathcal{G},\boldsymbol{k}} \right] \right] \right|\end{aligned}$$

holds. By construction of hierarchical surplus, we have $\boldsymbol{\Delta}^{\mathrm{stoc}}[\mathcal{Q}^{m(\boldsymbol{\beta})}[\Phi_{\mathcal{G},\boldsymbol{k}}]] = 0$ for all Chebyshev polynomials, $\Phi_{\mathcal{G},\boldsymbol{k}}$, such that $\exists\, j \in \mathcal{G} : k_j < m(\beta_j - 1)$ (i.e., for polynomials that are integrated exactly at least in one direction by both quadrature operators along that direction). Therefore, the previous sum reduces to the multi-index set $\boldsymbol{k} \geq m(\boldsymbol{\beta} - \mathbf{1})$. Furthermore, by the triangular inequality, we have

$$|\boldsymbol{\Delta}^{\mathrm{stoc}}[\mathcal{Q}^{m(\boldsymbol{\beta})} f]| \leq \sum_{\boldsymbol{k} \geq m(\boldsymbol{\beta} - \mathbf{1})} |f_{\boldsymbol{k}}| \left| \boldsymbol{\Delta}^{\mathrm{stoc}} \left[\mathcal{Q}^{m(\boldsymbol{\beta})} \left[\Phi_{\mathcal{G},\boldsymbol{k}} \right] \right] \right|.$$

Next, using the definitions of $\boldsymbol{\Delta}^{\text{stoc}}$ and $\text{Leb}_{m(\beta)}$ and recalling that Chebyshev polynomials of the first kind on $[-1,1]$ are bounded by 1, that $\text{Leb}_{m(\beta)} \leq \widetilde{\text{Leb}}_{m(\beta)} \leq 1$ for $\beta = 1, 2$ and $\text{Leb}_{m(\beta)} \leq \mathbb{L}$ for $\beta \geq 3$, we have

$$\begin{aligned}\left|\boldsymbol{\Delta}^{\text{stoc}}\Big[\mathcal{Q}^{m(\boldsymbol{\beta})}\big[\Phi_{\mathcal{G},\boldsymbol{k}}\big]\Big]\right| &= \left|\bigotimes_{j\in\mathcal{G}} \Delta_j^{\text{stoc}}[Q^{m(\beta_j)}[\phi_{k_j}]]\right| \\ &\leq \prod_{j\in\mathcal{G}} \widetilde{\text{Leb}}_{m(\beta_j)} \left\|\phi_{k_j}\right\|_{L^\infty([-1,1])} \leq \prod_{\beta_j\geq 3} \mathbb{L}. \qquad (31)\end{aligned}$$

We then bound $|f_{\boldsymbol{k}}|$ by Lemma 3 to obtain

$$\begin{aligned}|\boldsymbol{\Delta}^{\text{stoc}}[\mathcal{Q}^{m(\boldsymbol{\beta})} f]| &\leq \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \Bigg(\prod_{\beta_j\geq 3}\mathbb{L}\Bigg) \sum_{\boldsymbol{k}\geq m(\boldsymbol{\beta}-\boldsymbol{1})} \prod_{j\in\mathcal{G}} 2^{|k_j|_0}\rho_j^{-k_j} \\ &\leq \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \Bigg(\prod_{\beta_j\geq 3}\mathbb{L}\Bigg)\Bigg(\prod_{j\in\mathcal{G}} \sum_{k_j\geq m(\beta_j-1)} 2^{|k_j|_0}\rho_j^{-k_j}\Bigg) \\ &= \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \Bigg(\prod_{\substack{j\in\mathcal{J}\\ \beta_j\geq 3}}\mathbb{L}\Bigg)\Bigg(\prod_{\substack{j\notin\mathcal{J}\\ \beta_j\geq 3}}\mathbb{L}\Bigg) \\ &\quad\times \Bigg(\prod_{j\in\mathcal{G}\cap\mathcal{J}} \sum_{k_j\geq m(\beta_j-1)} 2^{|k_j|_0}\rho_j^{-k_j}\Bigg)\Bigg(\prod_{j\in\mathcal{G}\setminus\mathcal{J}} \sum_{k_j\geq m(\beta_j-1)} 2^{|k_j|_0}\rho_j^{-k_j}\Bigg).\end{aligned}$$

Next, we observe that $|k_j|_0 \leq \min\{1, k_j\}$ for $k_j \geq 0$ and $1 \leq \frac{1}{3}m(\beta_j - 1)$ for all $\beta_j \geq 3$. Then,

$$\begin{aligned}|\boldsymbol{\Delta}^{\text{stoc}}[\mathcal{Q}^{m(\boldsymbol{\beta})} f]| &\leq \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \mathbb{L}^{\#\mathcal{J}} \Bigg(\prod_{j\notin\mathcal{J}} \mathbb{L}^{\frac{1}{3}m(\beta_j-1)}\Bigg) \\ &\qquad 2^{\#\mathcal{J}} \Bigg(\prod_{j\in\mathcal{G}\cap\mathcal{J}} \sum_{k_j\geq m(\beta_j-1)} \rho_j^{-k_j}\Bigg)\Bigg(\prod_{j\in\mathcal{G}\setminus\mathcal{J}} \sum_{k_j\geq m(\beta_j-1)} 2^{k_j}\rho_j^{-k_j}\Bigg) \\ &\leq (2\mathbb{L})^{\#\mathcal{J}} \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \Bigg(\prod_{j\in\mathcal{G}} \sum_{k_j\geq m(\beta_j-1)} e^{-g(\rho_j)k_j}\Bigg) \\ &= (2\mathbb{L})^{\#\mathcal{J}} \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \Bigg(\prod_{j\in\mathcal{G}} \frac{1}{1-e^{-g(\rho_j)}}\Bigg)\Bigg(\prod_{j\in\mathcal{G}} e^{-g(\rho_j)m(\beta_j-1)}\Bigg) \\ &\leq C_{\text{SE}}(\boldsymbol{\rho}) \sup_{z\in\mathcal{E}_{\boldsymbol{\rho}}^{\mathcal{G}}} |f(\boldsymbol{z})| \prod_{j\geq 1} e^{-g(\rho_j)m(\beta_j-1)},\end{aligned}$$

where the last inequality is due to the fact that $m(\beta_j - 1) = 0$ whenever $j \notin \mathcal{G}$ or equivalently $\beta_j = 1$ and

$$\prod_{j \in \mathcal{G}} \frac{1}{1 - e^{-g(\rho_j)}} \leq \prod_{j > 1} \frac{1}{1 - e^{-g(\rho_j)}},$$

since $g(\rho_j) > 0$ for all $j \geq 1$. Note that $C_{\mathrm{SE}}(\boldsymbol{\rho})$ is independent of $\boldsymbol{\beta}$ and is bounded since

$$\prod_{j \geq 1} \frac{1}{1 - e^{-g(\rho_j)}} < \infty \Longleftrightarrow -\sum_{j \geq 1} \log(1 - e^{-g(\rho_j)}) < \infty \Longleftrightarrow \sum_{j \geq 1} e^{-g(\rho_j)} < \infty$$
$$\Longleftrightarrow \sum_{j \in \mathcal{J}} \frac{1}{\rho_j} + \sum_{j \notin \mathcal{J}} \frac{2\,(\mathbb{L})^{\frac{1}{3}}}{\rho_j} < \infty,$$

which was assumed. Moreover, to show that $\#\mathcal{J} < \infty$, note that $\sum_{j \in \mathcal{J}} \rho_j^{-1}$ is otherwise unbounded, which contradicts the first assumption of the theorem, namely $\sum_{j \geq 1} \rho_j^{-1} < \infty$. □

Remark 3 Sharper estimates could be obtained by exploiting the structure of the Chebyshev polynomials when bounding $\left|\Delta^{\mathrm{stoc}}[Q^{m(\beta_j)}[\phi_{k_j}]]\right|$ in (31) (for instance, the fact that $Q^{m(\beta_j)}[\phi_{k_j}] = 0$ whenever k_j is odd and larger than 1) rather than using the general bound $\Delta^{\mathrm{stoc}}[Q^{m(\beta_j)}[\phi_{k_j}]] \leq \widetilde{\mathrm{Leb}}_{m(\beta_j)} \left\|\phi_{k_j}\right\|_{L^\infty([-1,1])}$.

We are now almost in the position to prove the estimate on the error contribution (see Lemma 8); before doing this, we need another auxiliary lemma that gives conditions for the summability of certain sequences that will be considered in the proof of Lemma 8 as well as in the proof of the main theorem on the convergence of MISC.

Lemma 7 (Summability of stochastic rates) *Recall the definitions of* $\{\zeta_{0,j}\}_{j \geq 1}$ *in Lemma* 1, *of* $\{\zeta_{s,j}\}_{j \geq 1}$ *in Lemma* 2 *and of* $g(\cdot)$ *in Lemma* 6. *Under Assumption* A2, *for all* $s = 0, 1, \ldots, s_{\max}$, *the sequences* $\{e^{-g(\zeta_{s,j})}\}_{j \geq 1}$ *and* $\left\{\frac{1}{\zeta_{s,j}}\right\}_{j \geq 1}$ *are* ℓ^p*-summable for* $p \geq \tilde{p}_s = \frac{p_s}{1 - p_s}$, *with* $\tilde{p}_s < 1$.

Proof First note that, by definition of $g(\cdot)$, we have

$$\sum_{j \geq 1} e^{-p g(\zeta_{s,j})} \leq 2^p\, (\mathbb{L})^{3p} \sum_{j \geq 1} \zeta_{s,j}^{-p}.$$

Then, from (14)–(15), or (12)–(13) for $s = 0$, we can bound $2\tau_{s,j} \leq \zeta_{s,j}$ and obtain

$$\sum_{j \geq 1} \zeta_{s,j}^{-p} \leq 2^{-p} \sum_{j \geq 1} \tau_{s,j}^{-p} = 2^{-p} \left(\frac{\pi}{E_\delta \left\|\boldsymbol{b}_s\right\|_{\ell^{p_s}}^{p_s}}\right)^{-p} \sum_{j \geq 1} b_{s,j}^{(1-p_s)p}.$$

From Assumption A2, we know that $\boldsymbol{b}_s \in \ell^{p_s}$ for $p_s \leq \frac{1}{2}$, and therefore we have the condition

$$(1 - p_s)p \geq p_s \Rightarrow p \geq \frac{p_s}{1 - p_s} < 1.$$

□

Lemma 8 (Total error contribution) *Assume that the deterministic problem is solved with a method obtained by tensorizing piecewise multi-linear finite element spaces on each mesh, $\{\mathbb{T}_i\}_{i=1}^D$, discretizing each $\{\mathcal{B}_i\}_{i=1}^D$, and let h_i as in (17) be the mesh size of each $\{\mathbb{T}_i\}_{i=1}^D$. Then, under Assumptions A1 and A2, the error contribution, $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, of each difference operator, $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$, can be decomposed as*

$$\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} \leq \min_{s=0,1,\ldots,s_{\max}} C_s \Delta E_{\boldsymbol{\alpha}}^{\text{det}}(s) \Delta E_{\boldsymbol{\beta}}^{\text{stoc}}(s), \tag{32}$$

for a constant $C_s < \infty$ independent of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and

$$\Delta E_{\boldsymbol{\alpha}}^{\text{det}}(s) = 2^{-\boldsymbol{\alpha}\cdot\boldsymbol{r}_{\text{FEM}}(s\boldsymbol{q})}, \tag{33}$$

$$\Delta E_{\boldsymbol{\beta}}^{\text{stoc}}(s) = e^{-\sum_{j\geq 1} m(\beta_j - 1) g_{s,j}}, \tag{34}$$

with $g_{s,j} = g(\zeta_{s,j})$ as in Lemma 6 and $r_{\text{FEM}}(s\boldsymbol{q})_i = \min\{1, q_i s\}$ for $i = 1, \ldots, D$ with $\boldsymbol{q} \in \mathbb{R}_+^D$ s.t. $|\boldsymbol{q}| = 1$.

Proof Combining the definition of $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$, cf. (23), and the definition of $\Delta^{\text{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$, cf. (21), we have

$$\begin{aligned}
\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} = |\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]| &= \left|\boldsymbol{\Delta}^{\text{stoc}}[\boldsymbol{\Delta}^{\text{det}}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]]\right| = \left|\boldsymbol{\Delta}^{\text{stoc}}\left[\sum_{\boldsymbol{j}\in\{0,1\}^D} (-1)^{|\boldsymbol{j}|} F_{\boldsymbol{\alpha}-\boldsymbol{j},\boldsymbol{\beta}}\right]\right| \\
&= \left|\boldsymbol{\Delta}^{\text{stoc}}\left[\sum_{\boldsymbol{j}\in\{0,1\}^D} (-1)^{|\boldsymbol{j}|} \mathcal{Q}^{m(\boldsymbol{\beta})}[\Theta[u^{\boldsymbol{\alpha}-\boldsymbol{j}}(\cdot, \boldsymbol{y})]]\right]\right| \\
&= \left|\boldsymbol{\Delta}^{\text{stoc}}\left[\mathcal{Q}^{m(\boldsymbol{\beta})}\Theta\left[\sum_{\boldsymbol{j}\in\{0,1\}^D} (-1)^{|\boldsymbol{j}|} u^{\boldsymbol{\alpha}-\boldsymbol{j}}(\cdot, \boldsymbol{y})\right]\right]\right| \\
&= \left|\boldsymbol{\Delta}^{\text{stoc}}\left[\mathcal{Q}^{m(\boldsymbol{\beta})}\Theta[\boldsymbol{\Delta}^{\text{det}}[u^{\boldsymbol{\alpha}}(\cdot, \boldsymbol{y})]]\right]\right|.
\end{aligned}$$

We observe that $f(\boldsymbol{y}) = \Theta[\boldsymbol{\Delta}^{\text{det}}[u^{\boldsymbol{\alpha}}(\cdot, \boldsymbol{y})]]$ is a linear combination of some $u^{\boldsymbol{\alpha}}$ and that each of these $u^{\boldsymbol{\alpha}}$ is an $H_0^1(\mathcal{B}, \mathbb{C})$-holomorphic function, since the finite element approximations of u are holomorphic in the same complex region as u itself; hence, $f(\boldsymbol{y})$ is also holomorphic in the same region. Then, thanks to the summability properties in Lemma 7, we can apply Lemma 6 to $f(\cdot)$ in the polyellipses defined in Lemmas 1 and 2 by $\boldsymbol{\zeta}_s$ in (14) (or (12) for $s = 0$) and obtain

$$\begin{aligned}\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} &\le C_{\mathrm{SE}}(\boldsymbol{\zeta}_s) \sup_{z\in\mathcal{E}^{\mathcal{G}}_{\zeta_s}} |\Theta[\boldsymbol{\Delta}^{\det}[u^{\boldsymbol{\alpha}}(\boldsymbol{x},\boldsymbol{z})]]| e^{-\sum_{j\ge1} g(\zeta_{s,j})m(\beta_j-1)}\\ &\le C_{\mathrm{SE}}(\boldsymbol{\zeta}_s)\,\|\Theta\|_{H^{-1}(\mathcal{B})} \sup_{z\in\mathcal{E}^{\mathcal{G}}_{\zeta_s}} \left\|\boldsymbol{\Delta}^{\det}[u^{\boldsymbol{\alpha}}(\cdot,\boldsymbol{z})]\right\|_{H^1(\mathcal{B},\mathbb{C})} e^{-\sum_{j\ge1} g(\zeta_{s,j})m(\beta_j-1)},\end{aligned}$$

where $\mathcal{G}\subset\mathbb{N}_+$ denotes the support of $\boldsymbol{\beta}-\mathbf{1}$. Next, assuming that the spatial discretization consists of piecewise linear finite elements with spatial mesh sizes (17) and combining the a-priori bounds on the decay of the difference operators coming from the Combination Technique theory (see, e.g., [19, proof of Theorem 2]) with (4) and the fact that $u\in H^{1+s}(\mathcal{B})$ for any $s=0,1,\dots,s_{\max}$, we have the following bound for every $\boldsymbol{z}$ in the Bernstein polyellipse, $\mathcal{E}^{\mathcal{G}}_{\boldsymbol{\zeta}_s}$:

$$\left\|\boldsymbol{\Delta}^{\det}[u^{\boldsymbol{\alpha}}(\cdot,\boldsymbol{z})]\right\|_{H^1(\mathcal{B},\mathbb{C})} \le C_{CT}\,\|u(\cdot,\boldsymbol{z})\|_{\mathcal{H}^{\mathbf{1}+s\boldsymbol{q}}(\mathcal{B},\mathbb{C})}\,2^{-\sum_{i=1}^{D}\alpha_i\min\{1,q_i s\}} \tag{35}$$

$$\le C_{CT}C_{s,\boldsymbol{q}}\,\|u(\cdot,\boldsymbol{z})\|_{H^{1+s}(\mathcal{B},\mathbb{C})}\,2^{-\boldsymbol{\alpha}\cdot\boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})}, \tag{36}$$

for $\boldsymbol{q}\in\mathbb{R}^D_+$ s.t. q_i s.t. $|\boldsymbol{q}|=1$, some $C_{CT}>0$ independent of u, and where $C_{s,\boldsymbol{q}}$ is the embedding constant between $\mathcal{H}^{\mathbf{1}+s\boldsymbol{q}}(\mathcal{B},\mathbb{C})$ and $H^{1+s}(\mathcal{B},\mathbb{C})$. We then have the following bound:

$$\begin{aligned}\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} \le\; & C_{\mathrm{SE}}(\boldsymbol{\zeta}_s)\,\|\Theta\|_{H^{-1}(\mathcal{B})}\\ &\times C_{CT}C_{s,\boldsymbol{q}} \sup_{z\in\mathcal{E}^{\mathcal{G}}_{\zeta_s}} \|u(\cdot,\boldsymbol{z})\|_{H^{1+s}(\mathcal{B},\mathbb{C})}\, e^{-\sum_j g_{s,j}m(\beta_j-1)}2^{-\boldsymbol{\alpha}\cdot\boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})},\end{aligned}$$

where the constant, $C_{\mathrm{SE}}(\boldsymbol{\zeta}_s)$ is bounded independently of $\boldsymbol{\beta}$, thanks again to Lemma 6. The proof is then concluded by recalling that $\sup_{z\in\mathcal{E}^{\mathcal{G}}_{\zeta_s}}\|u(\cdot,\boldsymbol{z})\|_{H^{1+s}(\mathcal{B},\mathbb{C})}\le C_{s,u}$ independently of $\boldsymbol{\beta}$ and $\mathcal{G}$ due to Lemmas 1 and 2. □

Observe that the result in Lemma 8 gives a bound on $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ parametric on the vector $\boldsymbol{q}$. The optimal choice for such $\boldsymbol{q}$ will be discussed later on, in the proof of the main theorem, namely Theorem 10.

Remark 4 (Relaxing the simplifying assumption) We now clarify why the assumption $p_s<\frac{1}{2}$, for $s=0,1,\dots,s_{\max}$, is not essential. Due to a suboptimal choice of $\zeta_{s,j}$ in Lemma 2 (and Lemma 1 for $s=0$), the sequence $\{e^{-g_{s,j}}\}_{j\ge1}$ in Lemma 7, which is related to the solution u and which appears in the proof of the MISC convergence theorem, has worse summability $\tilde{p}_s=\frac{p_s}{1-p_s}$ than the sequence $\{b_{s,j}\}_{j\ge1}$, whose summability coefficient is p_s, and which is related to the diffusion coefficient, a. As we see in the main theorem stated below, we need $\tilde{p}_s<1$ to guarantee convergence of MISC, which implies $p_s<\frac{1}{2}$. By choosing the polyellipses in Lemmas 1 and 2 by the more elaborate strategy presented in [11], it would be possible to obtain the better summability $\tilde{p}_s=p_s$ for the sequence $\{e^{-g_{s,j}}\}_{j\ge1}$, which would only imply the less stringent condition $p_s<1$ and a better estimate for the MISC convergence rate. However, for ease of exposition, we maintain the sub-optimal choice, which is enough for the purpose of presenting the argument that proves convergence of MISC. The restriction $p_s < \frac{1}{2}$ formally prevents us from applying the MISC convergence analysis to diffusion coefficients with low spatial regularity. In practice, we see in Sect. 6 that the convergence estimates are numerically verified beyond this restriction.

Before proving the main theorem of this section, we finally need the following technical lemma.

Lemma 9 (Bounding a sum of double exponentials) *For $a > 0$, $b \geq 2$ and $0 \leq c < ab$,*

$$\sum_{k=1}^{\infty} e^{-ab^k+ck} \leq e^{-ab+\varepsilon(a,b,c)}$$

holds, where for each fixed c and b, $\varepsilon(\cdot, b, c)$ is a monotonically decreasing, strictly positive function with $\varepsilon(a, b, c) \to c$ as $a \to +\infty$.

Proof

$$\sum_{k\geq 1} e^{-ab^k+ck} = e^{-ab+c} + \sum_{k\geq 2} e^{-ab^k+ck} = e^{-ab+c} + \sum_{k\geq 1} e^{-ab^{k+1}+c(k+1)}$$
$$= e^{-ab}\left(e^c + e^c \sum_{k\geq 1} e^{-ab(b^k-1)+ck}\right).$$

We observe that for $b \geq 2$ we have $b^k - 1 \geq k$ for $k \geq 1$ integer. Therefore, $e^{-ab(b^k-1)} \leq e^{-abk}$ and we have

$$\sum_{k\geq 1} e^{-ab^k+ck} \leq e^{-ab}\left(e^c + e^c \sum_{k\geq 1} e^{k(c-ab)}\right) = e^{-ab}\left(e^c + \frac{e^{2c-ab}}{1-e^{c-ab}}\right).$$

Then,

$$\varepsilon(a, b, c) = \log\left(e^c + \frac{e^{2c-ab}}{1-e^{c-ab}}\right),$$

and we finish by verifying that the function, ε, has the required properties. □

We are now ready to state and prove our main result.

Theorem 10 (MISC convergence theorem) *Under Assumptions A1–A3, the profit-based MISC estimator, $\mathcal{M}_{\mathcal{I}}$, built using the set $\mathcal{I}$ defined in (29), Stochastic Collocation with Clenshaw–Curtis points as in (19)–(20), and spatial discretization obtained by tensorizing multi-linear piecewise finite element spaces on each mesh, $\{\mathbb{T}_i\}_{i=1}^{D}$, with mesh sizes h_i as in (17) for solving the deterministic problems, satisfies, for any $\delta > 0$,*

$$\left|\mathbb{E}[F] - \mathcal{M}_{\mathcal{I}}[F]\right| \leq C_\delta \text{Work}[\mathcal{M}_{\mathcal{I}}]^{-r_{\text{MISC}}+\delta},$$

for some constant $C_\delta > 0$ *that is independent of* $\mathrm{Work}[\mathcal{M}_\mathcal{I}]$ *and tends to infinity as* $\delta \to 0$. *Moreover,* $\mathrm{Work}[\mathcal{M}_\mathcal{I}]$ *is given by* (27)*, and*

$$r_{\mathrm{MISC}} = \max_{s=0,\ldots,s_{\max}} \begin{cases} r_{\det(s)}, & \textit{if } r_{\det}(s) \le \frac{1}{p_s} - 2, \\ \left(\frac{1}{p_0} - 2\right)\left(1 + \frac{1}{r_{\det}(s)}\left(\frac{1}{p_0} - \frac{1}{p_s}\right)\right)^{-1}, & \textit{otherwise,} \end{cases}$$

where

$$r_{\det}(s) = \min\left\{\frac{1}{\max_{i=1,\ldots,D} \gamma_i d_i}, \frac{s}{\sum_{j=1}^{D} \gamma_j d_j}\right\}.$$

Proof In this proof, we use Theorem 4. Therefore, we need to estimate the p-summability of the weighted profits for some $p < 1$. To this end, we use Lemma 8. Observe that Lemma 8 provides a family of bounds for each $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, depending on s; therefore we would then ideally choose the best s for each $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. However, this optimization problem is too complex and we simplify it by assuming that

- only two values of s will be considered, $s = 0$ and $s = s^*$ (which will not necessarily coincide with $s_{\max}$);
- the optimization between $s = 0$ and $s = s^*$ will not be carried out individually on each $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, but we will rather take a “convex combination” of the two corresponding estimates and choose the best outcome only at the end of the proof.

To this end, we first need to rewrite the result of the Lemma 8 in a more suitable form for any fixed $s^* \in \{1, 2, \ldots, s_{\max}\}$; note that from here on, with a slight abuse of notation, we drop the superscript * and simply use s to denote the fixed value. Thus, for such fixed s, consider the statement of Lemma 8, let $C_E = \max\{C_0, C_s\}$, $\chi_{j,s} = g_{s,j} \log_2 e$, and $\theta_{j,s} = (g_{0,j} - g_{s,j}) \log_2 e$ and combine (32)–(34), obtaining

$$\begin{aligned} \Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} &\le \min_{t=\{0,s\}} C_t \Delta E^{\det}_{\boldsymbol{\alpha}}(t) \Delta E^{\mathrm{stoc}}_{\boldsymbol{\beta}}(t) \\ &\le C_E \min_{\eta\in\{0,1\}} 2^{-\eta \boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha}} \prod_{j\ge1} e^{-m(\beta_j-1)[g_{s,j}+(1-\eta)(g_{0,j}-g_{s,j})]} \\ &= C_E \min_{\eta\in\{0,1\}} 2^{-\eta \boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha}} \prod_{j\ge1} 2^{-m(\beta_j-1)[g_{s,j}+(1-\eta)(g_{0,j}-g_{s,j})]\log_2 e} \\ &= C_E \min_{\eta\in\{0,1\}} 2^{-\eta \boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha}} \prod_{j\ge1} 2^{-m(\beta_j-1)[\chi_{j,s}+(1-\eta)\theta_{j,s}]} \\ &= C_E 2^{-\sum_{j\ge1} m(\beta_j-1)\chi_{j,s} - \max_{\eta\in\{0,1\}}\left(\eta \boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha} + \sum_{j\ge1} m(\beta_j-1)(1-\eta)\theta_{j,s}\right)} \\ &= C_E 2^{-\sum_{j\ge1} m(\beta_j-1)\chi_{j,s} - \max\left\{\sum_{j\ge1} m(\beta_j-1)\theta_{j,s},\ \boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha}\right\}}. \end{aligned}$$

for an arbitrary $\boldsymbol{q} \in \mathbb{R}^D_+$ with $|\boldsymbol{q}| = 1$ that we will choose later. We can now bound the weighted sum of the profits as follows:

$$
\begin{aligned}
&\sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathbb{N}_+^D\times\mathfrak{L}_+} P^p_{\boldsymbol{\alpha},\boldsymbol{\beta}}\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathbb{N}_+^D\times\mathfrak{L}_+} \Delta E^p_{\boldsymbol{\alpha},\boldsymbol{\beta}}\Delta W_{\boldsymbol{\alpha}}^{1-p}\Delta W_{\boldsymbol{\beta}}^{1-p}\\
&\le C_E^p C_W^{1-p} \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathbb{N}_+^D\times\mathfrak{L}_+} 2^{-p[\max\{\boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha},\sum_{j\ge1} m(\beta_j-1)\theta_{s,j}\}+\sum_{j\ge1} m(\beta_j-1)\chi_{s,j}]}\\
&\qquad\cdot 2^{(1-p)\sum_{i=1}^D \gamma_i d_i\alpha_i+(1-p)\sum_{j\ge1}(\beta_j-1)}\\
&= C_E^p C_W^{1-p} \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathbb{N}_+^D\times\mathfrak{L}_+} \min_{\lambda\in[0,1]} 2^{-p[\lambda\boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha}+(1-\lambda)\sum_{j\ge1} m(\beta_j-1)\theta_{s,j}+\sum_{j\ge1} m(\beta_j-1)\chi_{s,j}]}\\
&\qquad\cdot 2^{(1-p)\sum_{i=1}^D \gamma_i d_i\alpha_i+(1-p)\sum_{j\ge1}(\beta_j-1)}\\
&\le C_E^p C_W^{1-p} \min_{\lambda\in[0,1]} \sum_{[\boldsymbol{\alpha},\boldsymbol{\beta}]\in\mathbb{N}_+^D\times\mathfrak{L}_+} 2^{-p[\lambda\boldsymbol{r}_{\mathrm{FEM}}(s\boldsymbol{q})\cdot\boldsymbol{\alpha}+(1-\lambda)\sum_{j\ge1} m(\beta_j-1)\theta_{s,j}+\sum_{j\ge1} m(\beta_j-1)\chi_{s,j}]}\\
&\qquad\cdot 2^{(1-p)\sum_{i=1}^D \gamma_i d_i\alpha_i+(1-p)\sum_{j\ge1}(\beta_j-1)}\\
&= C_E^p C_W^{1-p} \min_{\lambda\in[0,1]} \left(\prod_{i=1}^D\sum_{k=1}^\infty 2^{-[p(\lambda r_{\mathrm{FEM}}(s\boldsymbol{q})_i+\gamma_i d_i)-\gamma_i d_i]k}\right)\\
&\qquad\cdot\left(\prod_{j=1}^\infty\sum_{k=1}^\infty 2^{-pm(k-1)((1-\lambda)\theta_{s,j}+\chi_{s,j})+(1-p)(k-1)}\right).
\end{aligned}
\tag{37}
$$

We then investigate under what conditions each of the two factors is finite (the constants C_E, C_W are bounded, cf. Lemmas 5 and 8). Before proceeding, we comment on the equations above. As we already mentioned at the beginning of the proof, here we are working in a suboptimal setting in which instead of choosing a different s for each $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, we restrict ourselves to choosing between only two values, $s=0$ or a certain $s>0$ (second line). Observe that we have an equality between the second and the third lines since 2^x is a monotone function of x. Hence, the minimum is always attained at either $\lambda=0$ or $\lambda=1$. However, when it comes to switching the order of the sum and the minimum in the fourth line, i.e., bounding the sum by choosing the same λ to bound every term in the sum, allowing for fractional λ gives a tighter bound on the overall sum than just considering $\lambda\in\{0,1\}$. Roughly speaking, we are somehow "mimicking" the fact that the optimal bound of the sum would use a different value of λ for every term by choosing an overall λ that is between the two possible values.

To investigate the condition for which each of the two factors in (37) are bounded, we immediately have for the first factor for all $i=1,\ldots,D$ that

$$
\begin{aligned}
&p\left(\lambda r_{\mathrm{FEM}}(s\boldsymbol{q})_i+\gamma_i d_i\right)-\gamma_i d_i>0\\
&\quad\Longrightarrow\quad p>\frac{\gamma_i d_i}{\lambda r_{\mathrm{FEM}}(s\boldsymbol{q})_i+\gamma_i d_i}=\left(\lambda\frac{r_{\mathrm{FEM}}(s\boldsymbol{q})_i}{\gamma_i d_i}+1\right)^{-1}.
\end{aligned}
\tag{38}
$$

Our goal is to optimize the above constraint for the summability exponent p. To this end, we will minimize the right-hand side of (38), observing that it decreases with respect to $r_{\mathrm{FEM}}(s\boldsymbol{q})_i$. Hence, recalling the dependence of $r_{\mathrm{FEM}}(s\boldsymbol{q})_i$ (cf. Lemma 8) on the vector of weights $\boldsymbol{q}$, we consider

$$\begin{aligned} r_{\det}(s) &= \max_{\boldsymbol{q}\in\mathbb{R}^D_+, |\boldsymbol{q}|=1} \min_{i=1,\dots,D} \frac{r_{\text{FEM}}(s\boldsymbol{q})_i}{\gamma_i d_i} \\ &= \max_{\boldsymbol{q}\in\mathbb{R}^D_+, |\boldsymbol{q}|=1} \min_{i=1,\dots,D} \min\left\{\frac{1}{\gamma_i d_i}, \frac{sq_i}{\gamma_i d_i}\right\} \\ &= \max_{\boldsymbol{q}\in\mathbb{R}^D_+, |\boldsymbol{q}|=1} \min\left(\frac{1}{\max_{i=1,\dots,D}\gamma_i d_i}, \min_{i=1,\dots,D}\frac{sq_i}{\gamma_i d_i}\right) \\ &= \min\left(\frac{1}{\max_{i=1,\dots,D}\gamma_i d_i}, \max_{\boldsymbol{q}\in\mathbb{R}^D_+, |\boldsymbol{q}|=1}\min_{i=1,\dots,D}\frac{sq_i}{\gamma_i d_i}\right) \end{aligned}$$

and observe that the second argument of the min is maximized with respect to $\boldsymbol{q}$ by making $\frac{sq_i}{\gamma_i d_i}$ constant over i, i.e.,

$$q_i = \frac{\gamma_i d_i}{\sum_{j=1}^D \gamma_j d_j} \Rightarrow r_{\det}(s) = \min\left\{\frac{1}{\max_{i=1,\dots,D}\gamma_i d_i}, \frac{s}{\sum_{j=1}^D \gamma_j d_j}\right\}.$$

With this optimal choice, then (38) becomes simply

$$p > (\lambda r_{\det}(s) + 1)^{-1}. \tag{39}$$

For the second factor, denoting the generic term of the inner sum as $a_{j,k}$ for brevity and observing that $a_{j,1} = 1$ for every j, we have

$$\begin{aligned} \prod_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{j,k} \le \prod_{j=1}^{\infty}\left(1 + \sum_{k=2}^{\infty} a_{j,k}\right) = \exp\left(\sum_{j=1}^{\infty}\log\left(1+\sum_{k=2}^{\infty}a_{j,k}\right)\right) \\ \le \exp\left(\sum_{j=1}^{\infty}\sum_{k=2}^{\infty}a_{j,k}\right). \end{aligned}$$

We thus only have to discuss the convergence of the sum

$$\begin{aligned} &\sum_{j=1}^{\infty}\sum_{k=2}^{\infty} 2^{-pm(k-1)[(1-\lambda)\theta_{s,j}+\chi_{s,j}]+(1-p)(k-1)} \\ &\quad= \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} 2^{-pm(k)[(1-\lambda)\theta_{s,j}+\chi_{s,j}]+(1-p)k} \\ &\quad\le \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} 2^{-p2^{k-1}[(1-\lambda)\theta_{s,j}+\chi_{s,j}]+(1-p)k}, \end{aligned} \tag{40}$$

where the last step is a consequence of the fact that, for Clenshaw–Curtis points, $m(k) \geq 2^{k-1}$ for $k \geq 1$, cf. (20). Moreover, $(1-\lambda)\theta_{s,j} + \chi_{s,j} \geq 0$. To study the summability of (40), we want to use Lemma 9 to bound the inner sum in (40). First, we rewrite

$$\begin{aligned}
&\sum_{k=1}^{\infty} 2^{-p2^{k-1}[(1-\lambda)\theta_{s,j}+\chi_{s,j}]+(1-p)k} \\
&\quad = \sum_{k=1}^{\infty} \exp\left(-p\frac{\log 2}{2}[(1-\lambda)\theta_{s,j} + \chi_{s,j}]2^k + (1-p)k\log 2\right) \\
&\quad = \sum_{k=1}^{\infty} \exp\left(-a2^k + ck\right) \\
&\quad \text{with } a = p\frac{\log 2}{2}[(1-\lambda)\theta_{s,j} + \chi_{s,j}] > 0, \quad c = (1-p)\log 2 > 0,
\end{aligned}$$

where we have used the notation in Lemma 9. Note that this lemma holds true under the assumptions that $a > 0$ and $0 \leq c < 2a$, where the latter has to be verified as follows

$$\begin{aligned}
2a > c \Leftrightarrow p\log 2[(1-\lambda)\theta_{s,j} + \chi_{s,j}] > (1-p)\log 2 \\
\Leftrightarrow (1-\lambda)\theta_{s,j} + \chi_{s,j} > \frac{(1-p)}{p},
\end{aligned}$$

which is true whenever

$$\chi_{s,j} > r_{\det}(s),$$

due to (39), $\theta_{s,j} \geq 0$ and $\lambda \leq 1$. Define $\overline{\mathcal{J}} = \{j \geq 1 \;:\; \chi_{s,j} \leq r_{\det}(s)\}$ which has a finite cardinality since $\chi_{s,j} \to \infty$ as $j \to \infty$. Resuming from (40), we have, due to Lemma 9,

$$\begin{aligned}
\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} 2^{-p2^{k-1}[(1-\lambda)\theta_{s,j}+\chi_{s,j}]+(1-p)k} &\leq C(\overline{\mathcal{J}}) + \sum_{j\notin\overline{\mathcal{J}}}\sum_{k=1}^{\infty} \exp\left(-a2^k + ck\right) \\
&\leq C(\overline{\mathcal{J}}) + \sum_{j\notin\overline{\mathcal{J}}} e^{-2a+\varepsilon(a,2,c)},
\end{aligned}$$

where $C(\overline{\mathcal{J}})$ is bounded, since $\#\overline{\mathcal{J}} < \infty$, and $\varepsilon(a,2,c)$ is a monotonically decreasing function with limit $c = (1-p)\log 2$ independent of j. Therefore, the previous series converges if and only if

$$\sum_{j\notin\overline{\mathcal{J}}}^{\infty} e^{-2a} = \sum_{j\notin\overline{\mathcal{J}}}^{\infty} e^{-p\log 2[(1-\lambda)\theta_{s,j}+\chi_{s,j}]} = \sum_{j\notin\overline{\mathcal{J}}}^{\infty} 2^{-p[(1-\lambda)\theta_{s,j}+\chi_{s,j}]}$$

converges. Inserting the expression of $\theta_{s,j}$ and $\chi_{s,j}$, we get

$$\begin{aligned}\sum_{j\notin\overline{\mathcal{J}}}^{\infty} 2^{-p[(1-\lambda)\theta_{s,j}+\chi_{s,j}]} &= \sum_{j\notin\overline{\mathcal{J}}}^{\infty} 2^{-p[(1-\lambda)(g_{0,j}-g_{s,j})+g_{s,j}]\log_2 e}\\ &= \sum_{j\notin\overline{\mathcal{J}}}^{\infty} e^{-p[(1-\lambda)(g_{0,j}-g_{s,j})+g_{s,j}]}\\ &= \sum_{j\notin\overline{\mathcal{J}}}^{\infty} e^{-p(1-\lambda)g_{0,j}} e^{-p\lambda g_{s,j}}.\end{aligned}$$

After applying the Hölder inequality in the previous summation with exponents $\eta_1^{-1}+\eta_2^{-1}=1$, we need to simultaneously ensure the boundedness of the following sums:

$$\sum_{j\notin\overline{\mathcal{J}}}^{\infty} e^{-p(1-\lambda)g_{0,j}\eta_2} \quad\text{and}\quad \sum_{j\notin\overline{\mathcal{J}}}^{\infty} e^{-p\lambda g_{s,j}\eta_1}.$$

Recalling the summability result in Lemma 7, we understand that the following two conditions must hold:

$$\begin{cases} p(1-\lambda)\eta_2 \geq \dfrac{p_0}{1-p_0} \\ p\lambda\eta_1 \geq \dfrac{p_s}{1-p_s} \end{cases} \Rightarrow \begin{cases} p \geq \dfrac{p_0}{1-p_0}\dfrac{1}{1-\lambda}\dfrac{1}{\eta_2} \\ p \geq \dfrac{p_s}{1-p_s}\dfrac{1}{\lambda}\left(1-\dfrac{1}{\eta_2}\right), \end{cases}$$

which closes the discussion of the summability of the second factor of (37) for a fixed s. Recalling the constraint (39) coming from the first factor of (37), we finally have to solve the following optimization problem:

$$p > \min_{\lambda\in[0,1],1\leq\eta_2} \max\left\{(\lambda r_{\text{det}}(s)+1)^{-1}, \frac{p_s}{1-p_s}\frac{1}{\lambda}\left(1-\frac{1}{\eta_2}\right), \frac{p_0}{1-p_0}\frac{1}{1-\lambda}\frac{1}{\eta_2}\right\}$$

i.e., we have to choose η_2 and λ to minimize the lower bound on p. We first optimally select η_2 given λ, i.e., we take $\eta_2=\eta_2^*$ such that

$$\frac{p_s}{1-p_s}\frac{1}{\lambda}\left(1-\frac{1}{\eta_2^*}\right) = \frac{p_0}{1-p_0}\frac{1}{1-\lambda}\frac{1}{\eta_2^*} \Rightarrow \eta_2^* = 1+\frac{1-p_s}{p_s}\frac{p_0}{1-p_0}\frac{\lambda}{1-\lambda}$$

Substituting back, we obtain

$$\frac{p_s}{1-p_s}\frac{1}{\lambda}\frac{\eta_2^*-1}{\eta_2^*} = \left(\frac{1}{p_0}-1+\lambda\left(\frac{1}{p_s}-\frac{1}{p_0}\right)\right)^{-1},$$

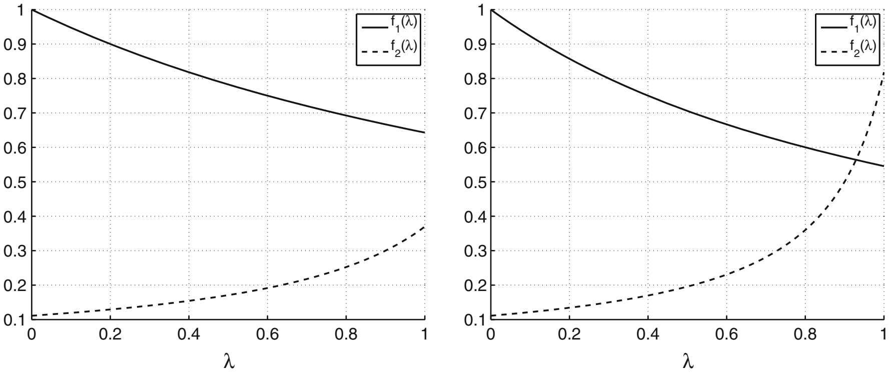


Fig. 3 Illustration of the optimization problem (41). As observed in the proof, f_1 is decreasing with λ while f_2 is increasing with λ. *Left* case 1 of the proof: the minmax point is $\lambda = 1$; *Right* case 2 of the proof: the minmax point is $\lambda < 1$

so that the minimization problem reads

$$p > \min_{\lambda \in [0,1]} \max \{ f_1(\lambda, s), f_2(\lambda, s) \},$$

$$f_1(\lambda, s) = (\lambda r_{\text{det}}(s) + 1)^{-1}, \quad f_2(\lambda, s) = \left(\frac{1}{p_0} - 1 + \lambda \left(\frac{1}{p_s} - \frac{1}{p_0} \right) \right)^{-1}. \quad (41)$$

Now, we recall that $p_0 \le p_s$. Hence, $f_2(\lambda, s)$ is increasing with λ. Conversely, $f_1(\lambda, s)$ is decreasing with λ since $r_{\text{det}}(s)$ is a positive number. Furthermore, notice that we cannot have $f_1(\lambda, s) < f_2(\lambda, s)$ for all $\lambda \in [0, 1]$. Indeed, the previous condition is equivalent to $f_1(0, s) \le f_2(0, s)$, i.e., $1 \le \frac{p_0}{1-p_0} \Rightarrow p_0 \ge \frac{1}{2}$, which does not satisfy Assumption A2. Note that, in this case, the lower bound for p in (41) is minimized for $\lambda = 0$, implying that $p > \frac{p_0}{1-p_0} > 1$, i.e., the method does not converge (cf. the statement of Theorem 4). Thus, we have only two cases (see also Fig. 3):

Case 1 $f_1(\lambda, s) > f_2(\lambda, s)$ for all $\lambda \in [0, 1]$, which means that the convergence of the method is dictated by the spatial discretization. Given that f_1 is decreasing and f_2 is increasing, the previous condition is equivalent to $f_1(1, s) \ge f_2(1, s)$, i.e., $r_{\text{det}}(s) \le \frac{1}{p_s} - 2$. In this case, the lower bound (41) is minimized for $\lambda = 1$, and we have $p > (r_{\text{det}}(s) + 1)^{-1}$.

Case 2 There exists $\lambda^* \in (0, 1)$ such that $f_1(\lambda, s) = f_2(\lambda, s)$. This condition is equivalent to the two conditions

$$\begin{cases} f_1(0, s) \ge f_2(0, s) \\ f_1(1, s) \le f_2(1, s) \end{cases} \Rightarrow \begin{cases} 1 \le \frac{1}{p_0} - 1 \\ r_{\text{det}}(s) \ge \frac{1}{p_s} - 2. \end{cases}$$

Solving for λ^* yields

$$\lambda^* = \frac{\frac{1}{p_0} - 2}{r_{\text{det}}(s) + \frac{1}{p_0} - \frac{1}{p_s}} > 0$$

which yields $p > \overline{p}$, where

$$\begin{aligned}\overline{p} &= \left(\left(\frac{\frac{1}{p_0} - 2}{r_{\det}(s) + \frac{1}{p_0} - \frac{1}{p_s}}\right) r_{\det}(s) + 1\right)^{-1} \\ &= \left(1 + \left(\frac{1}{p_0} - 2\right)\left(1 + \frac{1}{r_{\det}(s)}\left(\frac{1}{p_0} - \frac{1}{p_s}\right)\right)^{-1}\right)^{-1}.\end{aligned}$$

Since the previous computations were carried out for a fixed s, we can take the minimum over all possible values of s. Then, we can apply Theorem 4 and derive the convergence estimate,

$$\left|\mathbb{E}[F] - \mathcal{M}_{\mathcal{I}}[F]\right| \leq C_P(p) \text{Work}[\mathcal{M}_{\mathcal{I}}]^{1-1/p},$$

where $1 - 1/p = 1 - \max_{s=0,\ldots,s_{\max}}(1/\overline{p}) + \delta$ for any $\delta > 0$, which we reformulate as

$$\left|\mathbb{E}[F] - \mathcal{M}_{\mathcal{I}}[F]\right| \leq C_P\left(\frac{1}{1 + r_{\text{MISC}} - \delta}\right) \text{Work}[\mathcal{M}_{\mathcal{I}}]^{-r_{\text{MISC}}+\delta},$$

with $r_{\text{MISC}} = \max_{s=0,\ldots,s_{\max}}(1/\overline{p}) - 1$. □

5 Analysis of Example 1

In this section, we determine the value of $s_{\max}$ and the sequence $\{p_s\}_{s=0}^{s_{\max}}$ for Example 1. Since we will work with localized quantities of interest far from the boundary, cf. equation (45) written below, we believe that the effect of the boundary is negligible and the regularity $s_{\max}$ is mainly limited by the summability properties of κ. See "Appendix 2" for a slightly modified problem on the same domain where we can prove that the regularity is only limited by the summability properties of κ. Let us define the following family of auxiliary functions,

$$\Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}(\boldsymbol{x}) = \prod_{i=1}^{d} (\cos(\pi k_i x_i))^{\ell_i} (\sin(\pi k_i x_i))^{1-\ell_i},$$

so that κ from (8) can be written as

$$\begin{aligned}\kappa(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{\boldsymbol{k}\in\mathbb{N}^d} A_{\boldsymbol{k}} \sum_{\boldsymbol{\ell}\in\{0,1\}^d} y_{\boldsymbol{k},\boldsymbol{\ell}} \Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}(\boldsymbol{x}) \\ &= \sum_{j=0}^{\infty} \sum_{\{\boldsymbol{k}\in\mathbb{N}^d \,:\, |\boldsymbol{k}|=j\}} A_{\boldsymbol{k}} \sum_{\boldsymbol{\ell}\in\{0,1\}^d} y_{\boldsymbol{k},\boldsymbol{\ell}} \Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}(\boldsymbol{x}).\end{aligned}$$

Based on this expression, for $s \geq 0$, we analyze the summability of $\left\{A_{\boldsymbol{k}}\|D^{\boldsymbol{s}}\Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}\|_{L^\infty(\mathcal{B})}\right\}$ for $|\boldsymbol{s}| = s$ to determine the permissible values of p_s. First, for $|\boldsymbol{s}| = s$, observe that for a constant c independent of $\boldsymbol{k}$ we have

$$\|D^{\boldsymbol{s}}\Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}(\boldsymbol{x})\|_{L^\infty(\mathcal{B})} = \prod_{j=1}^{d}\left(\pi k_j\right)^{s_j} \leq c|\boldsymbol{k}|^s.$$

Then, for all $s \geq 0$, we have

$$\begin{aligned}
&\sum_{j=0}^{\infty}\sum_{\{\boldsymbol{k}\in\mathbb{N}^d\,:\,|\boldsymbol{k}|=j\}}\sum_{\boldsymbol{\ell}\in\{0,1\}^d} A_{\boldsymbol{k}}^{p_s}\|D^{\boldsymbol{s}}\Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}\|_{L^\infty(\mathcal{B})}^{p_s}\\
&\leq c2^d\sum_{j=0}^{\infty}\sum_{\{\boldsymbol{k}\in\mathbb{N}^d\,:\,|\boldsymbol{k}|=j\}} 2^{p_s\frac{|\boldsymbol{k}|_0}{2}}|\boldsymbol{k}|^{p_s s}(1+|\boldsymbol{k}|^2)^{-\frac{p_s\left(\nu+\frac{d}{2}\right)}{2}}\\
&\leq c2^d + c2^{d+p_s\frac{d}{2}}\sum_{j=1}^{\infty}\sum_{\{\boldsymbol{k}\in\mathbb{N}^d\,:\,|\boldsymbol{k}|=j\}} j^{-p_s\left(\nu+\frac{d}{2}-s\right)}\\
&= c2^d + c\frac{2^{d+p_s\frac{d}{2}}}{(d-1)!}\sum_{j=1}^{\infty} j^{-p_s\left(\nu+\frac{d}{2}-s\right)}\prod_{i=1}^{d-1}(j+i)\\
&= c2^d + c\frac{2^{d+p_s\frac{d}{2}}}{(d-1)!}\sum_{j=1}^{\infty} j^{-p_s\left(\nu+\frac{d}{2}-s\right)+d-1}\left(1+\frac{d-1}{j}\right)^{d-1}\\
&\leq c2^d + \frac{c2^{d+p_s\frac{d}{2}}d^{d-1}}{(d-1)!}\sum_{j=1}^{\infty} j^{-p_s\left(\nu+\frac{d}{2}-s\right)+d-1}.
\end{aligned}$$

From here, we obtain the bound

$$p_s > \left(\frac{\nu}{d}+\frac{1}{2}-\frac{s}{d}\right)^{-1}, \tag{42}$$

for all $s \geq 0$. Moreover, imposing $p_0 < \frac{1}{2}$ and $p_{s_{\max}} < \frac{1}{2}$ gives the bounds

$$\nu > \frac{3d}{2} \qquad \text{and} \qquad s_{\max} < \nu - \frac{3d}{2}, \tag{43}$$

respectively. Since $\Upsilon_{\boldsymbol{k},\boldsymbol{\ell}} \in C^\infty(\mathcal{B})$, the bounds in (43) are the only bounds on the value of $s_{\max}$. To determine an upper bound on the value of r_{MISC}, up to a small δ, we set $\gamma_1 = \cdots = \gamma_D = \gamma$ (motivated by the fact that all subdomains $\mathcal{B}_i$ are equal), we substitute $d_i = 1$ for $i = 1, \ldots, d$ and the lower bound of p_s in Theorem 10 and obtain after simplifying

$$
\begin{aligned}
r_{\text{MISC}} &= \max_{s=0,\ldots,s_{\max}} \begin{cases} \frac{\min(1,\frac{s}{d})}{\gamma} & \text{if } \frac{\min(1,\frac{s}{d})}{\gamma} \le \frac{\nu}{d} - \frac{3}{2} - \frac{s}{d}, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)\left(1 + \frac{\gamma}{\min(1,\frac{s}{d})}\frac{s}{d}\right)^{-1} & \text{if } \frac{\min(1,\frac{s}{d})}{\gamma} \ge \frac{\nu}{d} - \frac{3}{2} - \frac{s}{d}, \end{cases} \\
&= \max_{s=0,\ldots,s_{\max}} \begin{cases} \frac{s}{d\gamma} & \text{if } \frac{s}{d\gamma} \le \frac{\nu}{d} - \frac{3}{2} - \frac{s}{d} \text{ and } s \le d, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)(1+\gamma)^{-1} & \text{if } \frac{s}{d\gamma} \ge \frac{\nu}{d} - \frac{3}{2} - \frac{s}{d} \text{ and } s \le d, \\ \frac{1}{\gamma} & \text{if } \frac{1}{\gamma} \le \frac{\nu}{d} - \frac{3}{2} - \frac{s}{d} \text{ and } s \ge d, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)\left(1 + \frac{s\gamma}{d}\right)^{-1} & \text{if } \frac{1}{\gamma} \ge \frac{\nu}{d} - \frac{3}{2} - \frac{s}{d} \text{ and } s \ge d. \end{cases}
\end{aligned}
$$

Before continuing, we discuss the four branches of the previous expression. If $s_{\max} \le d$, then only the first two branches are valid. Since the rates in these two branches increase with s, the maximum is achieved for $s = s_{\max}$. If $s_{\max} \ge d$, then, since the rates in the third and fourth branches decrease with s, the maximum is achieved for $s = d$. Hence

$$
\begin{aligned}
r_{\text{MISC}} &= \begin{cases} \frac{s_{\max}}{d\gamma} & \text{if } \frac{s_{\max}(1+\gamma)}{d\gamma} \le \frac{\nu}{d} - \frac{3}{2} \text{ and } s_{\max} \le d, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)(1+\gamma)^{-1} & \text{if } \frac{s_{\max}(1+\gamma)}{d\gamma} \ge \frac{\nu}{d} - \frac{3}{2} \text{ and } s_{\max} \le d, \\ \frac{1}{\gamma} & \text{if } \frac{1}{\gamma} \le \frac{\nu}{d} - \frac{3}{2} - 1 \text{ and } s_{\max} \ge d, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)(1+\gamma)^{-1} & \text{if } \frac{1}{\gamma} \ge \frac{\nu}{d} - \frac{3}{2} - 1 \text{ and } s_{\max} \ge d, \end{cases} \\
&= \begin{cases} \frac{1}{\gamma}\left(\frac{\nu}{d} - \frac{3}{2}\right) & \text{if } \frac{1}{\gamma} \le 0 \text{ and } \frac{\nu}{d} \le \frac{5}{2}, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)(1+\gamma)^{-1} & \text{if } \frac{1}{\gamma} \ge 0 \text{ and } \frac{\nu}{d} \le \frac{5}{2}, \\ \frac{1}{\gamma} & \text{if } \frac{1}{\gamma} \le \frac{\nu}{d} - \frac{5}{2} \text{ and } \frac{\nu}{d} \ge \frac{5}{2}, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)(1+\gamma)^{-1} & \text{if } \frac{1}{\gamma} \ge \frac{\nu}{d} - \frac{5}{2} \text{ and } \frac{\nu}{d} \ge \frac{5}{2}, \end{cases} \\
&= \begin{cases} \gamma^{-1} & \text{if } \frac{\nu}{d} \ge \frac{1}{\gamma} + \frac{5}{2}, \\ \left(\frac{\nu}{d} - \frac{3}{2}\right)(1+\gamma)^{-1} & \text{if } \frac{\nu}{d} \le \frac{1}{\gamma} + \frac{5}{2}. \end{cases}
\end{aligned}
$$

In Fig. 4, we plot the upper bound of the rate of MISC work complexity, r_{MISC}, based on Theorem 10 and the following analysis variants:

Theory This is based on the summability properties of $\left\{A_{\boldsymbol{k}} \| D^{\boldsymbol{s}} \Upsilon_{\boldsymbol{k},\boldsymbol{\ell}} \|_{L^\infty(\mathcal{B})}\right\}$. We also use the value $r_{\text{FEM},i}(s) = 2\min\left(1, \frac{s}{d}\right)$ for all $i = 1, \ldots, d$. This is motivated by the fact that we expect to double the convergence rate of the underlying FEM method since we are considering convergence of a smooth linear functional of the solution.

Square summability Motivated by the arguments in Lemma 15 in the appendix, we believe that our results may be improved by instead considering the summability properties of $\left\{A_{\boldsymbol{k}}^2 \| D^{\boldsymbol{s}} \Upsilon_{\boldsymbol{k},\boldsymbol{\ell}} \|_{L^\infty(\mathcal{B})}^2\right\}$ for $|\boldsymbol{s}| \le s$. Similar calculations yield the bounds

$$
p_s > \left(\frac{2\nu}{d} + 1 - \frac{2s}{d}\right)^{-1}, \tag{44}
$$

and the corresponding conditions, $\nu > \frac{d}{2}$ and $s_{\max} < \nu - \frac{d}{2}$.

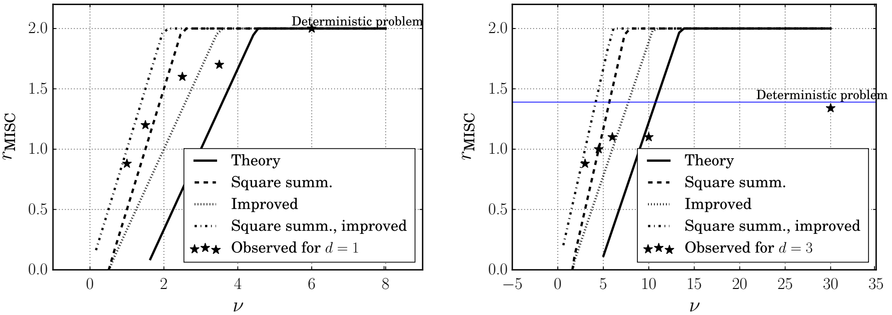


Fig. 4 The upper bound of the MISC rate, r_{MISC}, as predicted in Theorem 10 versus the observed rates when running the example detailed in Sect. 6. Refer to Sect. 5 for an explanation of the different curves. Also included are the observed convergence rates for a few values of ν and the observed convergence rate of MISC with no random variable and constant diffusion coefficient, a, as a *horizontal line*. The latter is referred to as the "deterministic problem" and shows more clearly the effect of the logarithmic factor in the work for $d > 1$, as shown in Fig. 9 and proved in [22, Theorem 1]

Improved As mentioned in Remark 4, we could in principle make our results sharper by taking $\tilde{p}_s = p_s$ instead of $\tilde{p}_s = \frac{p_s}{1-p_s}$. The modifications of Theorem 10 to account for these rates are straightforward. Moreover, when considering square summability, the conditions become $\nu > 0$ and $s_{\max} < \nu$.

We also include in Fig. 4 the observed convergence rates of MISC when applied to the example with different values of ν, as discussed below in Sect. 6, and the observed convergence rate of MISC when applied to the same problem with no random variables and a constant diffusion coefficient, a. In the latter case, MISC reduces to a deterministic combination technique [7]. Note that the rate of MISC with no random variables is an upper bound for the convergence rate of MISC with any $\nu > 0$.

From this figure, we can clearly see that the predicted rates in our theory are pessimistic when compared to the observed rates and that the suggested analysis of using the square summability or using the improved rates, $\tilde{p}_s$, might yield sharper bounds for the predicted work rates. On the other hand, we know from our previous work [22, Theorem 1] that the work degrades with increasing d with a log factor and in fact the expected work rate for maximum regularity when the number of random variables is finite is of $\mathcal{O}\left(W_{\max}^{-2} \log(W_{\max})^{d-1}\right)$. This can be seen Fig. 4 and $d = 3$, since in this case the observed work rate seems to be converging to a value less that 2.

6 Numerical Experiments

We now verify the effectiveness of the MISC approximation on some instances of the general elliptic equation (1), as well as the validity of the convergence analysis detailed in the previous sections. In particular, we consider the domain $\mathcal{B} = [0, 1]^d$ and the family of random diffusion coefficients specified in Example 1. In more detail, we consider a problem with one physical dimension ($d = 1$) and another with three dimensions ($d = 3$); in both cases, we set $\varsigma(\boldsymbol{x}) = 1$, and model the diffusion coefficient by the expansion (8) with different values of ν. Finally, the quantity of interest is a local average defined as

$$F(\boldsymbol{y}) = \frac{10}{(\sigma\sqrt{2\pi})^d} \int_{\mathcal{B}} u(\boldsymbol{x}, \boldsymbol{y}) \exp\left(- \frac{\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2}{2\sigma^2} \right) d\boldsymbol{x} \tag{45}$$

with $\sigma = 0.2$ and location $\boldsymbol{x}_0 = 0.3$ for $d = 1$ and $\boldsymbol{x}_0 = [0.3, 0.2, 0.6]$ for $d = 3$. The deterministic problems are discretized with a second-order centered finite differences scheme for which we expect to recover the same rate in the numerical experiments that we would obtain with piece-wise multi-linear finite elements on a structured mesh. We choose the mesh sizes in (17) and $h_{0,i} = 1/3$ for all $i = 1, \ldots, d$, and the resulting linear system is solved with GMRES with sufficient accuracy. With these values and using the coarsest discretization, $h_0 = 1/3$, in all dimensions, the coefficient of variation of the quantity of interest can be approximated to be between 90% and 110% depending on the number of dimensions, d, and the particular value of the parameter, ν, that we consider below. Finally, the quadrature points on the stochastic domain are the already-mentioned Clenshaw–Curtis points (see eq. (19) and (20)).

In the plots below, the computational work is compared in terms of the total number of degrees of freedom to avoid discrepancies in running time due to implementation details, i.e., using (27) and Assumption A3. Moreover, we set $\gamma = 1$ in (30), which is motivated by the fact that, for the tolerances we are interested in, we estimate that the cost of solving a linear system with GMRES is linear with respect to the number of degrees of freedom.

In order to evaluate the MISC estimator, we need to build the index set (29). To do that, we must be able to evaluate two quantities for every $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$: the work contribution, $\Delta W_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, and the error contribution, $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. Evaluating the work contribution is straightforward thanks to Assumption A3 and using $\gamma = 1$. On the other hand, evaluating the error contribution is more involved. We look at two options:

"brute-force" evaluation We compute $\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]$ for all $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ within some "universe" index set and set $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \left|\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right|$. Notice that this method is not practical since the cost of constructing the set, $\mathcal{I}$, would far dominate the cost of the MISC estimator. However, within some "universe" index set, this method would produce the best possible convergence and serve as a benchmark for other MISC sets within that universe.

"a-priori" evaluation We use Lemma 8 to bound $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. Using these bounds instead of exact values produces quasi-optimal index sets (cf. [3,33]). This method in turn requires the estimation of the parameters r_{FEM}, $\left\{g_{s,j}\right\}_{j \geq 1}$ for all $s = 0, \ldots, s_{\max}$. Since we use a second-order centered finite differences scheme and consider the convergence of a quantity of interest, we expect $r_{\mathrm{FEM}} = 2\min\left(1, \frac{\nu}{d}\right)$ as motivated by the "improved" analysis in the previous section and considering the summability properties of $\left\{A_{\boldsymbol{k}}^2 \|D^s \Upsilon_{\boldsymbol{k},\boldsymbol{\ell}}\|^2_{L^\infty(\mathcal{B})}\right\}$. This can also be validated numerically in the usual way by fixing all random variables to their expected value and checking the decay of $\Delta E_{\boldsymbol{\alpha},\boldsymbol{1}}$ with respect to $\boldsymbol{\alpha}$.

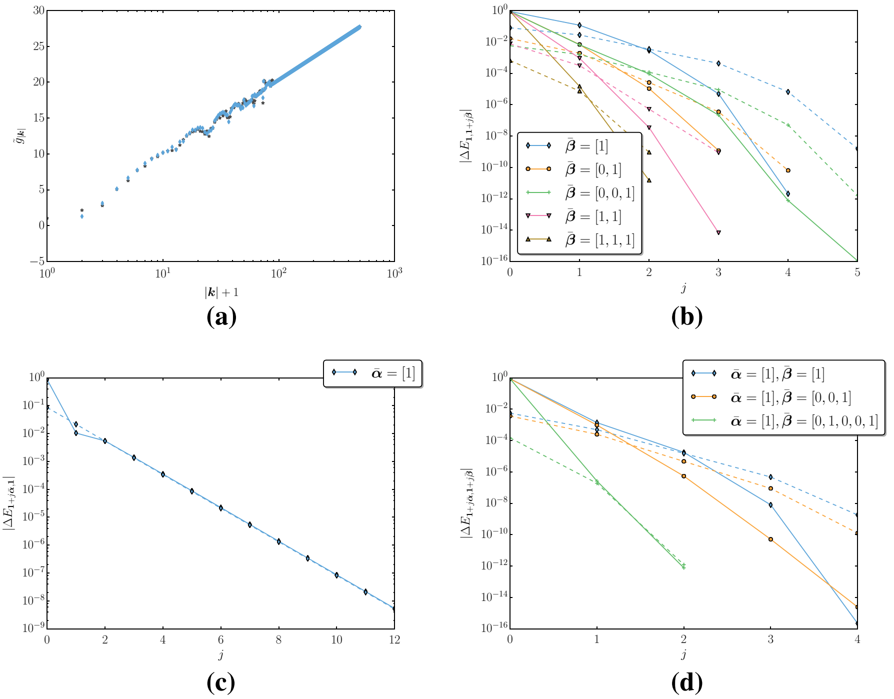
On the other hand, estimating $\left\{g_{s,j}\right\}_{j\ge1}$ for $s = 0, \dots, s_{\max}$ is more difficult since, in principle, we do not know a priori, for a given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which value of $s \in \{0, 1, \dots, s_{\max}\}$ yields the smallest estimate of $\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. Instead, we use a "simplified" model that was used in [22]:

$$\Delta E_{\boldsymbol{\alpha},\boldsymbol{\beta}} \le C e^{-\sum_{j\ge1} m(\beta_j-1)\tilde{g}_j} 2^{-|\boldsymbol{\alpha}| r_{\mathrm{FEM}}}, \tag{46}$$

where $\tilde{g}_j$ is some unknown function of $g_{s,j}$ for all $s = 0, 1, \dots, s_{\max}$. $\tilde{g}_j$ can be estimated given r_{FEM} and a set of evaluations of $\left|\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right|$ for some $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{I}^*$ by solving a least-squares problem to fit the linear model

$$\sum_{j\ge1} \tilde{g}_j m(\beta_j - 1) = -\log\left(\left|\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right|\right) - |\boldsymbol{\alpha}| r_{\mathrm{FEM}}, \qquad \text{for all } (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{I}^*.$$

For our example, these rates are plotted in Figs. 5a and 6a for $d = 1$ and $d = 3$, respectively. In our current implementation, the construction of the optimal MISC set, $\mathcal{I}$, is separate from the set $\mathcal{I}^*$. However, it is possible in principle to construct an



(a) **(b)**

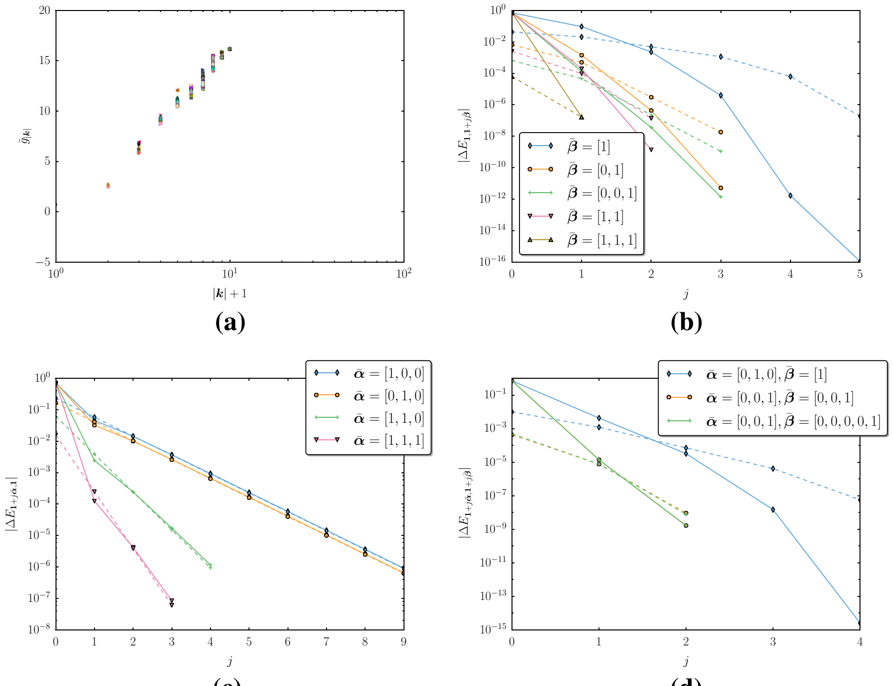
(c) **(d)**

Fig. 5 Example 1, $d = 1$ and $\nu = 2.5$. **a** A plot of the estimated stochastic rates, $\tilde{g}_j$, that are used in (46). Different markers correspond to different modes multiplying the same value of $A_{\boldsymbol{k}}$. **b–d** *solid lines* show the computed approximations of $\Delta E_{\mathbf{1},\mathbf{1}+j\bar{\boldsymbol{\beta}}}$, $\Delta E_{\mathbf{1}+j\bar{\boldsymbol{\alpha}},\mathbf{1}}$ and $\Delta E_{\mathbf{1}+j\bar{\boldsymbol{\alpha}},\mathbf{1}+j\bar{\boldsymbol{\beta}}}$, respectively, versus the model in (46) represented with *dashed lines*

algorithm in which the optimal MISC set, $\mathcal{I}$, is constructed iteratively by alternating between estimating rates given a set of indices and evaluating the MISC estimator.

Note that, in the current work, there are certain operations whose costs we do not track or compare. The first operation is the estimation of the stochastic rates, $\left\{\tilde{g}_j\right\}_{j\geq 1}$. The second operation is the construction of the optimal set given estimates of error and work contribution. We believe that the cost of these two operations can be reduced by using the previously mentioned iterative algorithm. The cost of these operations is thus dominated by the cost of evaluating the MISC estimator. The third operation is the assembly of the stiffness matrix, especially since it scales linearly with the number of random variables. While the cost of this operation is relevant to our discussion, it is usually dominated by the cost of the linear solver, at least for fine-enough discretizations.

Finally, we also compare MISC to the Multi-index Monte Carlo (MIMC) method as detailed in [23], for which $\mathcal{O}\left(W_{\max}^{-0.5}\right)$ convergence can be proved for Example 1 with $\gamma = 1, d \leq 3$ and sufficiently large ν (see "Appendix 1"). Moreover, when computing errors, we use the result obtained using a well-resolved MISC solution as a reference value.



(a) **(b)**

(c) **(d)**

Fig. 6 Example 1, $d = 3$ and $\nu = 4.5$. **a** A plot of the estimated stochastic rates, $\tilde{g}_j$, that are used in (46). Here different markers correspond to different modes multiplying the same value of $A_{\boldsymbol{k}}$. **b–d** *solid lines* show the computed approximations of $\Delta E_{\mathbf{1},\mathbf{1}+j\tilde{\boldsymbol{\beta}}}$, $\Delta E_{\mathbf{1}+j\tilde{\boldsymbol{\alpha}},\mathbf{1}}$ and $\Delta E_{\mathbf{1}+j\tilde{\boldsymbol{\alpha}},\mathbf{1}+j\tilde{\boldsymbol{\beta}}}$, respectively, versus the model in (46) represented with *dashed lines*

Figures 5b–d and 6b–d compare some computed values of $\left|\boldsymbol{\Delta}[F_{\boldsymbol{\alpha},\boldsymbol{\beta}}]\right|$ versus the model (46) using the estimated rates $r_{\text{FEM}} = 2\min\left(1, \frac{\nu}{d}\right)$ and $\left\{\tilde{g}_j\right\}_{j\geq 1}$. These plots show that the model (46) is a good fit for the case $d = 1$, $\nu = 2.5$ and $d = 3$, $\nu = 4.5$. Moreover, similar plots were produced for other values of d and ν that are not reported here but also show good fit. Figures 7 and 8 show

- the maximum spatial discretization level, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} \max(\boldsymbol{\alpha})$,
- the maximum quadrature level, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} \max(\boldsymbol{\beta})$,
- the index of the last activated random variable, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} \max_{\beta_j>1} j$,
- and the maximum number of jointly activated variables, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} |\boldsymbol{\beta}-\mathbf{1}|_0$.

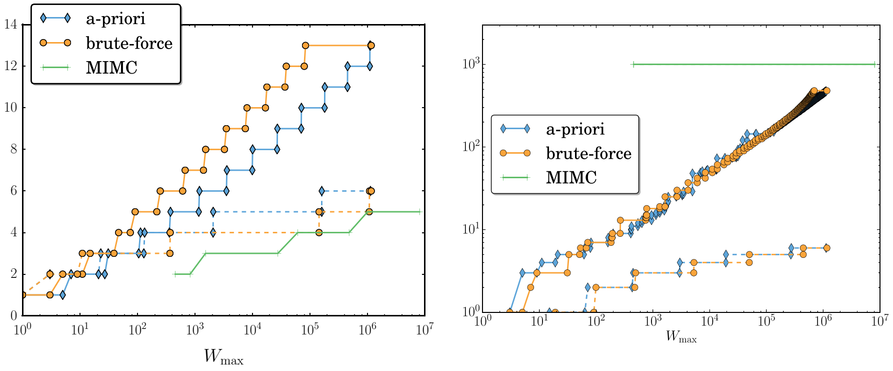


Fig. 7 Example 1, $d = 1$ and $\nu = 2.5$. This figure shows extreme values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ included in the MISC set, $\mathcal{I}$. Specifically, the *left-solid lines* are the maximum spatial discretization level, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} (\max(\boldsymbol{\alpha}))$, the *left-dashed lines* are the maximum quadrature level, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} (\max(\boldsymbol{\beta}))$, the *right-solid lines* are the index of the last activated random variable, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} \left(\max_{\beta_j>1} j\right)$, and the *right-dashed lines* are the maximum number of jointly activated variables, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} (|\boldsymbol{\beta}-\mathbf{1}|_0)$

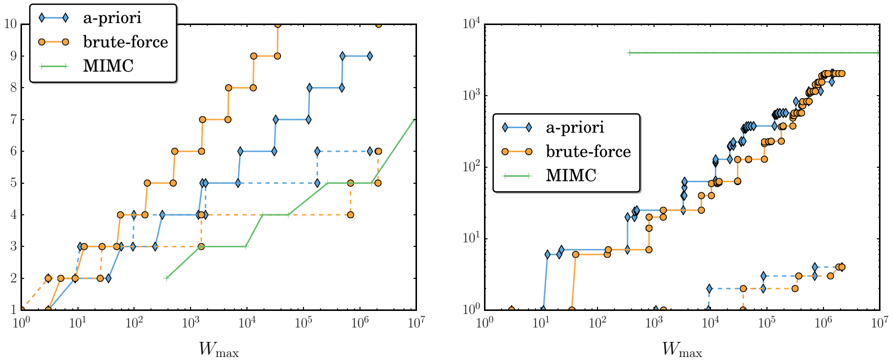
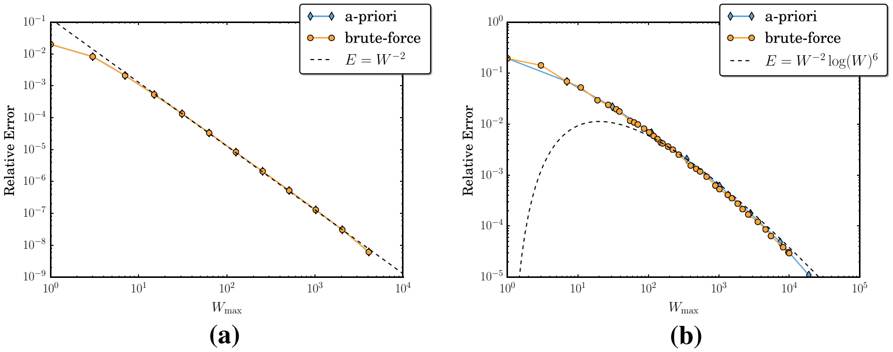
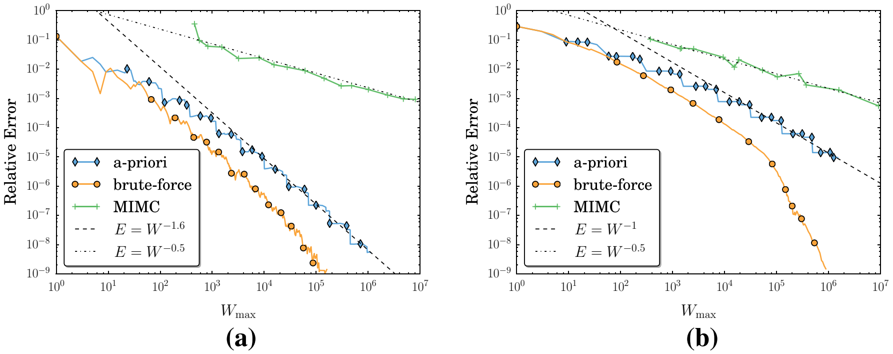


Fig. 8 Example 1, $d = 3$ and $\nu = 4.5$. This figure shows extreme values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ included in the MISC set $\mathcal{I}$. Specifically, the *left-solid lines* are the maximum spatial discretization level, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} (\max(\boldsymbol{\alpha}))$, the *left-dashed* are the maximum quadrature level, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} (\max(\boldsymbol{\beta}))$, the *right-solid* are the index of the last activated random variable, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} \left(\max_{\beta_j>1} j\right)$, and the *right-dashed* are the maximum number of jointly activated variables, $\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{I}} (|\boldsymbol{\beta}-\mathbf{1}|_0)$



(a) **(b)**

Fig. 9 Convergence results of MISC Example 1 with a constant diffusion coefficient, a (left, $d = 1$ and $\nu = 2.5$; right, $d = 3$ and $\nu = 4.5$). In this case, MISC is equivalent to a deterministic combination technique [7]. These plots show the non-asymptotic effect of the logarithmic factor for $d > 1$ (as discussed in [22, Theorem 1]) on the linear convergence fit in log–log scale



(a) **(b)**

Fig. 10 Convergence results of MISC versus MIMC when applied to Example 1 (left, $d = 1$ and $\nu = 2.5$; right, $d = 3$ and $\nu = 4.5$)

These values convey the size of the used index set, $\mathcal{I}$, for different values of $W_{\max}$.

Figure 4 shows the observed convergence rates of MISC for the cases $d = 1$ and $d = 3$ and different values of ν. This figure shows that the observed rates are better than those predicted by the theory developed in this work, which suggests that further improvement in the theory is possible (see Remark 4). Figures 9 and 10 show in greater details the observed convergence curves for $d = 1$, $\nu = 2.5$ and $d = 3$, $\nu = 4.5$ and their respective linear fit in log-log scale.

We recall that, as shown in [22, Theorem 1], the convergence rate of MISC with a finite number of random variables is $\mathcal{O}\left(W_{\max}^{-2}\log(W_{\max})^{d-1}\right)$. Compare this to the theory presented here that predicts, as $\nu \to \infty$, a convergence of $\mathcal{O}\left(W_{\max}^{-2+\epsilon}\right)$ for any $\epsilon > 0$. However, Fig. 9 shows that even for a problem with $d = 3$ and no random variables, MISC (which, in this case, becomes equivalent to a deterministic combination technique [7]) has an observed convergence rate that is closer to -1.38.

This is due to the effect of the logarithmic term that is nonzero for $d > 1$. Based on this, we should not expect a better convergence rate for $d = 3$ and any finite $\nu > 0$. This is also numerically validated in Fig. 10, which shows the full convergence curves for $d = 1$, $\nu = 2.5$ and $d = 3$, $\nu = 4.5$.

7 Conclusions

In this work, we analyzed the performance of the MISC method when applied to linear elliptic PDEs depending on a countable sequence of random variables. For ease of presentation, we worked on tensor product domains, but the results can be extended to more general domains and non-uniform meshes, as briefly mentioned Sect. 3. We proved a convergence result using a summability argument that shows that, in certain cases, the convergence of the method is essentially dictated by the convergence properties of the deterministic solver. We then applied the convergence theorem to derive convergence rates for the approximation of the expected value of a functional of the solution of an elliptic PDE with diffusion coefficient described by a random field, tracking the dependence of the convergence rate on the spatial regularity of the realizations of the random field. The theoretical findings are backed up by numerical experiments that show the dependence of the convergence rate on the regularity parameter. Future works includes extending the convergence analysis to higher-order finite element solvers and improving the estimates of the error contribution of each difference operator by taking into account the factorial terms appearing in the estimates for the size of the Chebyshev coefficients, cf. [3,11]. Moreover, the ideas in [12] can be extended to design an algorithm that iteratively estimates the optimal MISC set by alternating between optimizing the set and evaluating the estimator to ensure that the work to optimize the set is dominated by the work to evaluate the MISC estimator.

Acknowledgements F. Nobile and L. Tamellini received support from the Center for ADvanced MOdeling Science (CADMOS) and partial support by the Swiss National Science Foundation under the Project No. 140574 “Efficient numerical methods for flow and transport phenomena in heterogeneous random porous media”. L. Tamellini also received support from the Gruppo Nazionale Calcolo Scientifico - Istituto Nazionale di Alta Matematica “Francesco Severi” (GNCS-INDAM). R. Tempone is a member of the KAUST Strategic Research Initiative, Center for Uncertainty Quantification in Computational Sciences and Engineering. R. Tempone received support from the KAUST CRG3 Award Ref: 2281.

Appendix 1: Summability of Series Expansion

We start by recalling a useful multivariate Faà di Bruno formula taken from [13, Theorem 2.1].

Lemma 11 *Let $\mathcal{B} \subset \mathbb{R}^d$ be an open domain, $g : \mathcal{B} \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be functions of class $C^s(\mathcal{B})$ and denote $h = f \circ g : \mathcal{B} \to \mathbb{R}$. For any multi-index $\boldsymbol{i} \in \mathbb{N}^d$, $|\boldsymbol{i}| \le s$,* *and any $\boldsymbol{x} \in \mathcal{B}$,*

$$D^{\boldsymbol{i}} h(\boldsymbol{x}) = \boldsymbol{i}! \sum_{\lambda=1}^{|\boldsymbol{i}|} f^{(\lambda)}(g(\boldsymbol{x})) \sum_{r=1}^{\lambda} \sum_{p_r(\boldsymbol{i},\lambda)} \prod_{j=1}^{r} \frac{(D^{\boldsymbol{\ell}_j} g(\boldsymbol{x}))^{k_j}}{k_j!(\boldsymbol{\ell}_j!)^{k_j}}, \tag{47}$$

holds, where

$$p_r(\boldsymbol{i}, \lambda) = \{(k_j, \boldsymbol{\ell}_j) \in \mathbb{N} \times \mathbb{N}_0^d,\ j = 1, \ldots, r :\ \mathbf{0} \prec \boldsymbol{\ell}_1 \prec \boldsymbol{\ell}_2 \prec \cdots \prec \boldsymbol{\ell}_r, \\ \sum_{j=1}^{r} k_j = \lambda,\ \ \sum_{j=1}^{r} k_j \boldsymbol{\ell}_j = \boldsymbol{i}\}$$

and $\prec$ denotes the lexicographic ordering of multi-indices. The set $p_r(\boldsymbol{i}, \lambda)$ denotes the set of possible decompositions of $\boldsymbol{i}$ as a sum of λ multi-indices with $r \leq \lambda$ distinct multi-indices, $\boldsymbol{\ell}_j$, taken with multiplicity k_j such that $\sum_{j=1}^{r} k_j = \lambda$.

Also from [13, Corollary 2.9], we have that, for any $\boldsymbol{i} \in \mathbb{N}^d$,

$$\boldsymbol{i}! \sum_{r=1}^{\lambda} \sum_{p_r(\boldsymbol{i},\lambda)} \prod_{j=1}^{r} \frac{1}{k_j!(\boldsymbol{\ell}_j!)^{k_j}} = S_{|\boldsymbol{i}|,\lambda},$$

where $S_{n,k}$ is the *Stirling number of the second kind*, which counts the number of ways to partition a set of n objects into k non-empty subsets. Similarly, the *Bell number*, $B_n = \sum_{k=0}^{n} S_{n,k}$, counts the number of partitions of a set of n objects, whereas the *ordered Bell numbers* are defined by $\tilde{B}_n = \sum_{k=0}^{n} k! S_{n,k}$ and satisfy the recursive relation $\tilde{B}_n = \sum_{k=0}^{n-1} \binom{n}{k} \tilde{B}_k$. Clearly, $B_n \leq \tilde{B}_n$. Moreover, the bound

$$B_n \leq \tilde{B}_n \leq n!/(\log 2)^n \tag{48}$$

was given in [3, Lemma A.3]. We now use these results to show the following result

Lemma 12 *Let $\mathcal{B} \subset \mathbb{R}^d$ be an open-bounded domain and $\kappa \in C^s(\overline{\mathcal{B}})$ (real or complex valued) for $s \geq 0$. Then, $a = e^{\kappa} \in C^s(\overline{\mathcal{B}})$ and we have the estimate*

$$\|a\|_{C^s(\overline{\mathcal{B}})} \leq \frac{s!}{(\log 2)^s} \|a\|_{C^0(\overline{\mathcal{B}})} (1 + \|\kappa\|_{C^s(\overline{\mathcal{B}})})^s.$$

Proof Using formula (47), we have for any $\boldsymbol{i} \in \mathbb{N}^d$, $|\boldsymbol{i}| \leq s$ and any $\boldsymbol{x} \in \mathcal{B}$

$$|D^{\boldsymbol{i}} e^{\kappa(\boldsymbol{x})}| = \boldsymbol{i}! \sum_{\lambda=1}^{|\boldsymbol{i}|} e^{\kappa(\boldsymbol{x})} \sum_{r=1}^{\lambda} \sum_{p_r(\boldsymbol{i},\lambda)} \prod_{j=1}^{r} \frac{|D^{\boldsymbol{\ell}_j} \kappa(\boldsymbol{x})|^{k_j}}{k_j!(\boldsymbol{\ell}_j!)^{k_j}} \leq \|a\|_{C^0(\overline{\mathcal{B}})} \sum_{\lambda=1}^{|\boldsymbol{i}|} \|\kappa\|_{C^s(\overline{\mathcal{B}})}^{\lambda} S_{|\boldsymbol{i}|,\lambda} \\ \leq \|a\|_{C^0(\overline{\mathcal{B}})} (1 + \|\kappa\|_{C^s(\overline{\mathcal{B}})})^{|\boldsymbol{i}|} B_n.$$

The result then follows from the bound on the Bell numbers in (48). □

$L^p(\Gamma)$ Summability, Pointwise in Space

We now consider a diffusion coefficient as in Assumption A2:

$$a(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{\sum_{j\geq 1} \psi_j(\boldsymbol{x}) y_j\right\} = \prod_{j=1}^{\infty} e^{y_j \psi_j(\boldsymbol{x})}, \qquad \boldsymbol{x} \in \mathcal{B},$$

with y_j, $j \geq 1$, independent random variables, all uniformly distributed in $[-1, 1]$ and recall the definition of the sequence $\boldsymbol{b}_s = \{b_{s,j}\}_{j\geq 1}$, for all $s \in \mathbb{N}$ in (6).

Lemma 13 *If* $\boldsymbol{b}_0 \in \ell^2$ *then* $\mathbb{E}\big[a(\boldsymbol{x})^p\big] < \infty$ *for all* $0 < p < \infty$ *and* $\forall \boldsymbol{x} \in \mathcal{B}$.

Proof For any $\boldsymbol{x} \in \mathcal{B}$, we estimate the p-th moment of $a(\boldsymbol{x}, \boldsymbol{y})$, exploiting the independence of the random variables, y_j:

$$\mathbb{E}\big[a(\boldsymbol{x})^p\big] = \mathbb{E}\left[\prod_{j=1}^{\infty} e^{p y_j \psi_j(\boldsymbol{x})}\right] = \prod_{j=1}^{\infty} \mathbb{E}\Big[e^{p y_j \psi_j(\boldsymbol{x})}\Big] = \prod_{j=1}^{\infty} \frac{\sinh(p\psi_j(\boldsymbol{x}))}{p\psi_j(\boldsymbol{x})}$$
$$= \exp\left\{\sum_{j=1}^{\infty} \log\left(\frac{\sinh(p\psi_j(\boldsymbol{x}))}{p\psi_j(\boldsymbol{x})}\right)\right\}$$

where in the last two equalities we have implicitly assumed that $\sinh(z)/z = 1$ for $z = 0$. Setting $\theta_0(p; \boldsymbol{x}) = \prod_{j=1}^{\infty} \frac{\sinh(p\psi_j(\boldsymbol{x}))}{p\psi_j(\boldsymbol{x})}$ and observing that $\log(\sinh(z)/z) \sim z^2/6$, we have

$$\mathbb{E}\big[a(\boldsymbol{x})^p\big] = \theta_0(p; \boldsymbol{x}) < \infty \quad \forall \boldsymbol{x} \in \mathcal{B}, \quad 0 < p < \infty \qquad \Longleftrightarrow \qquad \sum_{j=1}^{\infty} \psi_j(\boldsymbol{x})^2 < \infty.$$

Since $\sum_{j=1}^{\infty} b_{0,j}^2 < \infty$ implies $\sum_{j=1}^{\infty} \psi_j(\boldsymbol{x})^2 < \infty$ for any $\boldsymbol{x} \in \mathcal{B}$, this concludes the proof. □

A similar result holds for higher-order derivatives of a.

Lemma 14 *Let* $s \in \mathbb{N}_+$. *If* $\boldsymbol{b}_s \in \ell^2$, *then for any* $\boldsymbol{i} \in \mathbb{N}^d$, $|\boldsymbol{i}| = s$, $\mathbb{E}\big[(D^{\boldsymbol{i}} a(\boldsymbol{x}))^{2p}\big] < \infty$ *for all* $0 < p < \infty$ *and* $\forall \boldsymbol{x} \in \mathcal{B}$.

Proof Since the calculations are tedious, we prove the result here for $s = 1$ only. Using the chain rule, we have

$$
\begin{aligned}
(\partial_{x_i} a(\boldsymbol{x}, \boldsymbol{y}))^{2p} &= \left(\sum_{j \geq 1} a(\boldsymbol{x}, \boldsymbol{y}) \partial_{x_i} \psi_j(\boldsymbol{x}) y_j \right)^{2p} \\
&= a(\boldsymbol{x}, \boldsymbol{y})^{2p} \sum_{\substack{\boldsymbol{q} \in \mathbb{N}^{\mathbb{N}} \\ |\boldsymbol{q}| = 2p}} (2p)! \prod_{j=1}^{\infty} \frac{1}{q_j!} (\partial_{x_i} \psi_j(\boldsymbol{x}) y_j)^{q_j} \\
&= \sum_{\substack{\boldsymbol{q} \in \mathbb{N}^{\mathbb{N}} \\ |\boldsymbol{q}| = 2p}} (2p)! \prod_{j=1}^{\infty} \frac{1}{q_j!} (\partial_{x_i} \psi_j(\boldsymbol{x}) y_j)^{q_j} e^{2p y_j \psi_j(\boldsymbol{x})}.
\end{aligned}
$$

Hence,

$$
\mathbb{E}\left[(\partial_{x_i} a(\boldsymbol{x}, \boldsymbol{y}))^{2p} \right] = \sum_{\substack{\boldsymbol{q} \in \mathbb{N}^{\mathbb{N}} \\ |\boldsymbol{q}| = 2p}} (2p)! \prod_{j=1}^{\infty} (\partial_{x_i} \psi_j(\boldsymbol{x}))^{q_j} \mathbb{E}\left[\frac{1}{q_j!} y_j^{q_j} e^{2p y_j \psi_j(\boldsymbol{x})} \right].
$$

We now distinguish between even or odd q_j. For even q_j, we have

$$
\mathbb{E}\left[\frac{1}{q_j!} y_j^{q_j} e^{2p y_j \psi_j(\boldsymbol{x})} \right] \leq \mathbb{E}\left[\frac{1}{q_j!} e^{2p y_j \psi_j(\boldsymbol{x})} \right] = \frac{1}{q_j!} \frac{\sinh(2p\psi_j(\boldsymbol{x}))}{2p\psi_j(\boldsymbol{x})},
$$

while for q_j odd we have

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{q_j!} y_j^{q_j} e^{2p y_j \psi_j(\boldsymbol{x})} \right] &= \frac{1}{q_j!} \int_{-1}^{1} \frac{1}{2} y^{q_j} e^{2p y \psi_j(\boldsymbol{x})} dy = \frac{1}{q_j!} \int_0^1 y^{q_j} \sinh(2p y \psi_j(\boldsymbol{x})) dy \\
&= \frac{1}{q_j!} \sum_{n=0}^{\infty} \frac{(2p\psi_j(\boldsymbol{x}))^{2n+1}}{(2n+1)!} \int_0^1 y^{2n+1+q_j} dy \\
&= \frac{1}{q_j!} \sum_{n=0}^{\infty} \frac{(2p\psi_j(\boldsymbol{x}))^{2n+1}}{(2n+1)!(2n+2+q_j)} \\
&\leq \frac{1}{(q_j+1)!} \sinh(2p|\psi_j(\boldsymbol{x})|) \leq \frac{2pb_{1,j}}{(q_j+1)!} \frac{\sinh(2p\psi_j(\boldsymbol{x}))}{2p\psi_j(\boldsymbol{x})}.
\end{aligned}
$$

Hence, defining the function

$$
f(q_j) = \begin{cases} \frac{1}{q_j!} & \text{for } q_j \text{ even}, \\ \frac{2pb_{1,j}}{(q_j+1)!} & \text{for } q_j \text{ odd}, \end{cases}
$$

we have

$$
\begin{aligned}
\mathbb{E}\Big[(\partial_{x_i} a(\boldsymbol{x},\boldsymbol{y}))^{2p}\Big] &\le \sum_{\substack{\boldsymbol{q}\in\mathbb{N}^{\mathbb{N}}\\|\boldsymbol{q}|=2p}} (2p)! \prod_{j=1}^{\infty} b_{1,j}^{q_j} f(q_j) \frac{\sinh(2p\psi_j(\boldsymbol{x}))}{2p\psi_j(\boldsymbol{x})}\\
&= \theta_0(2p;\boldsymbol{x}) \sum_{\substack{\boldsymbol{q}\in\mathbb{N}^{\mathbb{N}}\\|\boldsymbol{q}|=2p}} (2p)! \prod_{j=1}^{\infty} b_{1,j}^{q_j} f(q_j)\\
&\le \theta_0(2p;\boldsymbol{x}) \sum_{\substack{\boldsymbol{q}\in\mathbb{N}^{\mathbb{N}}\\|\boldsymbol{q}|=2p,\boldsymbol{q}\text{ even}}} (2p)!(1+2p)^{|\boldsymbol{q}|_0} \prod_{j=1}^{\infty} \frac{b_{1,j}^{q_j}}{q_j!}\\
&\le (1+2p)^p \theta_0(2p;\boldsymbol{x}) \sum_{\substack{\boldsymbol{q}\in\mathbb{N}^{\mathbb{N}}\\|\boldsymbol{q}|=p}} (2p)! \prod_{j=1}^{\infty} \frac{b_{1,j}^{2q_j}}{(2q_j)!}\\
&\le (1+2p)^p (2p)^p \theta_0(2p;\boldsymbol{x}) \sum_{\substack{\boldsymbol{q}\in\mathbb{N}^{\mathbb{N}}\\|\boldsymbol{q}|=p}} p! \prod_{j=1}^{\infty} \frac{(b_{1,j}^2)^{q_j}}{q_j!}\\
&= (1+2p)^p (2p)^p \theta_0(2p;\boldsymbol{x}) \left(\sum_{j\ge 1} b_{1,j}^2\right)
\end{aligned}
$$

from which we see that $\mathbb{E}\big[(\partial_{x_i} a(\boldsymbol{x},\boldsymbol{y}))^{2p}\big]$ is bounded for any $0 \le p < \infty$ and any $\boldsymbol{x} \in \mathcal{B}$ if $\boldsymbol{b}_1 \in \ell^2$. □

$L^p(\Gamma)$ Summability, Uniform in Space

Assuming now that $\boldsymbol{b}_s \in \ell^2$ so that the random field, a, is s-times differentiable in an $L^p(\Gamma)$ sense according to Lemma 14, we show that this implies some uniform $L^p(\Gamma)$ summability as detailed in the next lemma.

Lemma 15 *Let $s \in \mathbb{N}_+$. If $\boldsymbol{b}_s \in \ell^2$ then $\mathbb{E}\Big[\|a\|^p_{W^{\upsilon,\infty}(\mathcal{B})}\Big] < \infty$ for all $1 \le p < \infty$ and $\upsilon < s$.*

Proof We exploit the Sobolev embedding, $W^{\upsilon+\frac{d}{2q},2q}(\mathcal{B}) \subseteq W^{\upsilon,\infty}(\mathcal{B})$, for all $\upsilon \ge 0$ and $q \ge 1$. For $q \ge \max\{d/2(s-\upsilon), p/2\}$, we have

$$\mathbb{E}\left[\|a\|^p_{W^{\upsilon,\infty}(\mathcal{B})}\right] \leq \mathbb{E}\left[\|a\|^{2q}_{W^{s-\frac{d}{2q},\infty}(\mathcal{B})}\right] \lesssim \mathbb{E}\left[\|a\|^{2q}_{W^{s,2q}(\mathcal{B})}\right]$$

$$= \mathbb{E}\left[\sum_{|\boldsymbol{i}|\leq s} \int_{\mathcal{B}} (D^{\boldsymbol{i}} a(\boldsymbol{x}))^{2q} d\boldsymbol{x}\right] = \sum_{|\boldsymbol{i}|\leq s} \int_{\mathcal{B}} \mathbb{E}\left[(D^{\boldsymbol{i}} a(\boldsymbol{x}))^{2q}\right] d\boldsymbol{x} < \infty,$$

where the last term is bounded from Lemma 14. □

Now, we directly observe by taking $\upsilon = 0$ in the previous result that $a_{\max} = \|a\|_{L^\infty(\mathcal{B})}$ has bounded moments,

$$\mathbb{E}\left[a^p_{\max}\right] < \infty,$$

for all $1 \leq p < \infty$ and $0 < s$. Finally, by observing that, due to (2), in we have that $a_{\min} = \frac{1}{\|a^{-1}\|_{L^\infty(\mathcal{B})}}$ has the same distribution as $a_{\max}$. As a consequence, $a_{\min}$ has bounded moments as well. This implies in turn that (3) holds and thus problem (1) is well posed in the Bochner space, $L^p\left(\Gamma; H^1_0(\mathcal{B})\right)$. That is,

Corollary 16 (Well-posedness with log uniform coefficient) *We have for* $0 < \nu$ *that the problem in Example 1 is well posed in the Bochner space* $L^p\left(\Gamma; H^1_0(\mathcal{B})\right)$. *The corresponding solution, u, satisfies*

$$\|u\|_{L^p(\Gamma; H^1_0(\mathcal{B}))} \leq C\mathbb{E}\left[\frac{1}{a^p_{\min}}\right]^{1/p} \|\varsigma\|_{H^{-1}(\mathcal{B})}.$$

We observe that higher regularity of the solution, u, can be obtained by using larger values of s in Lemma 15. This in turn yields control on moments of $W^{\upsilon,\infty}(\mathcal{B})$ norms of the coefficient, a, and following, for instance, estimates similar to (2.10) in [18, Theorem 2.4], we can estimate moments of the $H^{1+s}(\mathcal{B})$ norm of the solution, u. These regularity estimates, once combined with pathwise error estimates for the combination technique, can be further used to show the corresponding ν-dependent convergence rates of MIMC [23], for Example 1, similar to what was presented in Sect. 5 for MISC in the current work.

Appendix 2: Shift Theorem for Problem (1)

Here, we seek to establish a *shift theorem* for the problem

$$\begin{cases} -\mathrm{div}(a(\boldsymbol{x})\,\nabla u(\boldsymbol{x})) = \varsigma(\boldsymbol{x}) & \text{in} \quad \mathcal{B} = [0,1]^D \\ u(\boldsymbol{x}) \;=\; 0 & \text{on} \quad \partial\mathcal{B}, \end{cases} \tag{49}$$

under suitable assumptions on a and ς.

With respect to problem (1), for convenience, we drop the dependence on the parameter vector, $\boldsymbol{y}$. We consider an odd periodic extension of ς, on $[-1,1]^D$, and an even periodic extension of the coefficient a on $[-1,1]^D$, named, respectively, $\tilde{\varsigma}$ and $\tilde{a}$. More precisely, for $\boldsymbol{j} = \{0,1\}^D$, we denote by $\boldsymbol{x}_{\boldsymbol{j}} = ((-1)^{j_1}x_1, \dots, (-1)^{j_D}x_D)$ and

$$\tilde{\varsigma}(\boldsymbol{x}_{\boldsymbol{j}} + 2\boldsymbol{k}) = (-1)^{|\boldsymbol{j}|}\varsigma(\boldsymbol{x}), \;\; \tilde{a}(\boldsymbol{x}_{\boldsymbol{j}} + 2\boldsymbol{k}) = a(\boldsymbol{x}), \;\; \forall \boldsymbol{x} \in [0,1]^2, \;\; \boldsymbol{j} \in \{0,1\}^D, \;\; \boldsymbol{k} \in \mathbb{N}^D.$$

The following Shift theorem holds for problem (49).

Lemma 17 *If the coefficient a is such that its periodic extension satisfies $\tilde{a} \in W^{s,\infty}(\mathbb{R}^D)$, $s \geq 0$ and $\varsigma \in C_0^\infty(\mathcal{B})$ then $u \in H^{s+1}(\mathcal{B})$.*

Proof We define the extended problem

$$\begin{cases} -\mathrm{div}(\tilde{a}(\boldsymbol{x})\,\nabla\tilde{u}(\boldsymbol{x})) = \tilde{\varsigma}(\boldsymbol{x}) & \text{in} \quad \tilde{\mathcal{B}} = [-1,1]^D \\ \int_{\tilde{\mathcal{B}}} u(\boldsymbol{x}) \;=\; 0 & \\ \text{periodic boundary conditions on } \partial\tilde{\mathcal{B}}. & \end{cases}$$

Since by assumption $\tilde{a} \in L^\infty(\mathbb{R}^D)$ and $\tilde{\varsigma} \in L^2(\tilde{\mathcal{B}})$, this problem has a unique solution, $\tilde{u} \in H^1_{per}(\tilde{\mathcal{B}}) \setminus \mathbb{R}$, where we denote with $H^s_{per}(\tilde{\mathcal{B}})$ the space of periodic functions with (periodic) square integrable derivatives up to order s. It is easy to check that the solution $\tilde{u}$ is odd, that is $\tilde{u}(\boldsymbol{x}_{\boldsymbol{j}}) = (-1)^{\boldsymbol{j}}\tilde{u}(\boldsymbol{x}), \forall \boldsymbol{x} \in [0,1]^D$, hence $\tilde{u} = 0$ (in the sense of traces) on $\partial\mathcal{B}$ and it coincides with the (unique) solution of (49) on $\mathcal{B}$. Moreover, standard elliptic regularity arguments allow us to say that $\tilde{u} \in H^s_{per}(\tilde{\mathcal{B}})$, hence $u \in H^s(\mathcal{B})$. □

References

1. I. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM Review, 52 (2010), 317–355.
2. A. Barth, C. Schwab, and N. Zollinger, *Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients*, Numerische Mathematik, 119 (2011), 123–161.
3. J. Beck, F. Nobile, L. Tamellini, and R. Tempone, *On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods*, Mathematical Models and Methods in Applied Sciences, 22 (2012), 1250023.
4. M. Bieri, *A sparse composite collocation finite element method for elliptic SPDEs.*, SIAM Journal on Numerical Analysis, 49 (2011), 2277–2301.
5. V. I. Bogachev, *Measure Theory, Vol. 1*, Springer Berlin Heidelberg, 2007.
6. H. J. Bungartz and M. Griebel, *Sparse grids*, Acta Numerica, 13 (2004), 147–269.
7. H. J. Bungartz, M. Griebel, D. Röschke, and C. Zenger, *Pointwise convergence of the combination technique for the Laplace equation*, East-West Journal of Numerical Mathematics, 2 (1994), 21–45.
8. J. Charrier, *Strong and weak error estimates for elliptic partial differential equations with random coefficients*, SIAM Journal on Numerical Analysis, 50 (2012), 216–246.
9. J. Charrier, R. Scheichl, and A. Teckentrup, *Finite element error analysis of elliptic pdes with random coefficients and its application to multilevel Monte Carlo methods*, SIAM Journal on Numerical Analysis, 51 (2013), 322–352.
10. A. Chkifa, *On the Lebesgue constant of Leja sequences for the complex unit disk and of their real projection*, Journal of Approximation Theory, 166 (2013), 176–200.

11. A. Cohen, R. DeVore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE'S*, Analysis and Applications, 9 (2011), 11–47.
12. N. Collier, A.- L. Haji- Ali, F. Nobile, E. von Schwerin, and R. Tempone, *A continuation multilevel Monte Carlo algorithm*, BIT Numerical Mathematics, 55 (2015), 399–432.
13. G. M. Constantine and T. H. Savits, *A multivariate Faà di Bruno formula with applications*, Transactions of the American Mathematical Society, 348 (1996), 503–520.
14. D. Dũng and M. Griebel, *Hyperbolic cross approximation in infinite dimensions*, Journal of Complexity, 33 (2016), 55–88.
15. B. Ganapathysubramanian and N. Zabaras, *Sparse grid collocation schemes for stochastic natural convection problems*, jcp, 225 (2007), 652–685.
16. M. B. Giles, *Multilevel Monte Carlo path simulation*, Operations Research, 56 (2008), 607–617.
17. W. J. Gordon and C. A. Hall, *Construction of curvilinear co-ordinate systems and applications to mesh generation*, International Journal for Numerical Methods in Engineering, 7 (1973), 461–477.
18. I. G. Graham, R. Scheichl, and E. Ullmann, *Mixed finite element analysis of lognormal diffusion and multilevel Monte Carlo methods*, Stochastic Partial Differential Equations: Analysis and Computations, (2015), 1–35.
19. M. Griebel and H. Harbrecht, *On the convergence of the combination technique*, in Sparse Grids and Applications - Munich 2012, J. Garcke and D. Pflüger, eds., vol. 97 of Lecture Notes in Computational Science and Engineering, Springer International Publishing, 2014, 55–74.
20. M. Griebel and S. Knapek, *Optimized general sparse grid approximation spaces for operator equations*, Mathematics of Computation, 78 (2009), 2223–2257.
21. M. Griebel, M. Schneider, and C. Zenger, *A combination technique for the solution of sparse grid problems*, in Iterative Methods in Linear Algebra, P. de Groen and R. Beauwens, eds., IMACS, Elsevier, North Holland, 1992, pp. 263–281.
22. A.- L. Haji- Ali, F. Nobile, L. Tamellini, and R. Tempone, *Multi-index stochastic collocation for random PDEs*, Computer Methods in Applied Mechanics and Engineering, 306 (2016), 95–122.
23. A.- L. Haji- Ali, F. Nobile, and R. Tempone, *Multi-index Monte Carlo: when sparsity meets sampling*, Numerische Mathematik, 132 (2015), 767–806.
24. A.- L. Haji- Ali, F. Nobile, E. von Schwerin, and R. Tempone, *Optimization of mesh hierarchies in multilevel Monte Carlo samplers*, Stochastic Partial Differential Equations: Analysis and Computations, 4 (2015), 76–112.
25. H. Harbrecht, M. Peters, and M. Siebenmorgen, *On multilevel quadrature for elliptic stochastic partial differential equations*, in Sparse Grids and Applications, vol. 88 of Lecture Notes in Computational Science and Engineering, Springer, 2013, 161–179.
26. M. Hegland, J. Garcke, and V. Challis, *The combination technique and some generalisations*, Linear Algebra and its Applications, 420 (2007), 249–275.
27. S. Heinrich, *Multilevel Monte Carlo methods*, in Large-Scale Scientific Computing, vol. 2179 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2001, 58–67.
28. T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs, *Isogeometric analysis: CAD, finite elements, nurbs, exact geometry and mesh refinement*, Computer Methods in Applied Mechanics and Engineering, 194 (2005), 4135–4195.
29. F. Y. Kuo, C. Schwab, and I. Sloan, *Multi-level Quasi-Monte Carlo Finite Element Methods for a Class of Elliptic PDEs with Random Coefficients*, Foundations of Computational Mathematics, 15 (2015), 411–449.
30. S. Martello and P. Toth, *Knapsack problems: algorithms and computer implementations*, Wiley-Interscience series in discrete mathematics and optimization, J. Wiley & Sons, 1990.
31. A. Narayan and J. D. Jakeman, *Adaptive Leja Sparse Grid Constructions for Stochastic Collocation and High-Dimensional Approximation*, SIAM Journal on Scientific Computing, 36 (2014), A2952–A2983.
32. F. Nobile, L. Tamellini, and R. Tempone, *Comparison of Clenshaw-Curtis and Leja Quasi-Optimal Sparse Grids for the Approximation of Random PDEs*, in Spectral and High Order Methods for Partial Differential Equations - ICOSAHOM '14, R. M. Kirby, M. Berzins, and J. S. Hesthaven, eds., vol. 106 of Lecture Notes in Computational Science and Engineering, Springer International Publishing, 2015, 475–482.
33. F. Nobile, L. Tamellini, and R. Tempone, *Convergence of quasi-optimal sparse-grid approximation of Hilbert-space-valued functions: application to random elliptic PDEs*, Numerische Mathematik, **134**(2) (2016), 343–388.

34. F. NOBILE, R. TEMPONE, AND C. WEBSTER, *An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM Journal on Numerical Analysis, 46 (2008), 2411–2442.
35. F. NOBILE, R. TEMPONE, AND C. WEBSTER, *A sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM Journal on Numerical Analysis, 46 (2008), 2309–2345.
36. C. SCHILLINGS AND C. SCHWAB, *Sparse, adaptive Smolyak quadratures for Bayesian inverse problems*, Inverse Problems, 29 (2013), 065011.
37. A. TECKENTRUP, P. JANTSCH, C. G. WEBSTER, AND M. GUNZBURGER, *A Multilevel Stochastic Collocation Method for Partial Differential Equations with Random Input Data*, SIAM/ASA Journal on Uncertainty Quantification, 3 (2015), 1046–1074.
38. H. W. VAN WYK, *Multilevel sparse grid methods for elliptic partial differential equations with random coefficients*, arXiv preprint arXiv:1404.0963, 2014.
39. G. W. WASILKOWSKI AND H. WOZNIAKOWSKI, *Explicit cost bounds of algorithms for multivariate tensor product problems*, Journal of Complexity, 11 (1995), 1–56.
40. D. XIU AND J. HESTHAVEN, *High-order collocation methods for differential equations with random inputs*, SIAM Journal on Scientific Computing, 27 (2005), 1118–1139.
41. C. ZENGER, *Sparse grids*, in Parallel Algorithms for Partial Differential Equations, W. Hackbusch, ed., vol. 31 of Notes on Numerical Fluid Mechanics, Vieweg, 1991, pp. 241–251.