

Modular Multifunctional Composite Structure for CubeSat Applications: Embedded Battery Prototype Thermal Analysis

*Original*

Modular Multifunctional Composite Structure for CubeSat Applications: Embedded Battery Prototype Thermal Analysis / Capovilla, G., Cestino, E., Reyneri, L., Valpiani, F.. - In: BATTERIES. - ISSN 2313-0105. - ELETTRONICO. - 11:5(2025). [10.3390/batteries11050172]

*Availability:*

This version is available at: 11583/2999563 since: 2025-04-27T20:55:05Z

*Publisher:*

MDPI

*Published*

DOI:10.3390/batteries11050172

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

## A Combination Technique for Optimal Control Problems Constrained by Random PDEs\*

Fabio Nobile<sup>†</sup> and Tommaso Vanzan<sup>‡</sup>

**Abstract.** We present a combination technique based on mixed differences of both spatial approximations and quadrature formulae for the stochastic variables to solve efficiently a class of optimal control problems (OCPs) constrained by random partial differential equations. The method requires to solve the OCP for several low-fidelity spatial grids and quadrature formulae for the objective functional. All the computed solutions are then linearly combined to get a final approximation which, under suitable regularity assumptions, preserves the same accuracy of fine tensor product approximations, while drastically reducing the computational cost. The combination technique involves only tensor product quadrature formulae, and thus the discretized OCPs preserve the (possible) convexity of the continuous OCP. Hence, the combination technique avoids the inconveniences of multilevel Monte Carlo and/or sparse grids approaches but remains suitable for high-dimensional problems. The manuscript presents an a priori procedure to choose the most important mixed differences and an analysis stating that the asymptotic complexity is exclusively determined by the spatial solver. Numerical experiments validate the results.

**Key words.** optimization under uncertainty, random PDEs, combination technique, sparse grids

**MSC codes.** 49M41, 65K05, 65D40, 90C15

**DOI.** 10.1137/22M1532263

**1. Introduction.** In this work, we propose a new framework to discretize and solve optimal control problems (OCPs) constrained by random partial differential equations (PDEs). This class of problems is increasingly more popular in the design of complex engineering systems, since a full knowledge of the physical PDE model is often not available, and the associated uncertainty is frequently modeled through random parameters. Consequently, the topic has drawn much attention in the last decade; see, e.g., the monographs [26, 34] and references therein.

We consider the following optimization problem with a random PDE constraint depending on a random vector  $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_N)$  taking values in a compact set  $\Gamma \subset \mathbb{R}^N$  and with density  $\rho$ ,

$$(1.1) \quad \begin{cases} \min_{u \in U} \mathbb{E}[F(y(\zeta))] + \frac{\nu}{2} \|u\|_U^2, \\ \text{where } y(\zeta) \in V \text{ solves} \\ \langle e(y(\zeta), \zeta), v \rangle = \langle Bu, v \rangle \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma, \end{cases}$$

\*Received by the editors November 1, 2022; accepted for publication (in revised form) April 3, 2024; published electronically June 26, 2024.

<https://doi.org/10.1137/22M1532263>

<sup>†</sup>CSQI Chair, École Polytechnique Fédérale de Lausanne, Lausanne, CH-1015, Switzerland ([fabio.nobile@epfl.ch](mailto:fabio.nobile@epfl.ch)).

<sup>‡</sup>Dipartimento di Scienze Matematiche, Politecnico di Torino, Torino, 10129, Italy ([tommaso.vanzan@polito.it](mailto:tommaso.vanzan@polito.it)).

where  $U$  and  $V$  are Hilbert spaces. The random vector accounts for the uncertainties in the PDE model, since coefficients, forcing terms, boundary/initial conditions, or shape of the domain may not be completely known either due to a lack of knowledge, measurement errors, or intrinsic randomness in the system.  $F$  is the convex quantity of interest to minimize. Examples are a tracking term which measures the distance of the state  $y(\zeta)$  from a desirable state  $y_d$ , the flux across a part of the boundary, the maximal deflection of a structure under external loads, or the aerodynamic forces over an airfoil. The random PDE (possibly nonlinear) is represented by  $e(\cdot, \cdot) : V \times \Gamma \rightarrow V'$ , and  $B : U \rightarrow V'$  is a linear bounded operator that describes how the control acts on the state (e.g., distributed or boundary control).

A standard approach to solve (1.1) is based on a sample average approximation (SAA) [42], which consists in replacing the (exact) expectation with an empirical approximation

$$\mathbb{E}[F(y(\zeta))] \approx \sum_{n=1}^M w_n F(y(\bar{\zeta}_n))$$

and consequently to collocate the  $\rho$ -a.e. PDE constraint onto the set of points  $\{\bar{\zeta}_n\}_{n=1}^M$ . The points are randomly generated in a Monte Carlo approach or chosen deterministically according to low discrepancy sets in a quasi-Monte Carlo quadrature [28]. The drawback of both approaches is that their discretization error slowly decays to zero with respect to the number of quadrature points  $M$ ; see [17, 33],

If the solution map  $\zeta \in \Gamma \rightarrow p(\zeta) \in V$ , with  $p$  being the adjoint variable, is sufficiently smooth, the stochastic collocation method (SCM) [1] is a valid alternative to Monte Carlo-based approximations. The SCM approximates the exact expectation in (1.1) with a quadrature formula obtained as the tensor product of one-dimensional (possibly Gaussian) quadrature operators, which may lead to the exponential convergence

$$\|u - u_{SC,\beta}\|_{L^2(D)} \leq C_{SC} \sum_{n=1}^N e^{-\tilde{g}_n \beta_n},$$

where  $\beta_n$  is the number of quadrature points used for the  $n$ th parameter  $\zeta_n$ , and  $\tilde{\mathbf{g}} = \{\tilde{g}_n\}_{n=1}^N$  are a set of coefficients depending on the region of holomorphy of the map  $\zeta \in \Gamma \rightarrow p(\zeta) \in V$  in the complex plane [1, 33]. See also [29, 30] for parametric regularity analyses of OCPs with parameter-dependent (stochastic) control  $u = u(\zeta)$ . Despite the exponential convergence with respect to  $\beta_n$ , the SCM suffers the curse of dimensionality if tensor product grids are used, since the number of collocation points grows exponentially with the dimension  $N$  of  $\zeta$  being  $M = \prod_{i=1}^N \beta_i$ . Sparse approximations have been extensively studied to approximate the state solution map  $\zeta \mapsto y(\zeta)$  (see, e.g., [7–9, 41]) and may lead to dimension-independent convergence rates in favorable cases. In the context of OCPs, sparse grids have been used in [23, 24] to reduce the number of samples for large values of  $N$ . On the other hand, [45] approximates the expectation in (1.1) using the multilevel Monte Carlo method (MLMC), which does not necessarily involve fewer samples but does reduce the computational cost by exploiting coarser discretization of the random PDE constraint. However, either approach is not always fully justified, since they both involve negative quadrature weights, which may destroy the convexity of the original OCP (1.1). Furthermore, even in the simple linear-quadratic case, negative weights introduce difficulties in developing efficient iterative solvers,

since the positive definiteness of the reduced optimality system cannot be guaranteed (thus preventing the use of conjugate gradient), while the preconditioners studied in [25, 36] for the full-space optimality system do require positive weights.

In this manuscript, we propose a combination technique (CT) to solve efficiently (1.1), inspired by the multi-index stochastic collocation method (MISC) for elliptic random PDEs in the context of forward uncertainty quantification (UQ) [18, 19] and by the seminal works concerning the solution of parametric high-dimensional PDEs [13, 14, 38]. The CT relies on a hierarchical representation of the optimal control specified by a set of multi-indices. Each multi-index corresponds to a level of discretization in the parameters  $\{\zeta_n\}_{n=1}^N$  and in the spatial approximation of the PDE constraint. The CT requires one to solve the OCP (1.1) several times, but each instance never uses a tensor product grid with many quadrature points and a fine PDE discretization at the same time. All the optimal controls computed on these coarser tensor product grids and meshes are then recombined to get a final approximation. Under suitable regularity assumptions, the CT can attain the same accuracy of the (full tensor) SCM solution, yet with a highly reduced computational cost. Thus, the CT is a valid alternative to both sparse grids and MLMC approximations, since it effectively reduces the number of quadrature points for high-dimensional stochastic input vectors (as sparse grid approximations do) and further reduces the computational cost by moving most of the computational work on coarser spatial grids (as MLMC does). Moreover, the CT solves exclusively OCPs discretized on tensor product grids, and thus the quadrature weights are positive and all the discretized OCPs are convex.

Key to the performance of the CT method is the choice of the set of multi-indices to be retained in the CT approximation. We here propose an a priori construction of such an index set based on a scalar quantity, called profit, assigned to each multi-index, as in [12, 19, 35]. The profit of a multi-index measures how advantageous it is to include it into the set, by taking into account both its error and its work contributions. Inherently in its construction, the CT approximation balances the spatial and stochastic error discretization according to this profit metric. We further present an analysis which characterizes the asymptotic complexity of the CT, under the assumption of exponential convergence of the tensor grid SCM, algebraic convergence of the PDE spatial discretization, and finite dimension of the random vector  $\zeta$ . We show that the asymptotic complexity of the method depends exclusively on the deterministic solver used for the random PDE, as for the MISC method [19] for forward UQ. Note that, while the complexity analysis is restricted to the finite-dimensional case, the computational framework presented here can be extended to the infinite-dimensional setting following, e.g., [6, 11] (see Remark 3).

Finally, although this work focuses on the SCM with Gaussian quadrature, we remark that the CT can be easily modified to accommodate other multilevel/multi-index quadrature formulas [21], such as multilevel/multi-index (quasi-)Monte Carlo methods (see Remark 2).

The rest of the manuscript is organized as follows. Section 2 introduces the class of problems considered. Section 3 introduces the CT approximation. Section 4 discusses the construction of the multi-index set which enters into the definition of the CT approximation. Two possible constructions are detailed: an adaptive procedure based on the dimension adaptive algorithm of [12], and an a priori construction based on some theoretical assumptions. A theoretical complexity analysis is further presented for the a priori construction (details of

the proofs are reported in the appendix). Section 5 discusses the validity of the assumptions for a model problem and presents numerical results that show the effectiveness of the CT.

**2. Problem setting.** Let  $\mathcal{D}$  be a Lipschitz bounded domain in  $\mathbb{R}^d$  and  $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_N)$  be an  $N$ -dimensional random vector whose components are mutually independent, and uniformly distributed on  $\Gamma := \times_{n=1}^N \Gamma_n$ , with probability density  $\rho(\zeta) = \prod_{n=1}^N \rho_n(\zeta_n) d\zeta_n$ ,  $\rho_n(\zeta_n) = \frac{1}{|\Gamma_n|}$ . Let  $\sigma_B(\Gamma)$  be the Borel  $\sigma$ -algebra over  $\Gamma$ , and let  $(\Gamma, \sigma_B(\Gamma), \rho(\zeta) d\zeta)$  be the complete probability space with uniform measure. In this manuscript, we consider the random PDE in weak form

$$(2.1) \quad \langle e(y(\zeta), \zeta), v \rangle = \langle \phi + Bu, v \rangle \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma,$$

where  $u \in U$ , and  $V$  and  $U$  are appropriate separable Hilbert spaces.

*Assumption 1.* Problem (2.1) is well-posed for  $\rho$ -a.e.  $\zeta \in \Gamma$ , and its solution, interpreted as a Hilbert space-valued function  $\zeta \in \Gamma \mapsto y(\zeta) \in V$ , belongs to the Bochner space

$$L^p_\rho(\Gamma; V) := \left\{ u : \Gamma \rightarrow V, \text{ strongly measurable s.t. } \int_\Gamma \|u(\zeta)\|_V^p d\rho(\zeta) < \infty \right\}$$

for some  $p \geq 1$ . Further, (2.1) defines a Gâteaux differentiable operator  $S_\zeta : U \rightarrow V$  such that  $y(\zeta) = S_\zeta(u)$  for  $\rho$ -a.e.  $\zeta$ .

In several applications, the interest lies in computing some convex and Fréchet differentiable quantity of interest of the random solution  $y(\zeta)$ , which we denote by  $F : L^p_\rho(\Gamma; V) \rightarrow L^q_\rho(\Gamma; \mathbb{R})$  for some  $q \geq 1$ . The goal is to optimally control the system, by minimizing the expected value of the quantity of interest, through an external control  $u \in U$ . An additional penalization on the energy of the control  $u$  can be added. In mathematical terms, this translates into the optimal control problem constrained by the random PDE

$$(2.2) \quad \begin{cases} \min_{u \in U} \mathbb{E}[F(y(\zeta))] + \frac{\nu}{2} \|u\|_U^2, \\ \text{where } y(\zeta) \in V \text{ solves} \\ \langle e(y(\zeta), \zeta), v \rangle = \langle \phi + Bu, v \rangle \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma, \end{cases}$$

with  $\nu \in \mathbb{R}^+$ . We denote the reduced objective functional by  $J : U \rightarrow \mathbb{R}$ , with  $J(u) := \mathbb{E}[F(S_\zeta(u))] + \frac{\nu}{2} \|u\|_U^2$ . We make the following working assumption.

*Assumption 2.* Problem (2.2) admits a solution  $u \in U$  which satisfies the optimality condition

$$(2.3) \quad \langle J'(u), v \rangle = \langle \nu u - B^* \mathbb{E}[p], v \rangle = 0 \quad \forall v \in U,$$

where  $B^*$  is the adjoint operator of  $B$ , and the state and adjoint variables  $(y, p)$  satisfy the optimality system

$$\begin{aligned} \langle e(y(\zeta), \zeta), v \rangle &= \left\langle \phi + BB^* \frac{\mathbb{E}[p]}{\nu}, v \right\rangle \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma, \\ \langle e^*(p(\zeta), \zeta), v \rangle &= \langle F'(y(\zeta)), v \rangle \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma, \end{aligned}$$

$F'$  being the Fréchet derivative of  $F$  and  $e^*$  being the adjoint operator of  $e$ . Further,  $J$  is twice Fréchet differentiable, and its Hessian evaluated in  $u$  is positive definite, with bounded inverse, and locally Lipschitz in a neighborhood of  $u$ .

Sufficient conditions for the validity of Assumption 2 can be found in, e.g., [26, 27, 34].

*Example 1.* To fix ideas, the reader may consider the elliptic random PDE that has been extensively studied in the context of UQ [1, 8, 31], namely

$$(2.4) \quad \langle e(y(\zeta), \zeta), v \rangle = \int_{\mathcal{D}} \kappa(x, \zeta) \nabla y(x, \zeta) \cdot \nabla v(x) dx = \int_{\mathcal{D}} \phi(x) v(x) dx \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma,$$

where  $V = H_0^1(\mathcal{D})$ . If the random diffusion field  $\kappa$  is strictly positive and bounded with probability 1, i.e., there exist  $\kappa_{\min} > 0$  and  $\kappa_{\max} < \infty$  such that

$$(2.5) \quad P(\kappa_{\min} \leq \kappa(x, \zeta) \leq \kappa_{\max} \quad \forall x \in \overline{\mathcal{D}}) = 1,$$

then (2.1) is well-posed and  $y \in L^\infty(\Gamma; V)$ . As an instance of (2.2), we may consider the tracking-type problem

$$(2.6) \quad \begin{cases} \min_{u \in L^2(\mathcal{D})} J(u) = \frac{1}{2} \mathbb{E} \left[ \|y(\zeta) - y_d\|_{L^2(\mathcal{D})}^2 \right] + \frac{\nu}{2} \|u\|_{L^2(\mathcal{D})}^2, \\ \text{where } y(\zeta) \in V \text{ solves} \\ \int_{\mathcal{D}} \kappa(x, \zeta) \nabla y(x, \zeta) \cdot \nabla v(x) dx = \int_{\mathcal{D}} (\phi(x) + u(x)) v(x) dx \quad \forall v \in V, \rho\text{-a.e. } \zeta \in \Gamma, \end{cases}$$

for which Assumption 2 is verified; see [26, section 6].

**3. Combination technique.** In this section, we present the CT to approximate the solution of (2.2). To do so, we introduce two multi-indices  $\beta$  and  $\alpha$ . The first one is associated to the quadrature formula used in the stochastic space, while the second one defines the level of approximation with respect to the physical variables.

Let  $\beta = (\beta_1, \beta_2, \dots, \beta_N) \in \mathbb{N}_+^N$ ,  $\mathbb{N}_+ := \{1, 2, \dots\}$  be a multi-index,  $m : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  a strictly increasing function, and  $\{\mathcal{Q}_j^{\beta_j}\}_{j=1}^N$  a set of one-dimensional quadrature operators defined as

$$(3.1) \quad \mathcal{Q}_j^{\beta_j} [F(y(\zeta))] = \sum_{n_j=1}^{m(\beta_j)} w_{n_j}^{\beta_j} F(y(\zeta_1, \dots, \zeta_{j-1}, \bar{\zeta}_{n_j}^{\beta_j}, \zeta_{j+1}, \dots, \zeta_N)) \approx \int_{\Gamma_j} F(y(\zeta_1, \dots, \zeta_N)) \rho_j(\zeta_j) d\zeta_j,$$

where  $(\bar{\zeta}_{n_j}^{\beta_j}, w_{n_j}^{\beta_j})_{n_j=1}^{m(\beta_j)}$  are the nodes and weights of the  $\beta_j$ th quadrature formula. To ensure good approximation properties, the nodes are usually chosen according to the underlying probability measure  $\rho_n(\zeta_n) d\zeta_n$ . Standard choices for uniform random variables are Gauss–Legendre, Clenshaw–Curtis, or Leja points [39, 44]. The map  $m(\cdot)$  specifies how the number of quadrature nodes depends on  $\beta_j$ . Common choices are the linear map  $m(\beta_j) = \beta_j$  or the doubling map  $m(\beta_j) = 2^{\beta_j} - 1$ ; see [35]. The multidimensional quadrature operator is directly obtained as the tensorization of the one-dimensional ones,

$$(3.2) \quad \mathbb{E}^\beta [F(y(\zeta))] := (\mathcal{Q}_1^{\beta_1} \otimes \dots \otimes \mathcal{Q}_N^{\beta_N}) [F(y(\zeta))] = \sum_{n_1=1}^{m(\beta_1)} \dots \sum_{n_N=1}^{m(\beta_N)} F(y(\bar{\zeta}_{n_1}^{\beta_1}, \dots, \bar{\zeta}_{n_N}^{\beta_N})) w_{n_1}^{\beta_1} \dots w_{n_N}^{\beta_N}.$$

Denoting with  $\Lambda_\beta \subset \Gamma$  the set of all quadrature nodes in (3.2) associated to the multi-index  $\beta$ , the semidiscretization of (2.2) is

$$(3.3) \quad \begin{cases} \min_{u \in U} \frac{1}{2} \mathbb{E}^\beta [F(y(\zeta))] + \frac{\nu}{2} \|u\|_U^2, \\ \text{where } y(\zeta) \in V \text{ solves} \\ \langle e(y(\zeta), \zeta), v \rangle = \langle \phi + Bu, v \rangle, \quad \forall v \in V, \forall \zeta \in \Lambda_\beta. \end{cases}$$

Next, let  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_D) \in \mathbb{N}^D$  be a second multi-index. In general,  $D$  does not necessarily coincide with the number of physical dimensions. The multi-index  $\alpha$  can tune the level of discretization along each physical dimension for meshes whose level of refinement can be set independently for each dimension (e.g., meshes obtained as the tensor product of one-dimensional meshes, structured meshes with a regular connectivity, or more general domains discretized using isogeometric analysis; see [3]). Otherwise,  $\alpha$  can be a simple scalar which tunes the reference mesh size across the whole domain  $\mathcal{D}$ , as in classical multilevel methods [21, 43, 45]. In a broader picture,  $\alpha$  could even represent time steps or describe a general set of multifidelity models [18, 37]. Nevertheless, for the sake of this manuscript, we suppose that the  $j$ th component of  $\alpha$  determines the mesh size  $h_{j, \alpha_j}$  along the  $j$ th physical dimension, and we denote with  $U^\alpha$  and  $V^\alpha$  the finite-dimensional approximations (e.g., finite element approximations on a mesh specified by  $\alpha$ ) of  $U$  and  $V$ .

Now, let  $u^{\alpha, \beta}$  be a solution of the fully discrete OCP

$$(3.4) \quad \begin{cases} \min_{u \in U^\alpha} \frac{1}{2} \mathbb{E}^\beta [F(y(\zeta))] + \frac{\nu}{2} \|u\|_U^2, \\ \text{where } y(\zeta) \in V^\alpha \text{ solves} \\ \langle e(y(\zeta), \zeta), v \rangle = \langle \phi + Bu, v \rangle \quad \forall v \in V^\alpha, \forall \zeta \in \Lambda_\beta, \end{cases}$$

which can be computed with any suitable optimization algorithm (see, e.g. [22]), with (3.4) being a finite-dimensional optimization problem. We assume that  $u^{\alpha, \beta} \xrightarrow{U} u$  as  $(\alpha, \beta) \rightarrow \infty$  (componentwise). Using, e.g., consistent finite element and stochastic collocation discretizations on tensor product grids, this is verified in the linear-quadratic case due the global convexity of the discretized objective functional; see, e.g., [17]. For a general formulation such as (2.2), which may involve local minima, sufficient conditions are provided by the Brezzi–Rappaz–Raviart theory [4] applied to the first-order optimality conditions; namely, the discretized gradient evaluated in  $u$  converges in norm to zero, the discretized objective functional is twice differentiable, whose Hessian evaluated in  $u$  is positive definite, with a uniformly (with respect to  $(\alpha, \beta)$ ) bounded inverse, and locally Lipschitz continuous in a neighborhood of  $u$ .

Next, we introduce the first-order difference operators for the deterministic and stochastic discretization parameters

$$(3.5) \quad \Delta_i^{\text{det}}[u^{\alpha, \beta}] = \begin{cases} u^{\alpha, \beta} - u^{\alpha - e_i, \beta} & \text{if } \alpha_i > 1, \\ u^{\alpha, \beta} & \text{if } \alpha_i = 1, \end{cases}$$

$$(3.6) \quad \Delta_i^{\text{stoc}}[u^{\alpha, \beta}] = \begin{cases} u^{\alpha, \beta} - u^{\alpha, \beta - e_i} & \text{if } \beta_i > 1, \\ u^{\alpha, \beta} & \text{if } \beta_i = 1, \end{cases}$$

where  $\mathbf{e}_i$  is the  $i$ th canonical vector. The deterministic and stochastic hierarchical surpluses are defined as the tensorized product of their respective first-order differences,

$$(3.7) \quad \Delta^{\det}[u^{\alpha,\beta}] = \otimes_{i=1}^D \Delta_i^{\det}[u^{\alpha,\beta}] = \Delta_1^{\det} \left[ \Delta_2^{\det} \left[ \dots \Delta_D^{\det}[u^{\alpha,\beta}] \right] \right] = \sum_{\mathbf{j} \in \{0,1\}^D} (-1)^{|\mathbf{j}|} u^{\alpha-\mathbf{j},\beta},$$

$$(3.8) \quad \Delta^{\text{stoc}}[u^{\alpha,\beta}] = \otimes_{i=1}^N \Delta_i^{\text{stoc}}[u^{\alpha,\beta}] = \Delta_1^{\text{stoc}} \left[ \Delta_2^{\text{stoc}} \left[ \dots \Delta_N^{\text{stoc}}[u^{\alpha,\beta}] \right] \right] = \sum_{\mathbf{j} \in \{0,1\}^N} (-1)^{|\mathbf{j}|} u^{\alpha,\beta-\mathbf{j}},$$

where we used the standard notation  $u^{\alpha,\beta} = 0$  if any of the  $\alpha_j$  or  $\beta_j$  is zero. Moreover,  $|\cdot|$  denotes the  $l_1$  norm of the vector (in this case the number of nonzero components).

Given a multi-index set  $\mathcal{I} \subset \mathbb{N}_+^{D+N}$ , the CT approximation  $\mathcal{M}_{\mathcal{I}}(u)$  for the optimal control  $u$ , solution of (2.2), is defined as the truncated telescopic sum,

$$(3.9) \quad u = \sum_{(\alpha,\beta) \in \mathbb{N}_+^{N+D}} \Delta[u^{\alpha,\beta}] \approx \sum_{(\alpha,\beta) \in \mathcal{I}} \Delta[u^{\alpha,\beta}] =: \mathcal{M}_{\mathcal{I}}(u),$$

where  $\Delta[u^{\alpha,\beta}] := \Delta^{\text{stoc}}[\Delta^{\det}[u^{\alpha,\beta}]]$ . In this work, we assume that the infinite series converges absolutely in  $U$ , namely that  $u^{\alpha,\beta} \xrightarrow{U} u$  as  $(\alpha,\beta) \rightarrow \infty$  (componentwise) and that  $\{\|\Delta u^{\alpha,\beta}\|_U\}_{(\alpha,\beta) \in \mathbb{N}_+^{N+D}} \in \ell^1(\mathbb{N}_+^{N+D})$ . In the rest of the manuscript, we assume the existence of specific bounds in Assumption 5 and carry out a complexity analysis under these hypotheses in section 4.2. Their validity is checked numerically on an example in section 5.

*Remark 1 (on the convergence of the infinite series).* Sufficient conditions for the convergence of the minimizers of the discretized OCPs to  $u$  have been previously recalled. On the other hand, the summability of the norms of the hierarchical surpluses may require additional hypotheses, and its analysis involves the cumbersome analytical derivation of suitable bounds. This is the subject of current endeavors in the linear-quadratic case. We remark that the validity of similar hierarchical expansions has been demonstrated in several contexts and it relies on the mix regularity with respect to the spatial and stochastic variables [13, 14, 18, 20, 35].

As in sparse grids [5], the rationale behind the truncation is that not all hierarchical surpluses  $\Delta[u^{\alpha,\beta}]$  have the same relevance. Therefore, under suitable assumptions (discussed in section 4.1), it is possible to retain only a few hierarchical surpluses while preserving accuracy at a reduced computational cost. The CT approximation (3.9) admits the following equivalent formulation [5]:

$$(3.10) \quad \mathcal{M}_{\mathcal{I}}(u) = \sum_{(\alpha,\beta) \in \mathcal{I}} c_{\alpha,\beta} u^{\alpha,\beta}, \quad c_{\alpha,\beta} = \sum_{\mathbf{j} \in \{0,1\}^{N+D}: (\alpha,\beta)+\mathbf{j} \in \mathcal{I}} (-1)^{|\mathbf{j}|},$$

which motivates the name “combination technique,” since  $\mathcal{M}_{\mathcal{I}}(u)$  is obtained as a combination of the solution of (3.4) for different multi-indices  $(\alpha,\beta)$ . Further, the coefficients  $c_{\alpha,\beta}$  are zero whenever  $(\alpha,\beta) + \mathbf{j} \in \mathcal{I} \forall \mathbf{j} \in \{0,1\}^{N+D}$ . Hence, only a few multi-indices  $(\alpha,\beta) \in \mathcal{I}$  actually contribute to the CT approximation.

*Remark 2 (sparse grids vs. combination technique).* Sparse grid approximations of OCPs under uncertainty [23, 24] rely on first-order differences of the one-dimensional quadrature operators

$$\tilde{\Delta}_n [\mathcal{Q}_n^{\beta_n}] := \mathcal{Q}_n^{\beta_n} - \mathcal{Q}_n^{\beta_n-1}, \quad n = 1, \dots, N,$$

and on the related hierarchical surpluses  $\tilde{\Delta} [\mathcal{Q}^\beta] := (\tilde{\Delta}_1 \otimes \dots \otimes \tilde{\Delta}_N) [\mathcal{Q}^\beta]$ . The sparse grid quadrature is then  $\tilde{\mathbb{E}}_{\mathcal{I}} = \sum_{\beta \in \mathcal{I}} \tilde{\Delta} [\mathcal{Q}^\beta]$ , and the sparse grid semidiscrete OCP reads as

$$(3.11) \quad \begin{cases} \min_{u \in U} \frac{1}{2} \tilde{\mathbb{E}}_{\mathcal{I}} [F(y(\zeta))] + \frac{\nu}{2} \|u\|_U^2, \\ \text{where } y(\zeta) \in V \text{ solves} \\ \langle e(y(\zeta), \zeta), v \rangle = \langle \phi + u, v \rangle \quad \forall v \in V, \forall \zeta \in \tilde{\Lambda}_\beta, \end{cases}$$

where  $\tilde{\Lambda}_\beta$  collects all the quadrature points involved by the sparse grid quadrature. It is well known that the weights of sparse grid quadrature formulae can be negative. This may break the convexity of the OCP. In contrast, if we fix a mesh discretization, the CT replaces the sparse approximation of the expectation operator with a hierarchical representation of  $u$ , using first-order differences of the solution  $u^\beta$  of the OCP computed with the tensor product quadrature determined by  $\beta$ . Hence, each OCP involves only quadratures with positive weights and the (reduced on  $u$ ) discrete OCPs remain convex.

*Remark 3 (multilevel/multi-index (quasi-)Monte Carlo methods).* The CT framework covers as particular instances other multilevel quadrature rules [21]. For instance, let  $\alpha$  be a scalar ( $D = 1$ ) setting the overall mesh size. Then, given a finest approximation level  $L$ , we may consider the telescopic sum

$$u \approx u^L = \sum_{\alpha=1}^L \Delta_1^{\det} [u^\alpha] = \frac{1}{\nu} B^* \sum_{\alpha=1}^L \mathbb{E} [p^\alpha - p^{\alpha-1}].$$

The expectation of the difference  $p^\alpha - p^{\alpha-1}$  is now prone to be approximated with a level-dependent (quasi-)Monte Carlo quadrature formula obtaining a multilevel (quasi-)Monte Carlo method to solve (2.2), which again preserves the (possible) convexity by avoiding negative quadrature weights, since  $p^\alpha$  and  $p^{\alpha-1}$  are the adjoint variables of two distinct OCPs discretized with the same (quasi-)Monte Carlo samples but on different spatial meshes. Notice that if  $\alpha \in \mathbb{N}^D$ ,  $D > 1$ , the CT can accommodate the multi-index (quasi-)Monte Carlo method [20, 40].

**4. Construction of the multi-index set  $\mathcal{I}$ .** It is clear that the multi-index set  $\mathcal{I}$  determines the computational efficiency of the CT approximation. Heuristically,  $\mathcal{I}$  should contain very few multi-indices that simultaneously lead to fine physical discretizations and to quadrature formulae with a large number of quadrature nodes, since they would require the solution of very expensive OCPs. In contrast,  $\mathcal{I}$  should contain “sparse” multi-indices which lead to fine discretizations/high level quadrature only across very few spatial and stochastic dimensions, while having coarse spatial/stochastic discretizations for most of the variables.

A well-known strategy to find a (quasi-)optimal multi-index set  $\mathcal{I}$  is to recast its construction as a knapsack optimization problem; see, e.g., [12, 19, 35]. To do so, we introduce the concepts of work contribution and error contribution associated to the hierarchical surplus

$\Delta[u^{\alpha,\beta}]$ . The work contribution  $\Delta W_{\alpha,\beta}$  measures the computational cost required to add  $\Delta[u^{\alpha,\beta}]$  to  $\mathcal{M}_{\mathcal{I}}(u)$ . In formulae, we set

$$\Delta W_{\alpha,\beta} = \text{Work} [\mathcal{M}_{\mathcal{I} \cup \{\alpha,\beta\}}(u)] - \text{Work} [\mathcal{M}_{\mathcal{I}}(u)] = \text{Work} [\Delta[u^{\alpha,\beta}]],$$

which implies

$$\text{Work} [\mathcal{M}_{\mathcal{I}}(u)] = \sum_{(\alpha,\beta) \in \mathcal{I}} \Delta W_{\alpha,\beta}.$$

The error contribution  $\Delta E_{\alpha,\beta}$  measures instead how much the error  $u - \mathcal{M}_{\mathcal{I}}(u)$  varies if  $\Delta[u^{\alpha,\beta}]$  is added to the estimator  $\mathcal{M}_{\mathcal{I}}$ ,

$$\Delta E_{\alpha,\beta} = \|u - \mathcal{M}_{\mathcal{I} \cup \{\alpha,\beta\}} - u + \mathcal{M}_{\mathcal{I}}\|_U = \|\mathcal{M}_{\mathcal{I} \cup \{\alpha,\beta\}}(u) - \mathcal{M}_{\mathcal{I}}(u)\|_U = \|\Delta[u^{\alpha,\beta}]\|_U.$$

It follows that the error of the CT approximation is bounded by the sum of error contributions not included in the set  $\mathcal{I}$ , that is,

$$(4.1) \quad \text{Error} [\mathcal{M}_{\mathcal{I}}(u)] = \|u - \mathcal{M}_{\mathcal{I}}(u)\|_U = \left\| \sum_{(\alpha,\beta) \notin \mathcal{I}} \Delta[u^{\alpha,\beta}] \right\|_U \leq \sum_{(\alpha,\beta) \notin \mathcal{I}} \Delta E_{\alpha,\beta},$$

provided that the series converges. To construct  $\mathcal{I}$ , we wish to solve a binary knapsack problem, i.e., to maximize the sum of error contributions included in  $\mathcal{I}$  subject to a maximum work  $W_{\max}$  available,

$$(4.2) \quad \begin{aligned} \max_{\chi_{\alpha,\beta} \in \{0,1\}} \quad & \sum_{(\alpha,\beta) \in \mathbb{N}_+^{D+N}} \Delta E_{\alpha,\beta} \chi_{\alpha,\beta} \\ \text{s.t.} \quad & \sum_{(\alpha,\beta) \in \mathbb{N}_+^{D+N}} \Delta W_{\alpha,\beta} \chi_{\alpha,\beta} \leq W_{\max}. \end{aligned}$$

Problem (4.2) leads to a *quasi-optimal* multi-index set  $\mathcal{I}$ , as the sum of the error contributions is only an upper bound of the actual error of the CT approximation. Moreover, the solution of (4.2) is computationally unfeasible. A valid alternative is to relax the integer constraint on  $\chi_{\alpha,\beta}$  and solve the relaxed problem using the Dantzig algorithm [32], which consists in:

- 1 For each hierarchical surplus, compute the profit  $P_{\alpha,\beta} = \frac{\Delta E_{\alpha,\beta}}{\Delta W_{\alpha,\beta}}$ .
- 2 Sort the hierarchical surpluses by decreasing profit.
- 3 Add the hierarchical surpluses to the multi-index set  $\mathcal{I}$  in such an order until the work constraint is satisfied.

Given a tolerance  $\varepsilon$ , the quasi-optimal multi-index set is thus given by

$$(4.3) \quad \mathcal{I}(\varepsilon) = \left\{ (\alpha,\beta) \in \mathbb{N}_+^{N+D} : P_{\alpha,\beta} \geq \varepsilon \right\}.$$

Notice that the above discussion assumes that the error and work contributions are known, but this is not generally the case. To build  $\mathcal{I}$  in practice, a possible solution is an adaptive procedure, such as the one proposed in [12]. The general algorithm is described in Algorithm 4.1.

---

**Algorithm 4.1.** Adaptive incremental construction of  $\mathcal{I}$ .

---

**Require:** Tolerance  $\varepsilon > 0$ .

- 1:  $\mathbf{v} = (1, \dots, 1) \in \mathbb{N}_+^{D+N}$ .
  - 2:  $\mathcal{E} = \{\mathbf{v}\}$ ,  $\mathcal{I} = \emptyset$ ,  $\mathcal{M}_{\mathcal{I}}[u] = 0$ ,  $\text{Err} = 0$ .
  - 3: Compute  $u^{\mathbf{v}}$ .
  - 4: Compute  $\Delta[u^{\mathbf{v}}]$ ,  $\Delta E_{\mathbf{v}}$ ,  $\Delta W_{\mathbf{v}}$ , and  $P_{\mathbf{v}}$  numerically.
  - 5: Set  $\text{Err} = \text{Err} + \Delta E_{\mathbf{v}}$ .
  - 6: **while**  $\text{Err} > \varepsilon$  **do**
  - 7: Select multi-index  $\mathbf{v} = \text{argmax}_{\ell \in \mathcal{E}} P_{\ell}$ .
  - 8: Set  $\mathcal{M}_{\mathcal{I}}[u] = \mathcal{M}_{\mathcal{I}}[u] + \Delta[u^{\mathbf{v}}]$ .
  - 9: Set  $\mathcal{I} = \mathcal{I} \cup \mathbf{v}$ ,  $\mathcal{E} = \mathcal{E} \setminus \mathbf{v}$ , and  $\text{Err} = \text{Err} - \Delta E_{\mathbf{v}}$ .
  - 10: **for**  $k = 1, \dots, D + N$  **do**
  - 11:  $\mathbf{l} = \mathbf{v} + \mathbf{e}_k$ .
  - 12: **if**  $\mathbf{l} - \mathbf{e}_i \in \mathcal{I}$  for every  $i = 1, \dots, D + N$  **then**
  - 13:  $\mathcal{E} = \mathcal{E} \cup \mathbf{l}$ .
  - 14: Compute  $u^{\mathbf{l}}$ .
  - 15: Compute  $\Delta[u^{\mathbf{l}}]$ ,  $\Delta E_{\mathbf{l}}$ ,  $\Delta W_{\mathbf{l}}$ , and  $P_{\mathbf{l}}$  numerically.
  - 16: Set  $\text{Err} = \text{Err} + \Delta E_{\mathbf{l}}$ .
  - 17: **end if**
  - 18: **end for**
  - 19: **end while**
  - 20: Output:  $\mathcal{I}$ ,  $\mathcal{M}_{\mathcal{I}}[u]$ , and  $\text{Err}$ .
- 

This iterative algorithm computes numerically the error contributions  $\Delta E_{\alpha, \beta} = \|\Delta u^{\alpha, \beta}\|_U$  for every multi-index  $(\alpha, \beta)$  in the reduced margin of a current index set, where the reduced margin of a set  $\mathcal{I}$  of multi-indices in  $\mathbb{N}_+^n$  is

$$\mathcal{R}_{\mathcal{I}} := \{\mathbf{v} \in \mathbb{N}_+^n : \mathbf{v} - \mathbf{e}_k \in \mathcal{I} \text{ for all } k \in \{1, \dots, n\}, \text{ with } v_k > 1\}.$$

The algorithm then updates  $\mathcal{I} = \mathcal{I} \cup (\bar{\alpha}, \bar{\beta})$ , where  $(\bar{\alpha}, \bar{\beta})$  is the multi-index in  $\mathcal{R}_{\mathcal{I}}$  associated to the largest profit. The adaptive construction ends when the sum of the error contributions in the reduced margin is below a desired tolerance, mimicking the error upper bound (4.1).

Despite its simplicity, this adaptive algorithm may not be efficient. The cost of computing the CT approximation may be dominated by the cost of constructing the index set  $\mathcal{I}$ , since evaluating numerically an error contribution (line 15 of Algorithm 4.1) for a given multi-index in the reduced margin requires one to solve an OCP (line 14). Notice that once  $u^{\mathbf{v}}$  has been computed,  $\Delta u^{\mathbf{v}}$  requires only a linear combination of previously computed solutions. A possible alternative is to use an a priori ansatz for  $\Delta E_{\alpha, \beta}$  and  $\Delta W_{\alpha, \beta}$ , so that the computation of the profits in the margin is reduced to the evaluation of a simple algebraic formula and an OCP is solved exclusively when a selected multi-index is added to  $\mathcal{I}$ . This variant is described by Algorithm 4.2. We will denote the construction of  $\mathcal{I}$  based on these adaptive/a priori frameworks as “incremental CT.”

---

**Algorithm 4.2.** A priori incremental construction of  $\mathcal{I}$ .

---

**Require:** Tolerance  $\varepsilon > 0$ .

- 1:  $\mathbf{v} = (1, \dots, 1) \in \mathbb{N}_+^{D+N}$ .
  - 2:  $\mathcal{E} = \{\mathbf{v}\}$ ,  $\mathcal{I} = \emptyset$ ,  $\mathcal{M}_{\mathcal{I}}[u] = 0$ ,  $\text{Err} = 0$ .
  - 3: Compute  $\Delta E_{\mathbf{v}}$ ,  $\Delta W_{\mathbf{v}}$ , and  $P_{\mathbf{v}}$  using ansatzes.
  - 4: Set  $\text{Err} = \text{Err} + \Delta E_{\mathbf{v}}$ .
  - 5: **while**  $\text{Err} > \varepsilon$  **do**
  - 6:   Select multi-index  $\mathbf{v} = \text{argmax}_{\ell \in \mathcal{E}} P_{\ell}$ .
  - 7:   Compute  $\mathbf{u}^{\mathbf{v}}$  and  $\Delta[u^{\mathbf{v}}]$ .
  - 8:   Set  $\mathcal{M}_{\mathcal{I}}[u] = \mathcal{M}_{\mathcal{I}}[u] + \Delta[u^{\mathbf{v}}]$ .
  - 9:    $\mathcal{I} = \mathcal{I} \cup \mathbf{v}$  and  $\mathcal{E} = \mathcal{E} \setminus \mathbf{v}$  and  $\text{Err} = \text{Err} - \Delta E_{\mathbf{v}}$ .
  - 10: **for**  $k = 1, \dots, D + N$  **do**
  - 11:    $\mathbf{l} = \mathbf{v} + \mathbf{e}_k$ .
  - 12:   **if**  $\mathbf{l} - \mathbf{e}_i \in \mathcal{I}$  for every  $i = 1, \dots, D + N$  **then**
  - 13:      $\mathcal{E} = \mathcal{E} \cup \mathbf{l}$ .
  - 14:     Compute  $\Delta E_{\mathbf{l}}$ ,  $\Delta W_{\mathbf{l}}$ , and  $P_{\mathbf{l}}$  using ansatzes.
  - 15:     Set  $\text{Err} = \text{Err} + \Delta E_{\mathbf{l}}$ .
  - 16:   **end if**
  - 17: **end for**
  - 18: **end while**
  - 19: Output:  $\mathcal{I}$  and  $\mathcal{M}_{\mathcal{I}}[u]$ .
- 

We remark that a priori ansatzes on  $\Delta E_{\alpha, \beta}$  and  $\Delta W_{\alpha, \beta}$  permit us to build the multi-index  $\mathcal{I}$  set beforehand using formula (4.3). In this way, one could use the CT formula (3.10), and reduce the number of OCPs that have to be solved, by considering only the multi-indices such that  $c_{\alpha, \beta} \neq 0$ . This approach is more efficient but less flexible, since it is inherently nonadaptive. If  $\mathcal{M}_{\mathcal{I}(\varepsilon)}[u]$  has been computed, it may not be possible to evaluate  $\mathcal{M}_{\mathcal{I}(\varepsilon')}[u]$ , with  $\varepsilon' > \varepsilon$ , as  $\mathcal{M}_{\mathcal{I}(\varepsilon')}[u]$  may involve multi-indices that were inactive (i.e.,  $c_{\alpha, \beta} = 0$ ) for  $\mathcal{M}_{\mathcal{I}(\varepsilon)}[u]$ . Further, it does not have an intrinsic stopping criterion, contrary to the adaptive/incremental CT procedure. One could also use a posteriori error estimators for  $\Delta E_{\alpha, \beta}$ , generalizing the work of [16] for random PDEs. However, the derivation of suitable a posteriori error estimators for OCP constrained by random PDEs combined with the CT is still an open problem.

In the next subsection, we assume explicit analytic expressions for the error and work contributions and discuss an a priori construction of the quasi-optimal set  $\mathcal{I}$ . These ansatzes contain few parameters that are problem-dependent and have to be estimated case by case. A complexity analysis of the CT approximation based on this proposed a priori construction is presented in subsection 4.2. The soundness of these expressions has been verified theoretically in the context of random PDEs under suitable regularity assumptions; see [18]. We will verify numerically their validity on a model problem in section 5.1.

**4.1. An a priori construction.** To build the set  $\mathcal{I}$  in a priori fashion, we need some hypothesis on the decay of the error and work contributions. Following the framework proposed in [19] for the MISC method, we make the following assumptions.

*Assumption 3.* The discretization parameters  $h_{n,\alpha_n}$  depend exponentially on the discretization level  $\alpha_n$ , while the number of collocation points grows linearly with the stochastic level  $\beta_n$ ,

$$(4.4) \quad h_{n,\alpha_n} = h_0 2^{-\alpha_n} \quad \text{and} \quad m(\beta_n) = \beta_n.$$

Notice that we assume that the level-to-nodes relation  $m(\beta_n)$  is linear, in contrast to several works concerning random PDEs; see, e.g., [19, 35], where the doubling map  $m(\beta_n) = 2^{\beta_n} - 1$  is considered. The doubling map is often used together with Clenshaw–Curtis nodes, since the quadrature/interpolation nodes are then nested, which permits one to recycle several precalculated solutions when computing the quadrature/interpolation for a higher level  $\beta$ . In our setting, each multi-index  $\beta$  requires the solution of a robust OCP, and none of the previously computed optimal controls (neither state or adjoint variables) can be recycled. Numerically, we have observed that the linear and doubling mappings lead to the same convergence rate, but the former has a better constant.

In addition to  $\Delta W_{\alpha,\beta}$  and  $\Delta E_{\alpha,\beta}$  associated to the mixed hierarchical surpluses, for a fixed  $\bar{\beta} \in \mathbb{N}_+^N$ , we denote with  $\Delta W_{\alpha,\bar{\beta}}^{\text{det}}$  and  $\Delta E_{\alpha,\bar{\beta}}^{\text{det}}$  the work and error contributions associated to the deterministic hierarchical surplus (3.7). Similarly, for a fixed  $\bar{\alpha} \in \mathbb{N}_+^D$ , we denote with  $\Delta W_{\bar{\alpha},\beta}^{\text{stoc}}$  and  $\Delta E_{\bar{\alpha},\beta}^{\text{stoc}}$  the work and error contributions associated to the deterministic hierarchical surplus (3.8).

*Assumption 4 (assumption on the work contributions).* The work contributions associated to the deterministic, stochastic, and mixed hierarchical surpluses satisfy<sup>1</sup>

$$(4.5) \quad \Delta W_{\alpha,\bar{\beta}}^{\text{det}} \leq C_{\text{work}}^{\text{det},\bar{\beta}} \prod_{n=1}^D 2^{\alpha_n \tilde{\gamma}_n},$$

$$(4.6) \quad \Delta W_{\bar{\alpha},\beta}^{\text{stoc}} \leq C_{\text{work}}^{\text{stoc},\bar{\alpha}} \prod_{n=1}^N (\beta_n + 1),$$

$$(4.7) \quad \Delta W_{\alpha,\beta} \leq C_{\text{work}} \left( \prod_{n=1}^D 2^{\alpha_n \tilde{\gamma}_n} \right) \left( \prod_{n=1}^N (\beta_n + 1) \right)$$

for some rates  $\tilde{\gamma}_n$ .

*Assumption 5 (assumption on the error contributions).* The error contributions associated to the deterministic, stochastic, and mixed hierarchical surpluses satisfy

$$(4.8) \quad \Delta E_{\alpha,\bar{\beta}}^{\text{det}} \leq C_{\text{error}}^{\text{det},\bar{\beta}} \prod_{n=1}^D 2^{-\alpha_n \tilde{r}_n},$$

$$(4.9) \quad \Delta E_{\bar{\alpha},\beta}^{\text{stoc}} \leq C_{\text{error}}^{\text{stoc},\bar{\alpha}} \prod_{n=1}^N e^{-\tilde{g}_n(\beta_n+1)},$$

$$(4.10) \quad \Delta E_{\alpha,\beta} \leq C_{\text{error}} \left( \prod_{n=1}^D 2^{-\alpha_n \tilde{r}_n} \right) \left( \prod_{n=1}^N e^{-\tilde{g}_n(\beta_n+1)} \right)$$

for some rates  $\tilde{r}_i$  and  $\tilde{g}_i$ .

<sup>1</sup>The work bound involves  $\prod_{n=1}^N (\beta_n + 1)$  instead of the more natural  $\prod_{n=1}^N \beta_n$  to simplify the proofs of Theorems 4.1 and 4.2. Nevertheless, this does not influence the asymptotic analysis, since  $\prod_{n=1}^N \beta_n \leq \prod_{n=1}^N (\beta_n + 1) \leq 2^N \prod_{n=1}^N \beta_n$ .

Assumptions 4 and 5 specify how the work and error contributions of the deterministic and stochastic hierarchical surpluses depend on the multi-indices  $\alpha$  and  $\beta$ . They assume a multiplicative structure of the work and error contributions of the mixed hierarchical surplus. Notice that Assumption 5 guarantees that  $\{\Delta E_{\alpha,\beta}\}_{(\alpha,\beta) \in \mathbb{N}_+^{N+D}} \in l^1(\mathbb{N}_+^{N+D})$ , so that, together with the convergence of the minimizers of the discrete OCPs to  $u$ , it validates the hierarchical representation (3.9). The discussion of the rationale behind these assumptions is postponed to section 5.1, where numerical evidence of their validity together with a procedure to estimate the rates  $\tilde{r}_i$  and  $\tilde{g}_i$  are presented.

**4.2. Complexity analysis.** In this subsection, we present a complexity analysis of the CT in two different settings. We first consider the simpler case in which the spatial discretization is fixed, i.e., the multi-index  $\alpha$  is an assigned, constant, multi-index  $\alpha = \bar{\alpha}$ . In other words, the CT combines the solutions of (3.3) for different quadrature formulae, but all solutions are computed on the same spatial mesh. Hence, the CT involves exclusively the stochastic hierarchical surpluses,

$$(4.11) \quad \mathcal{M}_{\mathcal{I}}(u) = \sum_{\beta \in \mathcal{I}} \Delta^{\text{stoc}}[u^{\bar{\alpha},\beta}].$$

In the second setting, we consider the more general case where both  $\alpha$  and  $\beta$  are variables in the CT expansion as in (3.9). In both cases, the complexity analysis is based on a “direct counting” argument that consists in summing the error contributions outside the index set  $\mathcal{I}$  and the work contributions inside  $\mathcal{I}$ ; see [19, 35].

Let us start with the first setting. As  $\bar{\alpha}$  is fixed, to each multi-index  $\beta$  we associate the profit

$$P_{\beta} = \frac{\Delta E_{\bar{\alpha},\beta}^{\text{stoc}}}{\Delta W_{\bar{\alpha},\beta}^{\text{stoc}}},$$

and given a tolerance  $\varepsilon > 0$ , the index set  $\mathcal{I}$  simplifies to

$$\mathcal{I}(\varepsilon) := \left\{ \beta \in \mathbb{N}_+^N : \frac{\Delta E_{\bar{\alpha},\beta}^{\text{stoc}}}{\Delta W_{\bar{\alpha},\beta}^{\text{stoc}}} \geq \varepsilon \right\}.$$

Due to Assumptions 4 and 5,  $\mathcal{I}(\varepsilon)$  coincides with

$$\mathcal{I}(L) := \{ \beta \in \mathbb{N}_+^N : \tilde{g} \cdot (\beta + \mathbf{1}) + |\log(\beta + \mathbf{1})|_1 \leq L \},$$

where  $|\log(\beta + \mathbf{1})|_1 = \sum_{i=1}^N \log(\beta_i + 1)$ , and  $L := L(\varepsilon) = -\log(\varepsilon \frac{C_{\text{work},\bar{\alpha}}^{\text{stoc}}}{C_{\text{error},\bar{\alpha}}^{\text{stoc}}})$ .

The next theorem provides a complexity result for the CT applied only on the stochastic discretization multi-index  $\beta$ . The proof is detailed in Appendix 7.1.

**Theorem 4.1.** *There exist constants  $C_1$  and  $C_2$  such that for any  $W_{\max}$  satisfying  $W_{\max} \geq \frac{|\tilde{g}|^{2N} C_1}{(2N)!}$ , and choosing  $\hat{L} = \sqrt[2N]{\frac{W_{\max}(2N)!}{C_1}} - |\tilde{g}|$ ,*

$$\text{Work} [\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] \leq W_{\max},$$

$$\text{Error} [\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] \leq C_2 e^{-2N \sqrt[2N]{\frac{W_{\max}(2N)!}{C_1}}} \left( \sqrt[2N]{\frac{W_{\max}(2N)!}{C_1}} + 1 \right)^{2N-1}.$$

Numerically, we observed that the simplified estimate, based on Stirling approximation,

$$(4.12) \quad \text{Error} [\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] \leq C e^{-\gamma N} 2^{2N} \sqrt{W_{\max}},$$

for  $C, \gamma \in \mathbb{R}^+$  captures well the convergence (see Figure 3). Notice that the complexity result expressed in (4.12) is the same one found in [35] for the complexity of sparse grid interpolants (on non-nested grids) for elliptic random PDEs.

We next consider the general case in which the combination technique is applied to both the spatial and the stochastic discretization parameters. Under Assumptions 4, 5 and defining  $\mathbf{r} = \log(2)\tilde{\mathbf{r}}$ ,  $\boldsymbol{\gamma} = \log(2)\tilde{\boldsymbol{\gamma}}$ , for a given tolerance  $\varepsilon$ , the quasi-optimal set (4.3) becomes

$$\mathcal{I}(L) := \left\{ (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{N}_+^{D+N} : (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha} + \tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})|_1 \leq L \right\},$$

with  $L = -\log(\varepsilon \frac{C_{\text{work}}}{C_{\text{error}}})$ . The next theorem provides an asymptotic convergence result for the CT applied to the OCP (3.4). The proof is detailed in Appendix 7.2.

**Theorem 4.2.** *Let  $r_j = \log(2)\tilde{r}_j$ ,  $\gamma_j = \log(2)\tilde{\gamma}_j$ ,  $j = 1, \dots, N$ ,  $\boldsymbol{\Theta} = (\frac{\gamma_1}{r_1+\gamma_1}, \dots, \frac{\gamma_D}{r_D+\gamma_D})$ ,  $\mu = \min_n \frac{r_n}{\gamma_n}$ ,  $\chi = \max_n \boldsymbol{\Theta}_n$ , and  $n(\boldsymbol{\Theta}, \chi) := \#\{n : \boldsymbol{\Theta}_n = \chi\}$ . There exists a constant  $C_W$  such that for any  $W_{\max}$  satisfying  $W_{\max} \geq C_W e^\chi$ , and setting*

$$(4.13) \quad L = L(W_{\max}) = \frac{1}{\chi} \left( \log \left( \frac{W_{\max}}{C_W} \right) - (n(\boldsymbol{\Theta}, \chi) - 1) \log \left( \frac{1}{\chi} \log \left( \frac{W_{\max}}{C_W} \right) \right) \right),$$

the combination technique solution satisfies

$$(4.14) \quad \text{Work} [\mathcal{M}_{\mathcal{I}(L(W_{\max}))}] \leq W_{\max},$$

$$(4.15) \quad \limsup_{W_{\max} \rightarrow \infty} \frac{\text{Error} [\mathcal{M}_{\mathcal{I}(L(W_{\max}))}]}{W_{\max}^{-\mu} (\log(W_{\max}))^{(\mu+1)(n(\boldsymbol{\Theta}, \chi)-1)}} = C < \infty.$$

Notice that Theorem 4.2 provides a characterization of the asymptotic complexity. As we discuss in section 5, the pre-asymptotic behavior can be slightly different. Second, the asymptotic complexity depends exclusively on the rates associated to the *spatial* work and error contribution. The asymptotic complexity is not affected by the rates of the stochastic variables. This can be understood noticing that, while both the spatial and the stochastic errors decay exponentially with respect to  $\alpha_i$  and  $\beta_i$ , the spatial work grows exponentially with respect to  $\alpha_i$ , whereas the stochastic work grows only linearly with respect to  $\beta_i$ . Third, the asymptotic complexity of the CT provided by Theorem 4.2 is identical to the MISC complexity result presented in [19], and thus the CT applied to OCPs have the same asymptotic convergence of the MISC method for the solution of PDEs in the context of forward UQ.

**Remark 4 (infinite-dimensional setting).** The complexity analysis presented here cannot be readily generalized to the infinite-dimensional setting ( $N = \infty$ ), since both theorems involve constants that explode as  $N \rightarrow \infty$ . To handle the infinite-dimensional case, a refined analysis is required based on the summability of the sequence  $\{\tilde{g}_n\}_{n=1}^\infty$  in appropriate  $l_p$  spaces, following, e.g., [18]. Nevertheless, from the implementation point of view, the CT can be extended

to the infinite-dimensional setting by modifying Algorithm 4.2 in the loop 10–17, so that only a subset of the reduced margin involving a finite number of dimensions is explored at each iteration. This improved algorithm would automatically balance the quadrature error, the spatial discretization error, and the truncation error that is commonly committed by considering only a finite number of random parameters (as in truncated Karhunen–Loève expansions of random fields). For more details, we refer the reader to [17] for an analysis concerning truncation error and to [6, 11, 15] for algorithmic aspects.

**5. Numerical section.** In this section, we test the effectiveness of the CT on a model problem. In all experiments, we consider the OCP (2.6) with force term  $\phi = 1$  and with the diffusion coefficient

$$(5.1) \quad \kappa(x, \zeta) = e^{\sum_{i=1}^N \zeta_n \lambda_n \psi_n(\mathbf{x})},$$

where  $\zeta_n \sim \mathcal{U}(-1, 1)$  and  $\lambda_n = \sqrt{3}e^{-0.6n}$ . The spatial domain is  $\mathcal{D} = (0, 1)^d$ , with  $d = 1, 2$ . In the one-dimensional case,  $\psi_n(x) = \phi_{i_1(n)}(x)$ , while in the two-dimensional case,  $\psi_n(\mathbf{x}) = \phi_{i_1(n)}(x_1)\phi_{i_2(n)}(x_2)$ , where

$$(5.2) \quad \phi_n(x) = \begin{cases} \sin(\frac{n}{2}\pi x) & \text{if } n \text{ is even,} \\ \cos(\frac{n-1}{2}\pi x) & \text{if } n \text{ is odd.} \end{cases}$$

The maps  $i_j : \mathbb{N}^+ \rightarrow \mathbb{N}^+$ ,  $j = 1, 2$ , are detailed in Table 1. Notice that (5.1) leads to a well-posed state equation, as

$$k_{\min} := e^{-\sum_{n=1}^N \lambda_n} \leq \kappa(x, \zeta) \leq e^{\sum_{n=1}^N \lambda_n} =: k_{\max},$$

so that Assumptions 1 and 2 are satisfied, even in the limit  $N \rightarrow \infty$ . As a quantity of interest, we consider  $F(y(\zeta)) = \frac{1}{2}\|y(\zeta) - y_d\|_{L^2(D)}^2$ , with  $y_d := \prod_{i=1}^d \sin(\pi x_i)$ . Note that the precise expression of  $y_d$  does not influence the stochastic regularity of the problem, since it is deterministic. However, a less regular target state may influence the spatial regularity. For reproducibility, data and codes are available in [46].

**5.1. Discussion on the validity of the assumptions.** In this subsection, we discuss the rationale behind the assumptions made in section 4.1 and verify them on the model problem described in Example 1.

To satisfy Assumption 3, we consider a spatial mesh obtained as the tensor product of one-dimensional meshes, each characterized by a uniform mesh size  $h_{i,\alpha_i} = 2^{-\alpha_i-1}$ , and hence  $h_0 = 2^{-1}$ . In one dimension, we employ the Lagrange  $\mathbb{P}^1$  finite element space, while in two dimensions, we use the  $\mathcal{Q}^1$  bilinear finite element space; see [10]. The total number of degrees of freedom on the mesh associated to  $\alpha$  is  $N_\alpha = \prod_{n=1}^D (2^{\alpha_n+1} - 1) \leq C_D \prod_{n=1}^D 2^{\alpha_n}$ . Further, we

**Table 1**  
Definition of the maps  $i_1$  and  $i_2$ .

$n$	1	2	3	4	5	6	7	8	9	10	...
$i_1(n)$	1	2	1	3	2	1	4	3	2	1	...
$i_2(n)$	1	1	2	1	2	3	1	2	3	4	...

consider a linear level-to-nodes map to define the tensor product quadrature rule. The total number of Gauss–Legendre quadrature points is  $M_\beta = \prod_{n=1}^N \beta_n$ .

Next, we analyze Assumption 4 on the work contributions. To compute the optimal control  $u^{\alpha,\beta}$ , we solve the full-space discrete optimality system (see [33, 36] for a derivation),

$$(5.3) \quad \begin{pmatrix} \mathcal{M} & 0 & \mathcal{A} \\ 0 & \nu M_s & -M_s E^\top \\ \mathcal{A} & -EM_s & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_p \\ 0 \\ \mathbf{f}_y \end{pmatrix},$$

where  $\mathcal{A} = \text{diag}(w_1 A(\bar{\zeta}_1), \dots, w_{M_\beta} A(\bar{\zeta}_{M_\beta}))$  and  $A(\bar{\zeta}_n) \in \mathbb{R}^{N_\alpha \times N_\alpha}$  is the stiffness matrix associated to the quadrature node  $\bar{\zeta}_n$  and  $w_n$  is its quadrature weight,  $M_s \in \mathbb{R}^{N_\alpha \times N_\alpha}$  is the mass matrix,  $\mathcal{M} = \text{diag}(w_1 M_s, \dots, w_{M_\beta} M_s)$ , and  $E = (w_1 I_s, \dots, w_{M_\beta} I_s)^\top$ , where  $I_s \in \mathbb{R}^{N_\alpha \times N_\alpha}$  is the identity matrix. Note that the dimensions of  $\mathcal{A}$ ,  $\mathcal{M}$ , and  $E$  depend on the multi-index  $(\alpha, \beta)$ . We solve this large (symmetric) saddle-point system using MINRES preconditioned by the block diagonal preconditioner analyzed in [36]. For a fixed value of the regularization parameter  $\nu$ , this preconditioner leads to a robust convergence with respect to the mesh size and the number of collocation points, so that the number of Krylov iterations can be considered constant. The major cost of each iteration is the preconditioning of the  $2M_\beta$  stiffness matrices. Alternatively to (5.3), we may perform a Schur complement on  $u$  and solve the reduced optimality system

$$(5.4) \quad \left( \nu M_s + \sum_{n=1}^{M_\beta} w_n (M_s A^{-1}(\bar{\zeta}_n) M_s A^{-1}(\bar{\zeta}_n) M_s) \right) \mathbf{u} = \mathbf{g}.$$

Each conjugate gradient iteration then requires one to invert  $2M_\beta$  stiffness matrices. In spite of the full-space/reduced approach used, the overall cost for solving the OCP then depends linearly on the number of collocation points  $M_\beta$ , and possibly nonlinearly (depending on the preconditioner used) on the size of the finite element space, that is,

$$(5.5) \quad \text{Work} [u^{\alpha,\beta}] \leq \widehat{C} \left( \prod_{n=1}^N \beta_n \right) \left( \prod_{n=1}^D 2^{\alpha_n} \right)^\theta$$

for a real parameter  $\theta$  and constant  $\widehat{C}$ . The deterministic and stochastic work contributions thus satisfy

$$\Delta W_{\alpha,\beta}^{\text{det}} = \sum_{j \in \{0,1\}^D} \text{Work} [u^{\alpha-j,\bar{\beta}}] \leq \widehat{C} \sum_{j \in \{0,1\}^D} \left( \prod_{n=1}^N \bar{\beta}_n \right) \left( \prod_{n=1}^D 2^{\alpha_n - j_n} \right)^\theta \leq C_{\text{work}}^{\text{det},\bar{\beta}} \left( \prod_{n=1}^D 2^{\alpha_n} \right)^\theta,$$

with  $C_{\text{work}}^{\text{det},\bar{\beta}} = \widehat{C} (1 + 2^{-\theta})^D (\prod_{n=1}^N \bar{\beta}_n)$  for a fixed  $\bar{\beta}$ , and

$$\begin{aligned} \Delta W_{\bar{\alpha},\beta}^{\text{stoc}} &= \sum_{j \in \{0,1\}^N} \text{Work} [u^{\bar{\alpha},\beta-j}] \leq \widehat{C} \sum_{j \in \{0,1\}^N} \left( \prod_{n=1}^N (\beta_n - j_n) \right) \left( \prod_{n=1}^D 2^{\bar{\alpha}_n} \right)^\theta \\ &\leq C_{\text{work}}^{\text{stoc},\bar{\alpha}} \left( \prod_{n=1}^N (\beta_n + 1) \right), \end{aligned}$$

where  $C_{\text{work}}^{\text{stoc}, \bar{\alpha}} = \widehat{C} 2^N (\prod_{n=1}^d 2^{\bar{\alpha}_n})^\theta$  for a given  $\bar{\alpha}$ . We are left to check (4.7), but indeed

$$\begin{aligned} \Delta W_{\alpha, \beta} &= \sum_{\mathbf{i} \in \{0,1\}^{D+N}} \text{Work} [u^{(\alpha, \beta) - \mathbf{i}}] \leq \sum_{\mathbf{k} \in \{0,1\}^D} \sum_{\mathbf{j} \in \{0,1\}^N} \widehat{C} \left( \prod_{n=1}^D 2^{\alpha_n - j_n} \right)^\theta \left( \prod_{n=1}^N (\beta_n - j_n) \right) \\ &\leq \widehat{C} 2^N \left( \prod_{n=1}^N (\beta_n + 1) \right) \sum_{\mathbf{k} \in \{0,1\}^D} \left( \prod_{n=1}^D 2^{\alpha_n - j_n} \right)^\theta \\ &\leq C_{\text{work}} \left( \prod_{n=1}^N (\beta_n + 1) \right) \left( \prod_{n=1}^D 2^{\alpha_n} \right)^\theta, \end{aligned}$$

with  $C_{\text{work}} = 2^N (1 + 2^\theta)^D \widehat{C}$ .

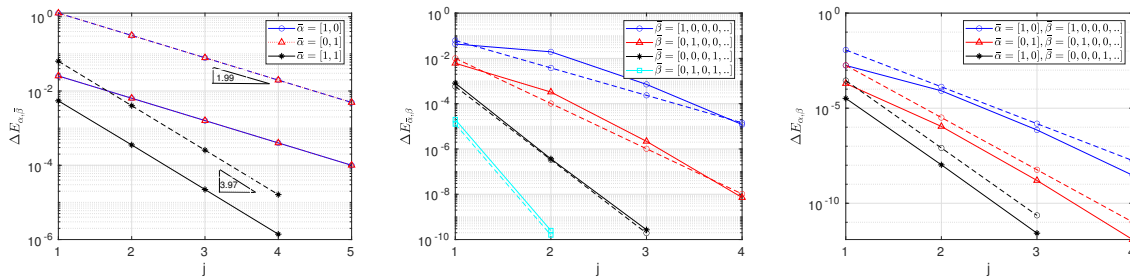
Notice that (4.5) and (4.7) are trivially satisfied setting  $\tilde{\gamma}_i = \theta$  for every  $i$ . In our numerical experiments, we used the full-space approach and inverted directly the  $2M_\beta$  matrices using the built-in MATLAB sparse banded matrix solver and found that  $\theta = 1$  gives a good description of the increase of the computational time with respect to the size of the finite element space.

Next, we focus on Assumption 5. On the one hand, (4.8) is a direct consequence of classical finite element error analysis combined with the spatial combination technique convergence theory, provided that  $u$  admits sufficient Sobolev mixed-order regularity [13, 14]. However, for general domains and/or with control constraints, the control might not enjoy such high regularity. In these cases, the CT could still be applied either on a fixed spatial mesh, or by letting  $\alpha$  be a scalar setting the overall mesh size, as in standard multilevel approaches [21, 45].

Concerning (4.9), it is well known that hierarchical surpluses satisfy such a hypothesis in the context of sparse grid interpolants for forward UQ problems, provided that the solution map of the random PDE is holomorphic in a Bernstein ellipse in the complex plane [8, 19, 35]. In our setting, we intuitively remark that (4.9) requires the map  $\zeta \in \Gamma \rightarrow p(\zeta) \in V$  to be holomorphic in a region of the complex plane, which in turn requires holomorphic regularity of the quantity of interest  $F$  and of the inverse of the operator  $e$  with respect to  $\zeta$ .

Finally, (4.10) assumes a product structure of the decay of the mixed spatial and stochastic surpluses. For forward UQ problems, this property has been extensively studied and exploited in several works; see, e.g., [3, 18–21, 43]. The rigorous extension of these results to the present context requires a highly technical theoretical analysis that is beyond the scope of this work, and it is the subject of a forthcoming manuscript.

Here we limit ourselves to verify (4.8), (4.9), and (4.10) numerically, and we estimate the rates  $\tilde{r}_i$  and  $\tilde{g}_i$ . To do so, we fix a multi-index  $\beta = \bar{\beta}$ , set  $\alpha = \mathbf{1} + j\bar{\alpha}$ , and fit the decay of  $\Delta E_{\alpha, \beta}$  for some different values of  $j$ . Similarly, we fix a  $\alpha = \bar{\alpha}$  and set  $\beta = \mathbf{1} + j\bar{\beta}$  and fit the decay of  $\Delta E_{\bar{\alpha}, \beta}$ . The mixed decays are fitted with the same procedure. The results are reported in Figure 1, which clearly shows that the assumptions are verified for the model problem. In our setting, we found  $\tilde{r}_i = 2$ , for all  $i$ , while the rates  $\tilde{g}_i$  are reported in Table 2. Notice that the estimated rates  $\tilde{g}_i$  may depend on the geometrical mesh used, but the dependence is quite mild (see [3] for a procedure to update the estimate of  $\tilde{g}_i$  on the fly).



**Figure 1.** Numerical validation of assumption (4.8), (4.9), and (4.10) on the decay of the error contributions. Spatial error contributions with  $\bar{\beta} = \mathbf{1}$  (left), stochastic error contributions with  $\bar{\alpha} = (3, 3)$  (center) for the first four random variables, and mixed spatial and stochastic error contributions (right). The solid lines are based on computed values, and the dashed lines are the fitted ansatzes.

**Table 2**

Fitted rates  $\tilde{g}_i$  for the first 10 random variables in the expansion of (5.1).

$\tilde{g}_1$	$\tilde{g}_2$	$\tilde{g}_3$	$\tilde{g}_4$	$\tilde{g}_5$	$\tilde{g}_6$	$\tilde{g}_7$	$\tilde{g}_8$	$\tilde{g}_9$	$\tilde{g}_{10}$	...
2.78	4.59	5.79	7.45	8.18	9.85	10.97	12.98	14.18	14.57	...

**5.2. Numerical tests.** We now show the performance of the CT to solve (2.6). To construct the multi-index sets, we rely on the sparse-grid MATLAB Kit [2]. We first consider the CT approximation (4.11) applied exclusively on the quadrature formula of the objective functional. For every multi-index  $\beta$ , we use the same physical discretization determined by a fixed  $\bar{\alpha}$ . More specifically, we compare five methods:

- Meth. 1 An “a priori incremental” algorithm in which  $\mathcal{I}$  is built adaptively by using the ansatzes (4.6) and (4.9) to compute the profits for all multi-indices in the reduced margin. The parameters  $\{\tilde{g}_n\}_n$  are estimated beforehand, and this cost is not considered. This algorithm corresponds to Algorithm 4.2.
- Meth. 2 The well-known adaptive algorithm [12] in which  $\mathcal{I}$  is built adaptively computing numerically  $\Delta E_{\bar{\alpha},\beta}$  (and thus also the profits  $P_{\bar{\alpha},\beta}$ ) for all multi-indices in the reduced margin. We keep track only of the work done to solve the OCPs that actually contribute to the CT solution (3.9), thus neglecting the cost of exploring the reduced margin (which might dominate the overall cost). In this way, the adaptive algorithm can be considered as a benchmark for the a priori CT. This algorithm corresponds to Algorithm 4.1.
- Meth. 3 An a priori algorithm based on the CT formulation (3.10). In particular, given the multi-index set  $\mathcal{I}^k$  built at the  $k$ th step of the a priori incremental algorithm (Meth. 1), we check for which  $\beta$ ,  $c_\beta \neq 0$ , and consequently compute the approximation using (3.10). The approximation will be identical to the solution of Meth. 1. However, the work performed will be less, as several multi-indices  $\beta$  do not contribute to the approximation.
- Meth. 4 An anisotropic tensor product approximation: given a sequence of levels  $L$ , we set  $\beta_n = 1 + \lfloor \frac{L}{\tilde{g}_n} \rfloor$ . Notice that the anisotropic tensor product approximation can be recast in the CT framework, using the multi-index set  $\mathcal{I}(L) = \{\beta \in \mathbb{N}^N : \max_n \tilde{g}_n(\beta_n - 1) \leq L\}$ ; see [2].

Meth. 5 A standard sparse grid method (see Remark 2) based on the multi-index set  $\mathcal{I}(L) = \{\beta \in \mathbb{N}^N : \sum_{i=1}^N \tilde{g}_n(\beta_n - 1) \leq L\}$  with Gauss–Legendre nodes and a linear level-to-node map.

Figure 2 shows the complexity (error vs. work) of the different methods for  $N = 2, 6, 10$  random variables and  $d = 2$ . The work reported in the  $x$ -axis corresponds to  $W = \sum_{\beta \in \mathcal{I}} M_\beta$  for Meth. 1 and Meth. 2,  $W = \sum_{\beta \in \mathcal{I}: c_\beta \neq 0} M_\beta$  for Meth. 3,  $W = M_\beta$  for Meth. 4, and  $W = |\tilde{\Lambda}_\beta|$ , i.e., the number of nodes involved in the sparse grid quadrature, for Meth. 5. The error corresponds to the  $L^2$  norm between the current approximation and a reference overkilled solution computed with Algorithm 4.2.

Notice that the a priori algorithms converge very similarly to the adaptive algorithm, thus confirming the effectiveness of the a priori construction based on the model-fitted ansatzes. Generally, the CT needs a few random variables in order to be more effective than anisotropic full tensor approximations, as sparse approximations are more efficient for sufficiently high dimensional problems [5].

The convergence of the CT is essentially unchanged moving from  $N = 6$  to  $N = 10$  random variables. Indeed, the random variables are increasingly less important, being weighted in (5.1) by  $\lambda_n$ , which tends to zero exponentially, and their rates  $\tilde{g}_i$  are increasingly larger (see Table 2). Hence, the addition of a new random variable requires only computing quite cheap hierarchical surplus  $\Delta^\beta u$ . This can be observed in the left and center panels of Figure 3, which show three components of the multi-indices  $\beta$  included by Meth. 2 for  $d = 2$  and  $N = 10$ . We clearly observe that whenever  $\beta_9$  or  $\beta_{10}$  are greater than 1, all the remaining components of

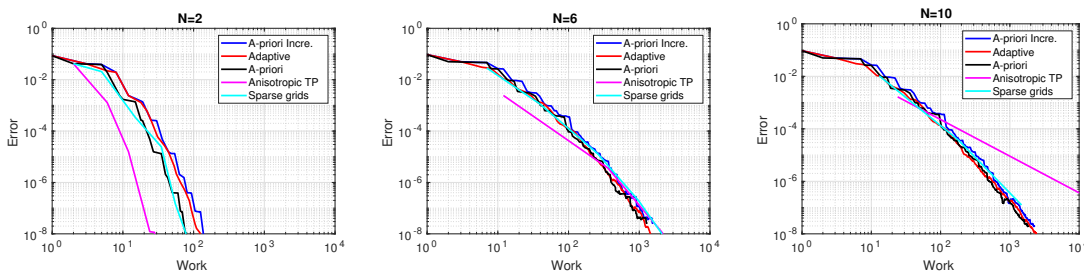


Figure 2. Convergence behavior of the different methods for  $N = 2, 6, 10$  random variables.

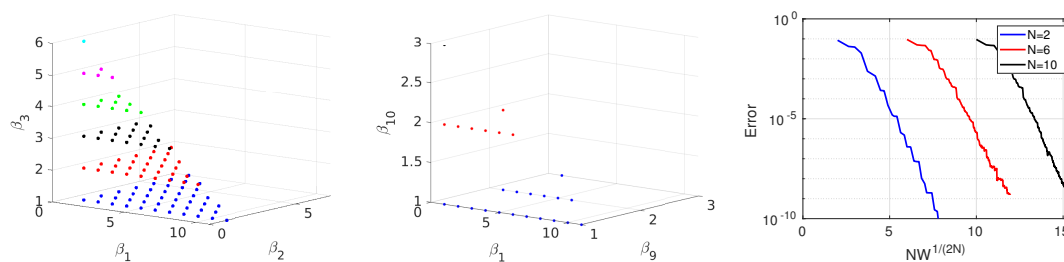
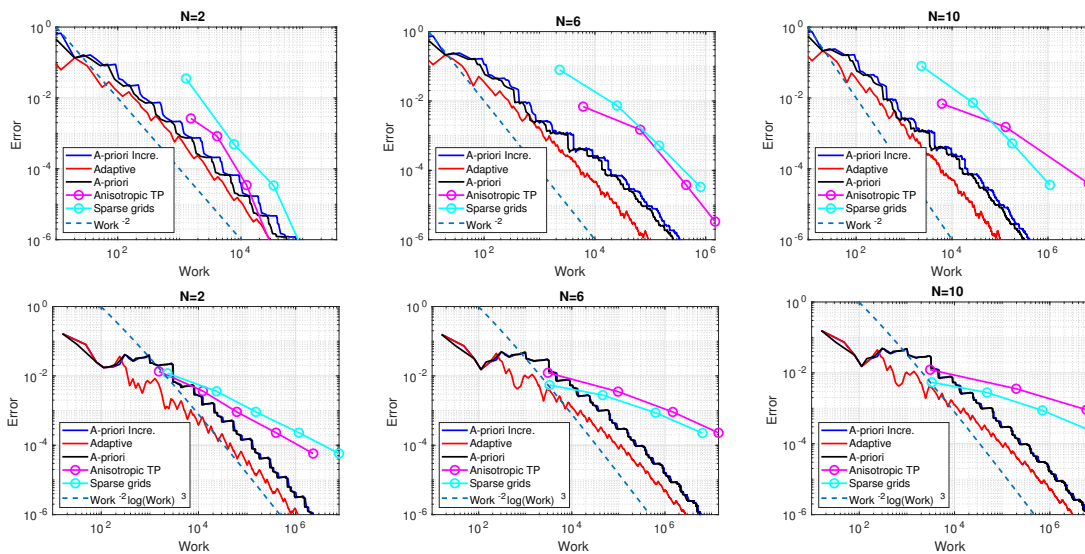


Figure 3. The left and center panels show the sparsity pattern of  $(\beta_1, \beta_2, \beta_3)$  and  $(\beta_1, \beta_9, \beta_{10})$  of Algorithm 4.2 for  $d = 2$  and  $N = 10$ . The right panel verifies numerically the simplified convergence estimate (4.12).



**Figure 4.** Convergence of the combined  $(\alpha, \beta)$  CT for a one-dimensional physical problem (top row) and a two-dimensional one (bottom row).

$\beta$  are very small, which implies that the corresponding OCP will involve very few quadrature nodes. In contrast, the addition of a new random variable, let's say the  $N + 1$ th, implies that the number of collocation points  $M_\beta$  for the anisotropic tensor product approximation is multiplied by  $\beta_{N+1} = 1 + \lfloor \frac{L}{g_{N+1}} \rfloor$ , and thus the single OCP that has to be solved becomes steadily larger. The classical sparse grid approach shows a computational complexity similar to that of the CT in all cases. However, due to the presence of negative weights, the full-space optimality system had to be solved with GMRES. The additional cost and memory usage due to the larger Krylov subspace have not been taken into account in the figure. Finally, the right panel of Figure 2 verifies instead the simplified complexity estimate (4.12), as the error decreases linearly in the log-log plot.

Next, we study the complexity of the CT applied to both spatial and stochastic variables. Figure 4 shows the convergence for the model problem set in a one-dimensional domain (top row) and a two-dimensional domain (bottom row). For Meth. 4 and Meth. 5, we consider a sequence of spatial meshes obtained by halving simultaneously the mesh size in each spatial direction and an increasing sequence of levels  $L$  for the stochastic quadrature. The work reported in the  $x$ -axis corresponds to  $W = \sum_{(\alpha, \beta) \in \mathcal{I}} (\prod_{n=1}^D 2^{\alpha_n+1}) M_\beta$  for Meth. 1 and Meth. 2,  $W = \sum_{\beta \in \mathcal{I}: c_\beta \neq 0} (\prod_{n=1}^D 2^{\alpha_n+1}) M_\beta$  for Meth. 3,  $W = (\prod_{n=1}^D 2^{\alpha_n+1}) M_\beta$  for Meth. 4, and  $W = (\prod_{n=1}^D 2^{\alpha_n+1}) |\Lambda_\beta|$  for Meth. 5. We calculate a reference solution by running Meth. 2 with a very small tolerance, and we linearly interpolate all the solutions computed (which live on different meshes) on a very fine mesh obtained by taking, for each physical dimension, the smallest mesh size used by Meth. 2. In two dimensions, the reference solution has more than  $2 \cdot 10^8$  degrees of freedom. To compute the error, we interpolate the current approximations on the reference mesh and calculate there the  $L^2$  norm of the difference with the reference solution using a mass lumped matrix. Notice that the interpolation step does not introduce

errors, as the finite element basis functions are linear in one dimension and bilinear in two dimensions, and the meshes are nested.

As  $\tilde{\gamma}_j = 1$  and  $\tilde{r}_j = 2$  for  $j = 1, 2$ , Theorem 4.2 predicts an *asymptotic* complexity of  $\text{Work}^{-2}$  for the one-dimensional setting and  $\text{Work}^{-2} \log(\text{Work})^3$  for the two-dimensional setting. For few random variables (e.g.,  $N = 2$ ), the asymptotic complexity is rapidly attained in both settings, while for larger numbers of random variables pre-asymptotic effects are non-negligible, and the asymptotic regime is achieved for larger values of  $W$ . Meth. 5 shows a worse complexity behavior than that of the CT, especially for  $d = 2$ , where we clearly see the benefits of the mixed spatial regularity. Finally, we emphasize that both the anisotropic tensor product and the sparse grid approximations are not able to achieve smaller errors on our workstation, since the size of their optimality system quickly saturates the memory storage. By solving several OCPs characterized by smaller optimality systems, the CT is able to overcome the limitations set by the workstation and achieve much smaller tolerances.

**6. Conclusions.** In this work, we proposed two variants of the combination technique to solve efficiently optimal control problems under uncertainty. The first one is based on the combination of solutions of several OCPs discretized using different tensor product quadrature formulae but on the same spatial mesh. The approach allows us to reduce the computational effort for high-dimensional problems, while avoiding the inconveniences of a direct sparse grid approximation of the objective functional. Then we applied the combination technique to both the spatial and the stochastic variables. Such a procedure allows us to automatically balance the error due to the spatial and stochastic discretization according to a profit rule, and it permits us to further reduce the computational cost by moving most computations on coarser grids. The generalization of this work to an infinite sequence of random variables, including a theoretical analysis for the decay of the error contributions, is currently ongoing and will be the subject of a forthcoming manuscript. Further efforts will be devoted to test the combination technique for OCPs involving control constraints and more general risk measures. The problem formulation (2.2) considered here is already sufficiently general to cover some common risk measures such as the mean-variance model. Of particular interest in risk-averse contexts is the CVaR (Conditional Value at Risk). Although the minimization of the CVaR can be formulated as a double minimization problem leading to an optimality condition of the form of (2.3), it also involves a discontinuous function of the random vector  $\zeta$  acting on the right-hand side of the adjoint equation, which might reduce the stochastic regularity. Thus, a CT based on a multilevel (quasi)-Monte Carlo quadrature seems promising, as it requires less smoothness than the stochastic collocation method.

**7. Appendix.** In this appendix, we detail the proofs of Theorems 4.1 and 4.2. Both proofs are based on a direct counting argument, and in particular the proof of Theorem 4.2 is an adaptation of the proof of Theorem 1 in [19] to a different multi-index set.

For a  $\mathbf{x} = (x_1, \dots, x_N) \in \mathbb{R}^N$ , let  $|\mathbf{x}| := \sum_{n=1}^N |x_n|$  and  $\log(\mathbf{x}) = (\log(x_1), \dots, \log(x_N))$ . The following technical lemmas are needed.

**Lemma 7.1 (Lemma 4 in [19]).** *Let  $f : (1, \infty)^N \rightarrow \mathbb{R}$  and  $g : (1, \infty)^N \rightarrow \mathbb{R}_+$ . If  $f$  and  $g$  are increasing, then*

$$\sum_{\alpha \in \mathbb{N}_+^N: f(\alpha) \leq 0} g(\alpha) \leq \int_{\mathbf{x} \in (1, \infty)^N: f(\mathbf{x}-1) \leq 0} g(\mathbf{x}) d\mathbf{x}.$$

If  $f$  and  $g$  are decreasing, then

$$\sum_{\alpha \in \mathbb{N}_+^N: f(\alpha) \leq 0} g(\alpha) \leq \int_{\mathbf{x} \in (1, \infty)^N: f(\mathbf{x}) \leq 0} g(\mathbf{x}-1) d\mathbf{x}.$$

**Lemma 7.2.** *The following bound holds true:*

$$\int_{\mathbf{x} \in (0, \infty)^N: |\mathbf{x}| \leq H} \mathbf{x} d\mathbf{x} = \frac{H^{2N}}{(2N)!}.$$

*Proof.* Perform the change of variables  $\mathbf{t} = \frac{\mathbf{x}}{H}$ , and use an induction argument over  $N$ . ■

**Lemma 7.3.** *The following bound holds true:*

$$\int_{\mathbf{x} \in (0, \infty)^N: |\mathbf{x}| \geq L} e^{-|\mathbf{x}| + |\log(\mathbf{x})|} d\mathbf{x} \leq e^{-L} (L+1)^{2N-1}.$$

*Proof.* The proof based on induction is part of the proof of Lemma 5 in [19]. ■

**Lemma 7.4 (Lemma 7 in [19]).** *Let  $k \in \mathbb{N}$ ,  $\mathbf{a} = (a_1, \dots, a_D) \in \mathbb{R}_+^D$ , and  $L > |\mathbf{a}|$ . Then*

$$\int_{\{\mathbf{x} \in \mathbb{R}_+^D: |\mathbf{x}| \leq L\}} e^{\mathbf{a} \cdot \mathbf{x}} (L - |\mathbf{x}|)^k d\mathbf{x} \leq \mathcal{U}_D(\mathbf{a}, k) e^{\max(\mathbf{a})L} L^{\eta(\mathbf{a}, \max(\mathbf{a}))-1},$$

where  $\eta(\mathbf{a}, \max(\mathbf{a})) := \#\{i: a_i = \max(\mathbf{a})\}$  and  $\mathcal{U}_D(\mathbf{a}, k)$  is a constant independent of  $L$ .

**Lemma 7.5 (Lemma B.3 in [20]).** *Let  $k \in \mathbb{N}$ ,  $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}_+^D$ , and  $L > |\mathbf{a}|$ . Then*

$$\int_{\{\mathbf{x} \in \mathbb{R}_+^D: |\mathbf{x}| > L\}} e^{-\mathbf{a} \cdot \mathbf{x}} \leq \mathcal{B}_D(\mathbf{a}) e^{-\min(\mathbf{a})L} L^{\eta(\mathbf{a}, \min(\mathbf{a}))-1},$$

where  $\mathcal{B}_D(\mathbf{a})$  is a constant independent on  $L$ .

### 7.1. Proof of Theorem 4.1.

**Theorem 7.6.** *There exist constants  $C_1$  and  $C_2$  such that for any  $W_{\max}$  satisfying  $W_{\max} \geq \frac{|\tilde{\mathbf{g}}|^{2N} C_1}{(2N)!}$ , and choosing  $\hat{L} = \sqrt[2N]{\frac{W_{\max}(2N)!}{C_1}} - |\tilde{\mathbf{g}}|$ ,*

$$\begin{aligned} \text{Work}[\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] &\leq W_{\max}, \\ \text{Error}[\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] &\leq C_2 e^{-2N \sqrt[2N]{\frac{W_{\max}(2N)!}{C_1}}} \left( \sqrt[2N]{\frac{W_{\max}(2N)!}{C_1}} + 1 \right)^{2N-1}. \end{aligned}$$

*Proof.* We start with the work estimate. The total work satisfies

$$\begin{aligned} \text{Work } [\mathcal{M}_{\mathcal{I}(L)}(u)] &= \sum_{\beta \in \mathcal{I}(L)} \Delta W_{\tilde{\alpha}, \beta}^{\text{stoc}} \leq C_{\text{work}}^{\text{stoc}, \tilde{\alpha}} \sum_{\{\beta \in \mathbb{N}_+^N: \tilde{\mathbf{g}} \cdot (\beta+1) + |\log(\beta+1)| \leq L\}} e^{|\log(\beta+1)|} \\ &\leq C_{\text{work}}^{\text{stoc}, \tilde{\alpha}} \int_{\{\beta \in \otimes_{i=1}^N (1, \infty): \tilde{\mathbf{g}} \cdot \beta + |\log(\beta)| \leq L\}} e^{|\log(\beta+1)|} d\beta, \end{aligned}$$

where the last inequality follows from Lemma 7.1.

Next, we perform the change of variables  $t_j = \tilde{g}_j \beta_j + \log(\beta_j)$ , setting  $\beta_j = q_j(t)$ , with  $q_j(t) \leq \frac{t_j}{\tilde{g}_j}$ , where

$$\begin{aligned} \text{Work } [\mathcal{M}_{\mathcal{I}(L)}(u)] &\leq C_{\text{work}}^{\text{stoc}, \tilde{\alpha}} \int_{\{\mathbf{t} \in \otimes_{i=1}^N (\tilde{g}_i, \infty)^N: |\mathbf{t}| \leq L\}} \prod_{i=1}^N \frac{q_i(t) + 1}{\tilde{g}_i q_i(t) + 1} q_i(t) dt \\ &\leq C_{\text{work}}^{\text{stoc}, \tilde{\alpha}} \int_{\{\mathbf{t} \in \otimes_{i=1}^N (\tilde{g}_i, \infty)^N: |\mathbf{t}| \leq L\}} \prod_{i=1}^N \frac{t_j/\tilde{g}_j + 1}{t_j/\tilde{g}_j + 1} \frac{t_j}{\tilde{g}_j} dt \\ &\leq C_1 \int_{\{\mathbf{t} \in \otimes_{i=1}^N (\tilde{g}_i, \infty)^N: |\mathbf{t}| \leq L\}} \prod_{i=1}^N (t_j + \tilde{g}_j) dt \\ &\leq C_1 \int_{\{\tilde{\mathbf{t}} \in \otimes_{i=1}^N (0, \infty)^N: |\tilde{\mathbf{t}}| \leq L + |\tilde{\mathbf{g}}|\}} \tilde{\mathbf{t}} d\tilde{\mathbf{t}} \\ &\leq C_1 \frac{(L + |\tilde{\mathbf{g}}|)^{2N}}{(2N)!}, \end{aligned}$$

where  $C_1 = C_{\text{work}}^{\text{stoc}, \tilde{\alpha}} \prod_{i=1}^N (\frac{1}{\tilde{g}_i})$  and in the last step we used Lemma 7.2.

We next consider the error term,

$$\begin{aligned} \text{Error } [\mathcal{M}_{\mathcal{I}(L)}(u)] &= \sum_{\beta \notin \mathcal{I}(L)} \Delta E_{\tilde{\alpha}, \beta}^{\text{stoc}} \\ &\leq C_{\text{error}}^{\text{stoc}, \tilde{\alpha}} \sum_{\{\beta \in \mathbb{N}_+^N: \tilde{\mathbf{g}} \cdot (\beta+1) + \log(\beta+1) > L\}} e^{-\tilde{\mathbf{g}} \cdot (\beta+1)} \\ &\leq C_{\text{error}}^{\text{stoc}, \tilde{\alpha}} \int_{\{\beta \in \otimes_{i=1}^N (1, \infty)^N: \tilde{\mathbf{g}} \cdot (\beta+1) + \log(\beta+1) > L\}} e^{-\tilde{\mathbf{g}} \cdot \beta} d\beta \\ &= C_{\text{error}}^{\text{stoc}, \tilde{\alpha}} e^{|\tilde{\mathbf{g}}|} \int_{\{\beta \in \otimes_{i=1}^N (1, \infty)^N: \tilde{\mathbf{g}} \cdot (\beta+1) + \log(\beta+1) > L\}} e^{-\tilde{\mathbf{g}} \cdot (\beta+1) + |\log(\beta+1)| - |\log(\beta+1)|} d\beta. \end{aligned}$$

The change of variables  $\tilde{g}_i(\beta_i + 1) + \log(\beta_i + 1) = t_j$ ,  $q(t_j) = \beta_i$ , with  $q(t_j) \leq \frac{t_j}{\tilde{g}_j}$ , leads to

$$\begin{aligned} \text{Error } [\mathcal{M}_{\mathcal{I}(L)}(u)] &\leq C_{\text{error}}^{\text{stoc}, \tilde{\alpha}} e^{|\tilde{\mathbf{g}}|} \int_{\{\mathbf{t} \in \otimes_{i=1}^N (2\tilde{g}_i + \log(2), \infty)^N: |\mathbf{t}| > L\}} e^{-|\mathbf{t}|} \left( \prod_{i=1}^N \frac{(q(t_i) + 1)^2}{\tilde{g}_i(q(t_i) + 1) + 1} \right) dt \\ &\leq C_{\text{error}}^{\text{stoc}, \tilde{\alpha}} e^{|\tilde{\mathbf{g}}|} \int_{\{\mathbf{t} \in \otimes_{i=1}^N (2\tilde{g}_i + \log(2), \infty)^N: |\mathbf{t}| > L\}} e^{-|\mathbf{t}|} \left( \prod_{i=1}^N \frac{t_j + \tilde{g}_j}{\tilde{g}_j^2} \right) dt \end{aligned}$$

$$\begin{aligned}
&\leq C_{\text{error}}^{\text{stoc}, \bar{\alpha}} \left( \prod_{i=1}^N \frac{e^{\tilde{g}_i}}{\tilde{g}_i^2} \right) \int_{\{\mathbf{t} \in \otimes_{i=1}^N (2\tilde{g}_i + \log(2), \infty)^N : |\mathbf{t}| > L\}} e^{-|\mathbf{t}| + |\log(\mathbf{t} + \tilde{\mathbf{g}})|} d\mathbf{t} \\
&\leq C_2 \int_{\{\mathbf{x} \in \otimes_{i=1}^N (3\tilde{g}_i + \log(2), \infty)^N : |\mathbf{x}| > L + |\tilde{\mathbf{g}}|\}} e^{-|\mathbf{x}| + |\log(\mathbf{x})|} d\mathbf{x} \\
&\leq C_2 \int_{\{\mathbf{x} \in \otimes_{i=1}^N (0, \infty)^N : |\mathbf{t}| > L + |\tilde{\mathbf{g}}|\}} e^{-|\mathbf{x}| + |\log(\mathbf{x})|} d\mathbf{x} \\
&\leq C_2 e^{-L - |\tilde{\mathbf{g}}|} (L + 1 + |\tilde{\mathbf{g}}|)^{2N-1},
\end{aligned}$$

where we used Lemma 7.3 and  $C_2 = C_{\text{error}}^{\text{stoc}, \bar{\alpha}} \left( \prod_{i=1}^N \frac{e^{2\tilde{g}_i}}{\tilde{g}_i^2} \right)$ .

Inserting the expression of  $\hat{L}$  into the bounds for the error and work contributions, it is immediate to verify that  $\text{Work} [\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] \leq W_{\max}$  and

$$\text{Error} [\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] \leq C_2 e^{-2N \sqrt{\frac{W_{\max}(2N)!}{C_1}}} \left( 2N \sqrt{\frac{W_{\max}(2N)!}{C_1}} + 1 \right)^{2N-1}. \quad \blacksquare$$

## 7.2. Proof of Theorem 4.2.

**Theorem 7.7.** *Let  $r_j = \log(2)\tilde{r}_j$ ,  $\gamma_j = \log(2)\tilde{\gamma}_j$ ,  $j = 1, \dots, N$ ,  $\Theta = (\frac{\gamma_1}{r_1 + \gamma_1}, \dots, \frac{\gamma_D}{r_D + \gamma_D})$ ,  $\mu = \min_n \frac{r_n}{\gamma_n}$ ,  $\chi = \max_n \Theta_n$ , and  $n(\Theta, \chi) := \#\{n : \Theta_n = \chi\}$ . There exists a constant  $C_W$  such that for any  $W_{\max}$  satisfying  $W_{\max} \geq C_W e^\chi$ , and setting*

$$(7.1) \quad L = L(W_{\max}) = \frac{1}{\chi} \left( \log \left( \frac{W_{\max}}{C_W} \right) - (n(\Theta, \chi) - 1) \log \left( \frac{1}{\chi} \log \left( \frac{W_{\max}}{C_W} \right) \right) \right),$$

the combination technique solution satisfies

$$(7.2) \quad \text{Work} [\mathcal{M}_{\mathcal{I}(L(W_{\max}))}] \leq W_{\max},$$

$$(7.3) \quad \limsup_{W_{\max} \rightarrow \infty} \frac{\text{Error} [\mathcal{M}_{\mathcal{I}(L(W_{\max}))}]}{W_{\max}^{-\mu} (\log(W_{\max}))^{(\mu+1)(n(\Theta, \chi)-1)}} = C < \infty.$$

*Proof.* Due to Assumption 4, the total work associated to  $\mathcal{I}(L)$  is bounded by

$$\begin{aligned}
\text{Work} [\mathcal{M}_{\mathcal{I}(L)}(u)] &= \sum_{(\alpha, \beta) \in \mathcal{I}(L)} \Delta W_{\alpha, \beta} \\
&\leq C_{\text{work}} \sum_{\{(\alpha, \beta) \in \mathbb{N}_+^{D+N} : (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha} + \tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})| \leq L\}} e^{\boldsymbol{\gamma} \cdot \boldsymbol{\alpha} + \log |\boldsymbol{\beta} + \mathbf{1}|} \\
&\leq C_{\text{work}} \int_{\{(\alpha, \beta) \in (1, \infty)^{D+N} : (\mathbf{r} + \boldsymbol{\gamma}) \cdot (\boldsymbol{\alpha} - \mathbf{1}) + \tilde{\mathbf{g}} \cdot \boldsymbol{\beta} + |\log(\boldsymbol{\beta})| \leq L\}} e^{\boldsymbol{\gamma} \cdot \boldsymbol{\alpha} + \log |\boldsymbol{\beta} + \mathbf{1}|} d\boldsymbol{\alpha} d\boldsymbol{\beta} \\
&= C_{\text{work}} \left( \prod_{i=1}^D \frac{e^{\gamma_i}}{r_i + \gamma_i} \right) \int_{\{(\bar{\alpha}, \beta) \in (0, \infty)^D \times (1, \infty)^N : |\bar{\alpha}| + \tilde{\mathbf{g}} \cdot \boldsymbol{\beta} + |\log(\boldsymbol{\beta})| \leq L\}} e^{\Theta \cdot \bar{\alpha} + \log |\boldsymbol{\beta} + \mathbf{1}|} d\bar{\alpha} d\boldsymbol{\beta},
\end{aligned}$$

where we used Lemma 7.1, performed the change of variables  $\bar{\alpha}_i = (\alpha_i - 1)(r_i + \gamma_i)$ , and defined  $\Theta_i = \frac{\gamma_i}{r_i + \gamma_i}$ . The change of variables  $\bar{\beta}_i = \beta_i \tilde{g}_i$  leads to

$$(7.4) \quad \text{Work } [\mathcal{M}_{\mathcal{I}(L)}(u)] \leq D_{\text{work}} \int_{\{(\bar{\alpha}, \bar{\beta}) \in (0, \infty)^D \times \otimes_{i=1}^N (\tilde{g}_i, \infty) : |\bar{\alpha}| + |\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})| \leq L\}} e^{\Theta \cdot \bar{\alpha} + |\log(\frac{\bar{\beta}}{\theta} + 1)|} d\bar{\alpha} d\bar{\beta},$$

where  $D_{\text{work}} := C_{\text{work}} (\prod_{i=1}^D \frac{e^{\gamma_i}}{r_i + \gamma_i}) (\prod_{i=1}^N \frac{1}{\tilde{g}_i})$ . Denoting with Int the right-hand side of (7.4) and using Lemma 7.4, we get

$$(7.5) \quad \begin{aligned} \text{Int} &= \int_{\{\bar{\beta} \in \otimes_{i=1}^N (\tilde{g}_i, \infty) : |\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})| \leq L\}} e^{|\log(\frac{\bar{\beta}}{\theta} + 1)|} \int_{\{\bar{\alpha} \in (0, \infty)^D : |\bar{\alpha}| \leq L - |\bar{\beta}| - |\log(\frac{\bar{\beta}}{\theta})|\}} e^{\Theta \cdot \bar{\alpha}} d\bar{\alpha} d\bar{\beta} \\ &\leq \int_{\{\bar{\beta} \in \otimes_{i=1}^N (\tilde{g}_i, \infty) : |\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})| \leq L\}} \mathcal{U}_D(\Theta, 0) e^{|\log(\frac{\bar{\beta}}{\theta} + 1)| + \chi(L - |\bar{\beta}| - |\log(\frac{\bar{\beta}}{\theta})|)} L^{\eta(\Theta, \chi) - 1} d\bar{\beta} \\ &\leq \mathcal{U}_D(\Theta, 0) e^{\chi L} \int_{\{\bar{\beta} \in \otimes_{i=1}^N (\tilde{g}_i, \infty) : |\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})| \leq L\}} e^{|\log(\frac{\bar{\beta}}{\theta} + 1)|} e^{-\chi(|\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})|)} L^{\eta(\Theta, \chi) - 1} d\bar{\beta} \\ &\leq \mathcal{U}_D(\Theta, 0) e^{\chi L} L^{\eta(\Theta, \chi) - 1} \int_{\{\bar{\beta} \in \otimes_{i=1}^N (\tilde{g}_i, \infty) : |\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})| \leq L\}} e^{|\log(\frac{\bar{\beta}}{\theta} + 1)|} e^{-\chi(|\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})|)} d\bar{\beta}. \end{aligned}$$

Plugging (7.5) into (7.4),

$$\text{Work } [\mathcal{M}_{\mathcal{I}(L)}(u)] \leq C_W L^{\eta(\Theta, \chi) - 1} e^{\chi L},$$

where  $C_W = D_{\text{work}} \mathcal{U}_D(\Theta, 0) \int_{\{\bar{\beta} \in \otimes_{i=1}^N (\tilde{g}_i, \infty) : |\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})| \leq L\}} e^{|\log(\frac{\bar{\beta}}{\theta} + 1)|} e^{-\chi(|\bar{\beta}| + |\log(\frac{\bar{\beta}}{\theta})|)} d\bar{\beta} < \infty$  as the latter integral is bounded for every  $L$ . We now focus on the error estimate,

$$(7.6) \quad \begin{aligned} \text{Error } [\mathcal{M}_{\mathcal{I}(\hat{L})}(u)] &\leq C_{\text{error}} \sum_{\{(\alpha, \beta) \in \mathbb{N}_+^{D+N} : (r + \gamma) \cdot \alpha + \tilde{g} \cdot (\beta + 1) + |\log(\beta + 1)| > L\}} e^{-r \cdot \alpha - \tilde{g} \cdot (\beta + 1)} \\ &= C_{\text{error}} \left( \sum_{\{(\alpha, \beta) \in \mathbb{N}_+^{D+N} : (r + \gamma) \cdot \alpha > L\}} e^{-r \cdot \alpha - \tilde{g} \cdot (\beta + 1)} \right. \\ &\quad \left. + \sum_{\{\alpha \in \mathbb{N}_+^D : (r + \gamma) \cdot \alpha \leq L\}} e^{-r \cdot \alpha} \sum_{\{\beta \in \mathbb{N}_+^N : \tilde{g} \cdot (\beta + 1) + |\log(\beta + 1)| > L - (r + \gamma) \cdot \alpha\}} e^{-\tilde{g} \cdot (\beta + 1)} \right). \end{aligned}$$

Using the change of variables  $\bar{\alpha}_i = (r_i + \gamma_i)\alpha_i$ , the first term can be bounded as

$$(7.7) \quad \begin{aligned} \sum_{\{(\alpha, \beta) \in \mathbb{N}_+^{D+N} : (r + \gamma) \cdot \alpha > L\}} e^{-r \cdot \alpha - \tilde{g} \cdot (\beta + 1)} &= \left( \sum_{\beta \in \mathbb{N}_+^N} \prod_{i=1}^D e^{-\tilde{g}_i(\beta_i + 1)} \right) \left( \sum_{\alpha \in \mathbb{N}_+^D : (r + \gamma) \cdot \alpha > L} e^{-r \cdot \alpha} \right) \\ &= \left( \prod_{i=1}^D \frac{e^{-2\tilde{g}_i}}{1 - e^{-\tilde{g}_i}} \right) \left( \sum_{\alpha \in \mathbb{N}_+^D : (r + \gamma) \cdot \alpha > L} e^{-r \cdot \alpha} \right) \end{aligned}$$

$$\begin{aligned}
&\leq \left( \prod_{i=1}^N \frac{e^{-2\tilde{g}_i}}{1 - e^{-\tilde{g}_i}} \right) \int_{\boldsymbol{\alpha} \in (1, \infty)^D: (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha} > L} e^{-\mathbf{r} \cdot \boldsymbol{\alpha} + r} d\boldsymbol{\alpha} \\
&\leq \left( \prod_{i=1}^N \frac{e^{-2\tilde{g}_i}}{1 - e^{-\tilde{g}_i}} \right) \left( \prod_{i=1}^D \frac{e^{r_i}}{r_i + \gamma_i} \right) \int_{\tilde{\boldsymbol{\alpha}} \in \otimes_{i=1}^D (r_i + \gamma_i, \infty): |\tilde{\boldsymbol{\alpha}}| > L} e^{-\Phi \cdot \tilde{\boldsymbol{\alpha}}} d\tilde{\boldsymbol{\alpha}} \\
&\leq C_{E,1} e^{-\delta L} L^{n(\Phi, \delta) - 1},
\end{aligned}$$

where we used Lemma 7.5 with

$$\Phi_i = \frac{r_i}{r_i + \gamma_i}, \quad \delta = \min_{i=1, \dots, N} \Phi_i, \quad \text{and} \quad C_{E,1} := \left( \prod_{i=1}^N \frac{e^{-2\tilde{g}_i}}{1 - e^{-\tilde{g}_i}} \right) \left( \prod_{i=1}^D \frac{e^{r_i}}{r_i + \gamma_i} \right) \mathcal{B}_D(\Phi).$$

Concerning the second term, we first define  $H := (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha}$  and  $\hat{L} := L - H$ , so that

$$\begin{aligned}
&\sum_{\{\boldsymbol{\beta} \in \mathbb{N}_+^N: \tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})| > \hat{L}\}} e^{-\tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1})} \leq \int_{\{\boldsymbol{\beta} \in \otimes_{i=1}^N (1, \infty): \tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})| > \hat{L}\}} e^{-\tilde{\mathbf{g}} \cdot \boldsymbol{\beta}} d\boldsymbol{\beta} \\
&= e^{|\tilde{\mathbf{g}}|} \int_{\{\boldsymbol{\beta} \in \otimes_{i=1}^N (1, \infty): \tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})| > \hat{L}\}} e^{-\tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})| - |\log(\boldsymbol{\beta} + \mathbf{1})|} d\boldsymbol{\beta}.
\end{aligned}$$

Setting  $t_i = \tilde{g}_i(\beta_i + 1) + \log(\beta_i + 1) = t_i$ ,  $q(t_i) = \beta_i$ , with  $q(t_i) \leq \frac{t_i}{\tilde{g}_i} - 1$ , leads to

$$\begin{aligned}
(7.8) \quad &\sum_{\{\boldsymbol{\beta} \in \mathbb{N}_+^N: \tilde{\mathbf{g}} \cdot \boldsymbol{\beta} + |\log(\boldsymbol{\beta} + \mathbf{1})| > \hat{L}\}} e^{-\tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1})} \\
&\leq \left( \prod_{i=1}^N e^{\tilde{g}_i} \right) \int_{\{\mathbf{t} \in \otimes_{i=1}^D (2\tilde{g}_i + \log(2), \infty): |\mathbf{t}| > \hat{L}\}} e^{-|\mathbf{t}|} \left( \prod_{i=1}^N \frac{(q(t_i) + 1)^2}{\tilde{g}_i(q(t_i) + 1)} \right) d\mathbf{t} \\
&\leq \left( \prod_{i=1}^N \frac{e^{\tilde{g}_i}}{\tilde{g}_i^2} \right) \int_{\{\mathbf{t} \in \otimes_{i=1}^D (2\tilde{g}_i + \log(2), \infty): |\mathbf{t}| > \hat{L}\}} e^{-|\mathbf{t}| + |\log(\mathbf{t})|} d\mathbf{t} \\
&\leq C_{E,2} e^{-\hat{L}} (\hat{L} + 1)^{2N-1},
\end{aligned}$$

where we used Lemma 7.3 in the last step and set  $C_{E,2} = \left( \prod_{i=1}^N \frac{e^{\tilde{g}_i}}{\tilde{g}_i^2} \right)$ . Inserting (7.8) into the second term of (7.6) and using Lemma 7.4,

$$\begin{aligned}
&\sum_{\{\boldsymbol{\alpha} \in \mathbb{N}_+^D: (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha} \leq L\}} e^{-\mathbf{r} \cdot \boldsymbol{\alpha}} \sum_{\{\boldsymbol{\beta} \in \mathbb{N}_+^N: \tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1}) + |\log(\boldsymbol{\beta} + \mathbf{1})| > L - (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha}\}} e^{-\tilde{\mathbf{g}} \cdot (\boldsymbol{\beta} + \mathbf{1})} \\
&\leq C_{E,2} \sum_{\{\boldsymbol{\alpha} \in \mathbb{N}_+^D: (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha} \leq L\}} e^{-L + \boldsymbol{\gamma} \cdot \boldsymbol{\alpha}} (L + 1 - (\mathbf{r} + \boldsymbol{\gamma}) \cdot \boldsymbol{\alpha})^{2N-1} \\
&\leq C_{E,2} \int_{\{\boldsymbol{\alpha} \in (1, \infty)^D: (\mathbf{r} + \boldsymbol{\gamma}) \cdot (\boldsymbol{\alpha} - \mathbf{1}) \leq L\}} e^{-L + \boldsymbol{\gamma} \cdot \boldsymbol{\alpha}} (L + 1 - (\mathbf{r} + \boldsymbol{\gamma}) \cdot (\boldsymbol{\alpha} - \mathbf{1}))^{2N-1} \\
&\leq C_{E,2} \left( \prod_{i=1}^D \frac{e^{\gamma_i}}{\gamma_i + r_i} \right) e^{-L} \int_{\{\mathbf{x} \in (0, \infty)^D: |\mathbf{x}| \leq L\}} e^{\Theta \cdot \mathbf{x}} (L + 1 - |\mathbf{x}|)^{2N-1} d\mathbf{x} \\
&\leq C_{E,3} e^{(\chi-1)L} L^{n(\Theta, \chi) - 1},
\end{aligned}$$

where  $C_{E,3} = C_{E,2}(\prod_{i=1}^D \frac{e^{\gamma_i}}{\gamma_i+r_i})\mathcal{U}_D(\Theta, 2N-1)$ . Since  $\delta = 1 - \chi$ ,  $n(\Theta, \chi) = n(\Phi, \delta)$ , and using the bounds for the two terms of (7.6), we conclude that

$$\text{Error} [\mathcal{M}_{\mathcal{I}(L)}(u)] \leq C_{E,1}e^{-\delta L}L^{\eta(\Phi,\delta)-1} + C_{E,3}e^{(\chi-1)L}L^{n(\Theta,\chi)-1} = C_{E,4}e^{(\chi-1)L}L^{n(\Theta,\chi)-1}.$$

Finally, a direct calculation shows that (7.1) leads to (7.2) and (7.3). ■

**Acknowledgment.** The second author is a member of GNCS (Gruppo Nazionale per il Calcolo Scientifico) of INdAM.

## REFERENCES

- [1] I. BABUŠKA, F. NOBILE, AND R. TEMPONE, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM Rev., 52 (2010), pp. 317–355, <https://doi.org/10.1137/100786356>.
- [2] J. BÄCK, F. NOBILE, L. TAMELLINI, AND R. TEMPONE, *Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: A numerical comparison*, in Spectral and High Order Methods for Partial Differential Equations, Springer, 2011, pp. 43–62.
- [3] J. BECK, L. TAMELLINI, AND R. TEMPONE, *IGA-based multi-index stochastic collocation for random PDEs on arbitrary domains*, Comput. Methods Appl. Mech. Engrg., 351 (2019), pp. 330–350.
- [4] F. BREZZI, J. RAPPAZ, AND P.-A. RAVIART, *Finite dimensional approximation of nonlinear problems: Part I: Branches of nonsingular solutions*, Numer. Math., 36 (1980/81), pp. 1–25.
- [5] H.-J. BUNGARTZ AND M. GRIEBEL, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.
- [6] A. CHKIFA, A. COHEN, AND C. SCHWAB, *High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs*, Found. Comput. Math., 14 (2014), pp. 601–633.
- [7] A. COHEN, R. DEVORE, AND C. SCHWAB, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs*, Anal. Appl., 9 (2011), pp. 11–47.
- [8] A. COHEN AND R. DEVORE, *Approximation of high-dimensional parametric PDEs*, Acta Numer., 24 (2015), pp. 1–159.
- [9] A. COHEN, R. DEVORE, AND C. SCHWAB, *Convergence rates of best n-term Galerkin approximations for a class of elliptic PDEs*, Found. Comput. Math., 10 (2010), pp. 615–646.
- [10] A. ERN AND J. L. GUERMOND, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer, New York, 2004.
- [11] O. G. ERNST, B. SPRUNGK, AND L. TAMELLINI, *Convergence of sparse collocation for functions of countably many Gaussian random variables (with application to elliptic PDEs)*, SIAM J. Numer. Anal., 56 (2018), pp. 877–905, <https://doi.org/10.1137/17M1123079>.
- [12] T. GERSTNER AND M. GRIEBEL, *Dimension-adaptive tensor-product quadrature*, Computing, 71 (2003), pp. 65–87.
- [13] M. GRIEBEL AND H. HARBRECHT, *On the convergence of the combination technique*, in Sparse Grids and Applications—Munich 2012, Springer, Cham, 2014, pp. 55–74.
- [14] M. GRIEBEL, M. SCHNEIDER, AND C. ZENGER, *A combination technique for the solution of sparse grid problems*, in Iterative Methods in Linear Algebra, North-Holland, Amsterdam, 1992, pp. 263–281.
- [15] M. GRIEBEL, AND U. SEIDLER, *A dimension-adaptive combination technique for uncertainty quantification*, Int. J. Uncertain. Quantif., 14 (2024), pp. 21–43.
- [16] D. GUIGNARD AND F. NOBILE, *A posteriori error estimation for the stochastic collocation finite element method*, SIAM J. Numer. Anal., 56 (2018), pp. 3121–3143, <https://doi.org/10.1137/17M1155454>.
- [17] P. A. GUTH, V. KAARNIOJA, F. Y. KUO, C. SCHILLINGS, AND I. H. SLOAN, *A quasi-Monte Carlo method for optimal control under uncertainty*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 354–383, <https://doi.org/10.1137/19M1294952>.
- [18] A.-L. HAJI-ALI, F. NOBILE, L. TAMELLINI, AND R. TEMPONE, *Multi-index stochastic collocation convergence rates for random PDEs with parametric regularity*, Found. Comput. Math., 16 (2016), pp. 1555–1605.

- [19] A.-L. HAJI-ALI, F. NOBILE, L. TAMELLINI, AND R. TEMPONE, *Multi-index stochastic collocation for random PDEs*, *Comput. Methods Appl. Mech. Engrg.*, 306 (2016), pp. 95–122.
- [20] A.-L. HAJI-ALI, F. NOBILE, AND R. TEMPONE, *Multi-index Monte Carlo: When sparsity meets sampling*, *Numer. Math.*, 132 (2016), pp. 767–806.
- [21] H. HARBRECHT, M. PETERS, AND M. SIEBENMORGEN, *On multilevel quadrature for elliptic stochastic partial differential equations*, in *Sparse Grids and Applications*, Springer, 2012, pp. 161–179.
- [22] M. HINZE, R. PINNAU, M. ULBRICH, AND S. ULBRICH, *Optimization with PDE Constraints*, *Math. Model. Theory Appl.* 23, Springer, New York, 2009.
- [23] D. P. KOURI, *An Approach for the Adaptive Solution of Optimization Problems Governed by Partial Differential Equations with Uncertain Coefficients*, Ph.D. thesis, Rice University, Houston TX, 2012.
- [24] D. P. KOURI, M. HEINKENSCHLOSS, D. RIDZAL, AND B. G. VAN BLOEMEN WAANDERS, *A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty*, *SIAM J. Sci. Comput.*, 35 (2013), pp. A1847–A1879, <https://doi.org/10.1137/120892362>.
- [25] D. P. KOURI AND D. RIDZAL, *Inexact trust-region methods for PDE-constrained optimization*, in *Frontiers in PDE-Constrained Optimization*, Springer, 2018, pp. 83–121.
- [26] D. P. KOURI AND A. SHAPIRO, *Optimization of PDEs with uncertain inputs*, in *Frontiers in PDE-Constrained Optimization*, Springer, 2018, pp. 41–81.
- [27] D. P. KOURI AND T. M. SUROWIEC, *Existence and optimality conditions for risk-averse PDE-constrained optimization*, *SIAM/ASA J. Uncertain. Quantif.*, 6 (2018), pp. 787–815, <https://doi.org/10.1137/16M1086613>.
- [28] D. P. KROESE, T. TAIMRE, AND Z. I. BOTEV, *Handbook of Monte Carlo Methods*, Wiley Ser. Probab. Statist., Wiley, 2013.
- [29] A. KUNOTH AND C. SCHWAB, *Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs*, *SIAM J. Control Optim.*, 51 (2013), pp. 2442–2471, <https://doi.org/10.1137/110847597>.
- [30] A. KUNOTH AND C. SCHWAB, *Sparse adaptive tensor Galerkin approximations of stochastic PDE-constrained control problems*, *SIAM/ASA J. Uncertain. Quantif.*, 4 (2016), pp. 1034–1059, <https://doi.org/10.1137/15M1041390>.
- [31] G. J. LORD, C. E. POWELL, AND T. SHARDLOW, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts Appl. Math., Cambridge University Press, 2014.
- [32] S. MARTELLO AND P. TOTH, *Knapsack Problems: Algorithms and Computer Implementations*, John Wiley & Sons, 1990.
- [33] M. MARTIN, S. KRUMSCHEID, AND F. NOBILE, *Complexity analysis of stochastic gradient methods for PDE-constrained optimal control problems with uncertain parameters*, *ESAIM Math. Model. Numer. Anal.*, 55 (2021), pp. 1599–1633.
- [34] J. MARTÍNEZ-FRUTOS AND F. P. ESPARZA, *Optimal Control of PDEs under Uncertainty: An Introduction with Application to Optimal Shape Design of Structures*, Springer, 2018.
- [35] F. NOBILE, L. TAMELLINI, AND R. TEMPONE, *Convergence of quasi-optimal sparse-grid approximation of Hilbert-space-valued functions: Application to random elliptic PDEs*, *Numer. Math.*, 134 (2016), pp. 343–388.
- [36] F. NOBILE AND T. VANZAN, *Preconditioners for robust optimal control problems under uncertainty*, *Numer. Linear Algebra Appl.*, 30 (2023), e2472, <https://doi.org/10.1002/nla.2472>.
- [37] B. PEHERSTORFER, K. WILLCOX, AND M. GUNZBURGER, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, *SIAM Rev.*, 60 (2018), pp. 550–591, <https://doi.org/10.1137/16M1082469>.
- [38] C. PFLAUM, *Diskretisierung elliptischer Differentialgleichungen mit dünnen Gittern*, Ph.D. thesis, Fakultät für Mathematik der Technischen Universität München, 1996.
- [39] L. REICHEL, *Newton interpolation at Leja points*, *BIT*, 30 (1990), pp. 332–346.
- [40] P. ROBBE, D. NUYENS, AND S. VANDEWALLE, *A multi-index quasi-Monte Carlo algorithm for lognormal diffusion problems*, *SIAM J. Sci. Comput.*, 39 (2017), pp. S851–S872, <https://doi.org/10.1137/16M1082561>.

- [41] C. SCHWAB AND C. J. GITTELSON, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, Acta Numer., 20 (2011), pp. 291–467.
- [42] A. SHAPIRO, D. DENTCHEVA, AND A. RUSZCZYŃSKI, *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*, SIAM, Philadelphia, 2021, <https://doi.org/10.1137/1.9781611976595>.
- [43] 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 J. Uncertain. Quantif., 3 (2015), pp. 1046–1074, <https://doi.org/10.1137/140969002>.
- [44] L. N. TREFETHEN, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87, <https://doi.org/10.1137/060659831>.
- [45] A. VAN BAREL AND S. VANDEWALLE, *Robust optimization of PDEs with random coefficients using a multilevel Monte Carlo method*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 174–202, <https://doi.org/10.1137/17M1155892>.
- [46] T. VANZAN AND F. NOBILE, *A combination technique for optimal control problems constrained by random PDEs - numerical tests*, Zenodo, 2024, <https://doi.org/10.5281/zenodo.10932946>.