

On the convergence of adaptive stochastic collocation for elliptic partial differential equations with affine diffusion

Martin Eigel¹, Oliver Ernst², Björn Sprungk³, Lorenzo Tamellini⁴

¹Weierstrass Institute, Berlin, Germany

²Department of Mathematics, TU Chemnitz, Germany

³Faculty of Mathematics and Computer Science, TU Bergakademie Freiberg, Germany

⁴Istituto di Matematica Applicata e Tecnologie Informatiche “E. Magenes”, Pavia, Consiglio Nazionale delle Ricerche, Italy

June 17, 2021

Abstract

Convergence of an adaptive collocation method for the parametric stationary diffusion equation with finite-dimensional affine coefficient is shown. The adaptive algorithm relies on a recently introduced residual-based reliable a posteriori error estimator. For the convergence proof, a strategy recently used for a stochastic Galerkin method with a hierarchical error estimator is transferred to the collocation setting. Extensions to other variants of adaptive collocation methods (including the now classical approach proposed in “T. Gerstner and M. Griebel, Dimension-adaptive tensor-product quadrature, Computing, 2003”) are explored.

Keywords. Random PDEs, parametric PDEs, sparse grids, stochastic collocation, adaptive algorithms, high-dimensional approximation, high-dimensional interpolation

AMS subject classifications: 65D05, 65D15, 65C30, 60H25

1 Introduction

Collocation methods are now a mainstay for solving equations containing high-dimensional parameters such as arise in uncertainty quantification (UQ) analyses of ordinary or partial differential equations (ODE/PDE) with uncertain model coefficients [MH03, XH05, BNT07]. It was realized early on that already moderately high-dimensional problems become tractable only when the approximations are based on sparse subspaces of the basic tensor product construction [NTW08b, NTW08a, BS09, MZ09, Bie11, BTNT12].

Subsequent work established that, under mild conditions, certain classes of random PDEs are tractable even in presence of countably many parameter variables [CDS10, CDS11, SS13, CCS14, CCS14, BCM17, ZS20, HS14, BCDVM, Che18, EST18]. These results prove that *there exists* a sequence of converging approximation operators (be they of collocation or Galerkin/projection nature) and derive the corresponding convergence rates. Such sequences of converging approximation operators can be sometimes estimated a priori as in [ZS20, Che18, EST18]. Another possible procedure is to rely instead on a posteriori adaptive strategies: the details of these strategies vary depending on the type of approximation operators (projection/collocation) and, moreover, these a posteriori adaptive strategies are often based on heuristics known to behave well in practice (even better than the a priori constructions) but for which a proof of convergence is often lacking.

For projection approaches, adaptive stochastic Galerkin finite element methods (ASGFEM), which control the discretization of both physical and parametric variables, are well-studied. The extensive research activity in the last years comprises in particular residual-based error estimators [EGSZ14, EGSZ15, EM16, EPS17] and hierarchical error estimators [BPS14, BS16, CPB19, BPRR19a]. The setting in these works is similar to the one considered here, i.e., linear elliptic PDEs with affine parametric coefficients. However, the cited works allow for a countably infinite expansion, which makes an additional dimension adaptivity necessary. With the employed Legendre chaos discretization for the parameter space, only the margin of an active set of polynomials has to be considered in the error estimator. The developed error estimators have been shown to be reliable and efficient, which for hierarchical estimators usually requires additional assumptions. Convergence of an ASGFEM algorithm was first shown in [EGSZ15] for a residual estimator and, using a different argument, in [BPRR19a] for a hierarchical estimator. A goal-oriented error

estimator was presented in [BPRR19b] and the more involved case of nonlinear coefficients and Gaussian parameters has only been considered recently in [EMPS20] with a low-rank hierarchical tensor discretization.

On the stochastic collocation side, the current literature discusses quite extensively algorithms for stochastic adaptivity, whereas much less attention has been devoted to (reliable) spatial adaptivity. To date, most adaptive sparse grid approximation schemes involve some variation of the basic procedure proposed by Gerstner and Griebel in [GG03], see also [Heg03]. This algorithm drives adaptivity in the parameter variables by exploring at each iteration a certain number of sparse subspaces admissible to the approximation and then evaluating for each of these an *error indicator*; this requires solving a certain number of PDEs. The subspace with the largest error indicator is selected and added to the approximation, and a new set of admissible sparse subspaces for the next enrichment step is generated. Several error indicators and variations of the selection strategy have been considered, see e.g. [Kli06, GK09, SS13, CCS14, NTTT16, FGB⁺20]. A crucial point is that these error indicators are *heuristics*. Conversely, the work [GN18] by Guignard and Nobile proposes a variation of the Gerstner–Griebel algorithm based on a reliable residual-based error *estimator* which can control adaptivity in both the physical and parametric variables. Another significant difference compared with typical indicator-based adaptive algorithms is that the procedure proposed in [GN18] evaluates the error estimator *without solving additional PDEs*. This allows significant computational savings with relative to the basic Gerstner–Griebel algorithm. For other works discussing spatial adaptivity in the context of stochastic collocation methods, see [SJ14, LSS19].

Guignard and Nobile give no convergence analysis in [GN18] for their proposed algorithm, and our contribution in this work is to close this gap. We do this by proving convergence of a slight modification of their algorithm (cf. Algorithm 3), thus establishing a convergence result for an adaptive sparse collocation method. This result is stated in Theorem 9. Our convergence analysis is based on a convergence theorem for abstract adaptive approximations (i.e., which covers both projection and collocation approximations, as well as other possible approximation strategies) w.r.t. the parameter variables. We derive this theorem by generalizing results given in [BPRR19a] on convergence of adaptive stochastic Galerkin methods. This approach for proving convergence requires that the employed error estimator possesses the property of *reliability*. In [GN18] Guignard and Nobile already established this property for their error estimator, but only for a specific model problem, namely, an elliptic PDE whose diffusion coefficient depends linearly on a finite number of parameters. Moreover, we also require the underlying univariate sequence of collocation points to be nested in order that the sparse collocation construction be interpolatory. Hence, our particular convergence result is also tied to these assumptions on the underlying PDE and collocation points. However, we believe that the general approach for establishing convergence of adaptive sparse collocation methods presented in this paper might be adapted to more general cases in the future. For instance, upon assuming that the error indicator used in the basic Gerstner–Griebel adaptive algorithm is indeed a reliable error estimator, we are able to prove convergence of this variant of the algorithm as well (see Theorem 10). We note that our analysis considers adaptivity in the parameter variables only, i.e., we focus on the semi-discrete setting. Finally, we mention the simultaneous and independent work [FS20], which also provides a convergence result (and a convergence rate) for adaptive stochastic collocation methods applied to an elliptic PDE with diffusion coefficient depending affinely on finitely many random variables. While the overall framework and the focus of that work is similar to ours, some differences are noteworthy: the algorithm for which [FS20] proves convergence is essentially the one discussed by Guignard and Nobile in [GN18] while we consider a different version and, in addition, we also provide a convergence proof for the original Gerstner–Griebel variant. Furthermore, the line of proof in [FS20], while similar to the present one, has of course some different technical aspects: in particular, our proof is valid for any choice of collocation points over the parameter space, whereas the proof in [FS20] assumes that Clenshaw–Curtis collocation points are used when constructing the sparse grid.

The remainder of this paper is structured as follows. Sections 2 and 3 contain preliminary information: in particular, Section 2 states the model problem and recalls the results in [BPRR19a] that will be instrumental for the rest of the work, while Section 3 gives details on the construction of adaptive sparse grid collocation schemes. Sections 4 and 5 contain our main results: Section 4 contains the statement of the specific adaptive collocation algorithm that we consider (i.e., our version of the Guignard–Nobile algorithm, see Algorithm 3), the associated convergence result (Theorem 9), the convergence result of the Gerstner–Griebel Algorithm (Theorem 10), and some discussion on computational aspects, while Section 5 contains the proof of the convergence result. Finally, conclusions and future research directions are outlined in Section 6.

2 Preliminaries

In this section we specify the model problem under consideration and recall basic properties of its solution. Furthermore, we discuss general adaptive approximations w.r.t. the parameter variables and state an abstract convergence result which provides the basis of our convergence analysis for adaptive sparse grid collocation.

2.1 Model Problem

We consider a common model problem arising in uncertainty propagation via random differential equations, i.e., the stationary diffusion equation containing a coefficient function which depends linearly on a high-dimensional parameter. Specifically, we wish to solve the parametric elliptic boundary value problem

$$-\nabla \cdot (a(\mathbf{y}) \nabla u(\mathbf{y})) = f, \quad \text{on } D \subset \mathbb{R}^d \quad (1a)$$

$$u(\mathbf{y}) = 0, \quad \text{on } \partial D. \quad (1b)$$

The domain $D \subset \mathbb{R}^d$ is assumed to be bounded and Lipschitz, $f \in L^2(D)$ and the coefficient $a(\mathbf{y}) \in L^\infty(D)$ is given by

$$a(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \sum_{m=1}^M a_m(\mathbf{x}) y_m, \quad \mathbf{y} \in \mathbf{\Gamma} := \Gamma^M, \Gamma := [-1, 1], \quad (2)$$

where $M \in \mathbb{N}$ is a finite number and $a_0, \dots, a_M \in L^\infty(D)$. The parametric domain $\mathbf{\Gamma}$ is equipped with a uniform product measure $\mu(d\mathbf{y}) := \bigotimes_{m=1}^M \frac{dy_m}{2}$, i.e., the components of \mathbf{y} can be viewed as i.i.d. uniform random variables over $\Gamma = [-1, 1]$. Further, we assume that the functions $a_0, \dots, a_M \in L^\infty(D)$ satisfy the *uniform ellipticity condition*

$$\sum_{m=1}^M |a_m(\mathbf{x})| \leq a_0(\mathbf{x}) - r, \quad \forall \mathbf{x} \in D, \quad (3)$$

for some $r > 0$. This implies that

$$a_{\min} := \min_{\mathbf{y} \in \mathbf{\Gamma}} \operatorname{ess\,inf}_{\mathbf{x} \in D} a(\mathbf{x}, \mathbf{y}) \geq r > 0. \quad (4)$$

We then define the constant

$$\alpha := 1 - \frac{a_{\min}}{\inf_{\mathbf{x} \in D} a_0(\mathbf{x})} \in (0, 1), \quad (5)$$

which will turn out to be important in Theorem 1 below. Due to the uniform ellipticity assumption, the weak solution $u(\mathbf{y}) \in \mathcal{H} = H_0^1(D)$ exists for any $\mathbf{y} \in \mathbf{\Gamma}$ and satisfies $u \in C(\mathbf{\Gamma}; \mathcal{H})$.

Polynomial expansions In order to approximate the solution u of (1), or rather the parameter-to-solution map $\mathbf{y} \mapsto u(\cdot, \mathbf{y}) \in \mathcal{H}$, we shall analyze polynomial expansions of u in the parameter $\mathbf{y} \in \mathbf{\Gamma}$,

$$u(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k} \in \mathcal{F}} u_{\mathbf{k}}(\mathbf{x}) P_{\mathbf{k}}(\mathbf{y}), \quad \mathcal{F} := \mathbb{N}_0^M, \quad u_{\mathbf{k}} \in \mathcal{H}, \quad (6)$$

where $P_{\mathbf{k}}(\mathbf{y}) = \prod_{m=1}^M P_{k_m}(y_m)$ is a finite product of univariate polynomials $P_k: \Gamma \rightarrow \mathbb{R}$ of degree k with $P_0 \equiv 1$. Two common choices for the basic polynomials P_k are

1. *Taylor polynomials*: $P_{\mathbf{k}}(\mathbf{y}) := \mathbf{y}^{\mathbf{k}} = \prod_{m=1}^M y_m^{k_m}$ where then

$$u_{\mathbf{k}}(\mathbf{x}) = t_{\mathbf{k}}(\mathbf{x}) := \frac{1}{\mathbf{k}!} \partial^{\mathbf{k}} u(\mathbf{x}, \mathbf{0}),$$

2. *Legendre polynomials*: $P_{\mathbf{k}}(\mathbf{y}) := L_{\mathbf{k}}(\mathbf{y}) = \prod_{m=1}^M L_{k_m}(y_m)$ with L_k denoting the k th $L_{\mu_1}^2$ -normalized Legendre polynomial w.r.t. the uniform distribution $\mu_1(dx) = \frac{dy}{2}$ on $\Gamma = [-1, 1]$ and

$$u_{\mathbf{k}}(\mathbf{x}) := \int_{\mathbf{\Gamma}} u(\mathbf{x}, \mathbf{y}) L_{\mathbf{k}}(\mathbf{y}) \mu(d\mathbf{y}).$$

Since $u \in C(\mathbf{\Gamma}; \mathcal{H}) \subset L_{\mu}^2(\mathbf{\Gamma}; \mathcal{H})$ we have that the expansion (6) using Legendre polynomials converges in $L_{\mu}^2(\mathbf{\Gamma}; \mathcal{H})$. The following result due to [BCM17] establishes under suitable assumptions an ℓ^p -summability of both Taylor and Legendre coefficients which, for instance, implies that the Taylor expansion (6) of u converges in $L^\infty(\mathbf{\Gamma}; \mathcal{H})$.

Algorithm 1 Generic adaptive algorithm

 $\Lambda_0 = \{\mathbf{0}\}$ $u_0 := S_{\Lambda_0} u$ **for** $n \in \mathbb{N}_0$ **do** Choose a *candidate set* of multi-indices $\mathcal{C}_n \subset \mathcal{F} \setminus \Lambda_n$ for enriching Λ_n

Evaluate estimates of the error contribution on the candidate set:

$$\eta_n(\mathbf{k}) = \eta(\mathbf{k}, u_n), \quad \mathbf{k} \in \mathcal{C}_n$$

 Determine *marked indices* $\mathcal{M}_n \subset \mathcal{C}_n$ (according to a given marking strategy based on $\eta_n(\mathbf{k})$); Set $\Lambda_{n+1} := \Lambda_n \cup \mathcal{M}_n$ Set $u_{n+1} := S_{\Lambda_{n+1}} u$.**end for**

Theorem 1 ([BCM17, Theorem 2.2 & 3.1, Corollary 2.3 & 3.2]). *Let the condition (3) for a as in (2) be satisfied. Then a unique solution u of the corresponding elliptic problem (1) exists and belongs to $C(\Gamma; \mathcal{H})$. Moreover, for any $\boldsymbol{\rho} := (\rho_m)_{m=1}^M$ with $1 < \rho_m < \alpha^{-1}$ with α as in (5)*

1. *the Taylor coefficients $t_{\mathbf{k}} \in \mathcal{H}$ of u satisfy $(\boldsymbol{\rho}^{\mathbf{k}} \|t_{\mathbf{k}}\|_{\mathcal{H}})_{\mathbf{k} \in \mathcal{F}} \in \ell^2(\mathcal{F})$,*
2. *and the Legendre coefficients $u_{\mathbf{k}} \in \mathcal{H}$ of u satisfy $(b_{\mathbf{k}}^{-1} \boldsymbol{\rho}^{\mathbf{k}} \|u_{\mathbf{k}}\|_{\mathcal{H}})_{\mathbf{k} \in \mathcal{F}} \in \ell^2(\mathcal{F})$ with $b_{\mathbf{k}} := \prod_{m=1}^M \sqrt{1 + 2k_m}$.*

Remark 2. *The authors of [BCM17] actually consider the infinite-dimensional noise case, i.e., with $M = \infty$ in (2), and prove the results stated in Theorem 1 under the assumption that*

$$\left\| \frac{\sum_{m=1}^{\infty} \rho_m |a_m|}{a_0} \right\|_{C(D)} < 1,$$

for a sequence $\boldsymbol{\rho} := (\rho_m)_{m \geq 1}$ with $\rho_m > 1$. Hence, Theorem 1 can be derived easily from this general case by setting $a_m(\mathbf{x}) \equiv 0$ and $\rho_m > 1$ arbitrarily for $m > M$:

$$\left\| \frac{\sum_{m=1}^{\infty} \rho_m |a_m|}{a_0} \right\|_{C(D)} = \left\| \frac{\sum_{m=1}^M \rho_m |a_m|}{a_0} \right\|_{C(D)} < \alpha^{-1} \left\| \frac{\sum_{m=1}^{\infty} |a_m|}{a_0} \right\|_{C(D)} \leq \alpha^{-1} (1 - a_{\min}) = 1.$$

2.2 Adaptive Polynomial Approximation

Given the decay rate stated in Theorem 1 for the norms of the coefficients $u_{\mathbf{k}}$ of the expansion (6), a polynomial approximation of u seems feasible. To this end, we consider the truncated expansions u_{Λ} based on a finite multi-index set $\Lambda \subset \mathcal{F}$,

$$u_{\Lambda} := S_{\Lambda} u = \sum_{\mathbf{k} \in \Lambda} \hat{u}_{\mathbf{k}} P_{\mathbf{k}}, \quad \hat{u}_{\mathbf{k}} \in \mathcal{H},$$

where S_{Λ} denotes a suitable *approximation operator* and $\hat{u}_{\mathbf{k}}$ are approximations to the true coefficients $u_{\mathbf{k}}$ of u (cf. (6)). For instance, S_{Λ} could be the operator associated with a Galerkin approach for approximating u using the finite-dimensional polynomial space

$$\mathcal{P}_{\Lambda}(\Gamma) := \text{span} \{P_{\mathbf{k}} : \mathbf{k} \in \Lambda\},$$

or, as we in our case later, the operator associated to sparse grid collocation based on Λ . At this point we do not need to further specify S_{Λ} .

We consider in particular an *adaptive approach* to compute such polynomial approximations u_{Λ} . More specifically, starting from an initial set $\Lambda_0 \subset \mathcal{F}$ we construct nested multiindex sets $\Lambda_n \subset \Lambda_{n+1}$, $n \in \mathbb{N}_0$, and compute the associated polynomial approximations $u_n := S_{\Lambda_n} u$ by the generic adaptive algorithm detailed in Algorithm 1. Again, we do not further specify how to compute the estimates $\eta_n(\mathbf{k}) = \eta(\mathbf{k}, u_n)$ at this point. Instead, we provide a fairly general convergence theorem for Algorithm 1, stating conditions on $\eta_n(\mathbf{k})$ that guarantee convergence of the algorithm.

The following theorem draws upon the work [BPRR19a] on the convergence of adaptive stochastic Galerkin methods. Specifically, it is a compact summary of a way of proving for convergence for stochastic Galerkin outlined in detail in [BPRR19a, Section 6 and 7], slightly modified to fit the application to adaptive sparse grid collocation. We state the theorem here and provide the proof at the end of the section.

Theorem 3 (cf. [BPRR19a]). *Let u_n denote the approximations constructed via Algorithm 1. Assume that*

1. *the total error estimator $\eta_n := \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_n(\mathbf{k})$ is reliable, i.e., there exists a constant $C < \infty$ independent of n such that*

$$\|u - u_n\| \leq C\eta_n,$$

where $\|\cdot\|$ denotes a suitable norm for functions $v: \Gamma \rightarrow \mathcal{H}$,

2. *there exists a sequence of non-negative numbers $(\eta_\infty(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}} \in \ell^1(\mathcal{F})$ such that for $(\hat{\eta}_n(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}}$ with $\hat{\eta}_n(\mathbf{k}) := \eta_n(\mathbf{k})$ for $\mathbf{k} \in \mathcal{C}_n \cup \Lambda_n$ and $\hat{\eta}_n(\mathbf{k}) = 0$ otherwise, we have*

$$\lim_{n \rightarrow \infty} \|\eta_\infty - \hat{\eta}_n\|_{\ell^1} = 0,$$

3. *there exists a constant $c > 0$ independent of n such that for all $\mathbf{k} \in \mathcal{C}_n \setminus \mathcal{M}_n$ we have*

$$\eta_n(\mathbf{k}) \leq c \sum_{\mathbf{i} \in \mathcal{M}_n} \eta_n(\mathbf{i}).$$

From these assumptions it follows that

$$\lim_{n \rightarrow \infty} \|u - u_n\| = 0.$$

Remark 4. *Before we prove the theorem, we comment on the second and third assumption:*

1. *The third assumption is generally easily to satisfy. For instance, simply choosing $\mathcal{M}_n := \arg \max_{\mathbf{k} \in \mathcal{C}_n} \eta_n(\mathbf{k})$ satisfies the assumption with $c = 1$.*
2. *For sparse grid collocation, the second assumption turns out to be the most difficult to verify. Moreover, it is probably the most cryptic assumption of the theorem. It can usually be verified as follows: assuming the sequence u_n has a limit u_∞ with corresponding error estimators $\eta_\infty(\mathbf{k}) := \eta(\mathbf{k}, u_\infty)$, conclude from $u_n \rightarrow u_\infty$ that $\|\eta_\infty - \hat{\eta}_n\|_{\ell^1} \rightarrow 0$ by exploiting continuity properties of the error estimator $\eta(\mathbf{k}, u_n)$ w.r.t. u_n . Note that in principle u_∞ is just the limit of u_n , but does not necessarily coincide with the actual solution of the PDE (1). The fact that $u_\infty = u$ is the assertion of the theorem.*
3. *As we will see in the proof of Theorem 3, the second assumption represents some kind of saturation of the reliable error estimators η_n : since $\|\eta_\infty - \hat{\eta}_n\|_{\ell^1} \rightarrow 0$ we have that*

$$\eta_n \leq \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k}) + \sum_{\mathbf{k} \in \mathcal{C}_n} |\eta_n(\mathbf{k}) - \eta_\infty(\mathbf{k})| \leq \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k}) + \|\hat{\eta}_n - \eta_\infty\|_{\ell^1}$$

converges to zero if $\sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k})$ does. Since $\|\hat{\eta}_n - \eta_\infty\|_{\ell^1} < \infty$ we can expect $\eta_\infty(\mathbf{k})$ to decay for large multi-indices \mathbf{k} . Thus, if \mathcal{C}_n tends to include increasingly larger multi-indices \mathbf{k} , then $\sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k})$ should decay to zero. This will be made rigorous in the subsequent proof.

The proof of Theorem 3 employs the following abstract lemma which was shown for the case $p = 2$ in [BPRR19a, Lemma 15]. Since their proof can be generalized to arbitrary $1 \leq p < \infty$ without any significant modification we simply state the result and refer to [BPRR19a, Lemma 15] for a detailed proof.

Lemma 5 (cf. [BPRR19a, Lemma 15]). *Let $\mathbf{z} = (z_k)_{k \in \mathbb{N}} \in \ell^p(\mathbb{N})$, $p \in [1, \infty)$, and $\mathbf{z}^{(n)} = (z_k^{(n)})_{k \in \mathbb{N}} \in \ell^p(\mathbb{N})$, $n \in \mathbb{N}_0$, be sequences of non-negative numbers satisfying $\lim_{n \rightarrow \infty} \|\mathbf{z} - \mathbf{z}^{(n)}\|_{\ell^p} = 0$. Assume further that there exists a continuous function $g: [0, \infty) \rightarrow [0, \infty)$ with $g(0) = 0$ and a sequence of nested subsets $\mathcal{J}_n \subset \mathbb{N}$, i.e., $\mathcal{J}_n \subset \mathcal{J}_{n+1}$, such that*

$$\forall n \in \mathbb{N}_0 \quad \forall k \notin \mathcal{J}_{n+1}: \quad z_k^{(n)} \leq g \left(\sum_{i \in \mathcal{J}_{n+1} \setminus \mathcal{J}_n} \left(z_i^{(n)} \right)^p \right).$$

Then $\lim_{n \rightarrow \infty} \sum_{k \notin \mathcal{J}_n} z_k^p = 0$.

Proof of Theorem 3. Since the error estimator is reliable, we only need to show that

$$\lim_{n \rightarrow \infty} \eta_n = \lim_{n \rightarrow \infty} \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_n(\mathbf{k}) = 0.$$

Due to

$$\sum_{\mathbf{k} \in \mathcal{C}_n} \eta_n(\mathbf{k}) \leq \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k}) + \sum_{\mathbf{k} \in \mathcal{C}_n} |\eta_n(\mathbf{k}) - \eta_\infty(\mathbf{k})| \leq \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k}) + \|\hat{\eta}_n - \eta_\infty\|_{\ell^1(\mathbb{N})},$$

as well as $\|\hat{\eta}_n - \eta_\infty\|_{\ell^1} \rightarrow 0$ by assumption, the statement of the theorem follows if

$$\lim_{n \rightarrow \infty} \sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k}) = 0.$$

In order to show this we apply Lemma 5 as follows: we identify the countable set \mathcal{F} with \mathbb{N} , η_∞ with \mathbf{z} and $\hat{\eta}_n$ with $\mathbf{z}^{(n)}$. Recall that by assumption $\|\hat{\eta}_n - \eta_\infty\|_{\ell^1} \rightarrow 0$. Thus, the first assumption of Lemma 5 is satisfied. Moreover, we identify the $\Lambda_n \subset \mathcal{F}$ with $\mathcal{J}_n \subset \mathbb{N}$. These sets are nested and $\mathcal{J}_{n+1} \setminus \mathcal{J}_n$ corresponds to \mathcal{M}_n . By our third assumption and the construction of $\hat{\eta}_n$ there holds for each $n \in \mathbb{N}$

$$\hat{\eta}_n(\mathbf{k}) \leq c \sum_{i \in \mathcal{M}_n} \hat{\eta}_n(i) \quad \forall \mathbf{k} \notin \Lambda_{n+1},$$

since $\hat{\eta}_n(\mathbf{k}) = 0$ for $\mathbf{k} \notin \mathcal{C}_n \cup \Lambda_n$ and $(\mathcal{C}_n \cup \Lambda_n) \setminus \Lambda_{n+1} = \mathcal{C}_n \setminus \mathcal{M}_n$. Thus, the second assumption of Lemma 5 is also satisfied with $g(s) = cs$. Hence, we can apply Lemma 5 to $\mathbf{z} \simeq \eta_\infty$ and $\mathbf{z}_n \simeq \hat{\eta}_n$ and obtain that

$$\lim_{n \rightarrow \infty} \sum_{\mathbf{k} \notin \Lambda_n} \eta_\infty(\mathbf{k}) = 0,$$

which by $\sum_{\mathbf{k} \in \mathcal{C}_n} \eta_\infty(\mathbf{k}) \leq \sum_{\mathbf{k} \notin \Lambda_n} \eta_\infty(\mathbf{k})$ concludes the proof. \square

3 Adaptive Sparse Collocation

We now introduce the sparse collocation approach and discuss how adaptive sparse grid algorithms can be derived from the abstract Algorithm 1. In particular, we show how to obtain the classical a-posteriori adaptive algorithm by Gerstner and Griebel [GG03] based on heuristic error *indicators* (as opposed to reliable error *estimators*, as proposed by Guignard and Nobile in [GN18]). As already discussed in the introduction, changing from indicators to estimators is key to proving convergence. Our version of the estimator-based algorithm by Guignard and Nobile and its convergence are then discussed in the subsequent sections.

Univariate interpolation nodes The first ingredient for any sparse grid construction is the choice of the underlying univariate sequences of collocation points. In this work, we consider nested point sequences: Let $(y_{(i)})_{i \in \mathbb{N}_0} \subset [-1, 1]$ denote a sequence of univariate interpolation nodes and define the associated node sets

$$\mathcal{Y}_k := \{y_{(i)} : i = 0, \dots, \mathbf{m}(k)\} \subset \Gamma, \quad k \in \mathbb{N}_0, \quad (7)$$

where $\mathbf{m} : \mathbb{N}_0 \rightarrow \mathbb{N}_0$ denotes the *growth function* of the sets \mathcal{Y}_k , i.e., $|\mathcal{Y}_k| = 1 + \mathbf{m}(k)$. We assume throughout that $\mathbf{m}(0) = 0$ and that \mathbf{m} is strictly increasing. Thus, we exclude *delayed sequences* of node sets with $\mathcal{Y}_k = \mathcal{Y}_{k+1}$ for certain k as sometimes employed for sparse grid methods, see [Pet03]. As an immediate consequence of our assumption, we have $\mathbf{m}(k) \geq k$ and $|\mathcal{Y}_k| \geq k + 1$. We later also use the *generalized inverse* of the growth function given for $i \in \mathbb{N}_0$ by

$$\mathbf{m}^{-1}(i) := \min\{k \in \mathbb{N}_0 : i \leq \mathbf{m}(k)\} \leq i, \quad (8)$$

which gives the index of the first node set \mathcal{Y}_k which contains $y_{(i)}$. A particularly convenient construction of such nested nodes is provided by Leja points. Leja sequences on $\Gamma = [-1, 1]$ are defined recursively by first choosing $y_{(0)} \in \Gamma$ and then setting

$$y_{(k)} = \arg \max_{y \in \Gamma} \prod_{i=0}^{k-1} |y - y_{(i)}|, \quad k \in \mathbb{N}_0, \quad (9)$$

see e.g. [Chk13, CCS14, Chk15, SS13, NTT15] and the references therein. The standard choice is to set $y_{(0)} = -1$; the rule (9) then leads to

$$y_{(0)} = -1, \quad y_{(1)} = 1, \quad y_{(2)} = 0, \quad y_{(3)} \approx -0.57735, \quad y_{(4)} \approx 0.65871, \quad \dots$$

Another common sequence, referred to as *R-Leja* (real Leja) points, is obtained by carrying out the Leja construction on the upper unit circle in the complex plane in place of $\Gamma = [-1, 1]$ and then projecting the sequence thus obtained onto the real line. This results in (see e.g. [Chk13] for a proof):

$$y_{(i)} = \cos \phi_{(i)}, \quad i \in \mathbb{N}_0,$$

$$\phi_{(0)} = 0, \quad \phi_{(1)} = \pi, \quad \phi_{(2)} = \pi/2, \quad \phi_{(2n+1)} = \frac{\phi_{(n+2)}}{2}, \quad \phi_{(2n+2)} = \phi_{(2n+2)} + \pi.$$

For both Leja and R-Leja nodes, we may utilize any strictly increasing growth function \mathbf{m} with $\mathbf{m}(0) = 0$ to construct nested node sets $\mathcal{Y}_k \subset \mathcal{Y}_{k+1}$ as in (7). The most common choice uses sets growing in unit increments, i.e., $\mathbf{m}(i) = i$.

Besides the Leja construction, *Clenshaw-Curtis* nodes are also popular collocation points. Here, the node sets \mathcal{Y}_k consist of the extrema of Chebyshev polynomials

$$\mathcal{Y}_0 = \{0\}, \quad \mathcal{Y}_k = \{-\cos(\pi i / \mathbf{m}(k)) : i = 0, \dots, \mathbf{m}(k)\}, \quad k \in \mathbb{N}.$$

Nestedness of the \mathcal{Y}_k is then achieved by the *doubling rule* $\mathbf{m}(k) = 2^k$ for $k \geq 1$. The corresponding sequence of nodes $(y_{(i)})_{i \in \mathbb{N}_0}$ is given, suitably arranged, by

$$\begin{aligned} y_{(0)} &= 0, \\ y_{(1)} &= -\cos(0), & y_{(2)} &= -\cos(\pi), \\ y_{(3)} &= -\cos(1/4\pi), & y_{(4)} &= -\cos(3/4\pi), \dots \end{aligned}$$

Sparse collocation We consider *hierarchical* sparse collocation based on nested sequences of node sets \mathcal{Y}_k as introduced above. Let $\mathcal{P}_k(\Gamma)$ denote the set of univariate polynomials on Γ of degree at most $k \in \mathbb{N}_0$. We can then define for any Hilbert space-valued continuous function $f: \Gamma \rightarrow \mathcal{H}$ two objects:

- a Lagrange interpolant $\mathcal{I}_k: C(\Gamma; \mathcal{H}) \rightarrow \mathcal{P}_{\mathbf{m}(k)}(\Gamma; \mathcal{H})$,
- a univariate *detail operator* $\Delta_k: C(\Gamma; \mathcal{H}) \rightarrow \mathcal{P}_{\mathbf{m}(k)}(\Gamma; \mathcal{H})$,

$$\Delta_0 = \mathcal{I}_0, \quad \Delta_k := \mathcal{I}_k - \mathcal{I}_{k-1}, \quad k \in \mathbb{N}.$$

With these definitions, we have that

$$\Delta_i f = 0 \quad \forall f \in \mathcal{P}_k(\Gamma, \mathcal{H}), \quad \forall i > \mathbf{m}^{-1}(k). \quad (10)$$

Since $\Delta_k f = \mathcal{I}_k f - \mathcal{I}_{k-1} f = \mathcal{I}_k(f - \mathcal{I}_{k-1} f)$, and due to the nestedness of the node sets $\mathcal{Y}_{k-1} \subset \mathcal{Y}_k$, the detail operators may be expressed as

$$\begin{aligned} \Delta_k f &= \sum_{i=\mathbf{m}(k-1)+1}^{\mathbf{m}(k)} [f(y_{(i)}) - \mathcal{I}_{k-1} f(y_{(i)})] \ell_i^{(\mathbf{m}(k))}, \\ \ell_i^{(\mathbf{m}(k))}(y) &:= \prod_{j=0, j \neq i}^{\mathbf{m}(k)} \frac{y - y_{(j)}}{y_{(i)} - y_{(j)}} \in \mathcal{P}_{\mathbf{m}(k)} \quad \text{for } i \in \{\mathbf{m}(k-1) + 1, \dots, \mathbf{m}(k)\}. \end{aligned}$$

It is therefore convenient to introduce the notation

$$h_i(y) := \ell_i^{(\mathbf{m}(k))}(y), \quad y \in \Gamma, \quad (11)$$

where $i \in \{\mathbf{m}(k-1) + 1, \dots, \mathbf{m}(k)\}$. The polynomials h_i , each associated to a node $y_{(i)}$, $i \in \mathbb{N}_0$, are called *hierarchical Lagrange polynomial*¹, $h_i \in \mathcal{P}_{\mathbf{m}(k)}$. The quantity $f(y_{(i)}) - \mathcal{I}_{k-1} f(y_{(i)}) = (f - \mathcal{I}_{k-1} f)(y_{(i)})$ is also called *hierarchical surplus*. Next, consider tensorized detail operators

$$\Delta_{\mathbf{i}} := \bigotimes_{m=1}^M \Delta_{i_m}, \quad \Delta_{\mathbf{i}}: C(\Gamma; \mathcal{H}) \rightarrow \mathcal{P}_{\mathbf{m}(\mathbf{i})}(\Gamma; \mathcal{H}),$$

where $\mathbf{m}(\mathbf{i}) = (\mathbf{m}(i_1), \dots, \mathbf{m}(i_M)) \in \mathbb{N}^M$ and

$$\mathcal{P}_{\mathbf{m}(\mathbf{i})} = \text{span}\{\mathbf{y}^j : j_m \leq \mathbf{m}(i_m) \text{ for } m = 1, \dots, M\}.$$

¹The difference from the standard Lagrange polynomials is that h_i is only defined for the nodes $y_{(i)}$ most recently added, with $i \in \{\mathbf{m}(k-1) + 1, \dots, \mathbf{m}(k)\}$, whereas the standard Lagrange polynomials are redefined for all $i \in \{1, \dots, \mathbf{m}(k)\}$ when new nodes are added.

Given a (finite) subset $\Lambda \subset \mathcal{F}$ we define the *sparse grid collocation operator* associated with the *sparse grid* \mathcal{Y}_Λ by

$$S_\Lambda := \sum_{\mathbf{i} \in \Lambda} \Delta_{\mathbf{i}}, \quad \mathcal{Y}_\Lambda := \bigcup_{\mathbf{i} \in \Lambda} \mathcal{Y}_{\mathbf{i}}, \quad \mathcal{Y}_{\mathbf{i}} := \mathcal{Y}_{i_1} \times \mathcal{Y}_{i_2} \times \dots \times \mathcal{Y}_{i_M}.$$

We require the multi-index sets $\Lambda \subset \mathcal{F}$ to be *downward-closed* (or *monotone*), which means that $\mathbf{i} \in \Lambda$ implies $\mathbf{i} - \mathbf{e}_m \in \Lambda$, where \mathbf{e}_m denotes the m th canonical unit multi-index. Downward-closedness of Λ implies three facts (see e.g. [EST18]): First,

$$\mathcal{Y}_\Lambda = \{\mathbf{y}_{(\mathbf{j})} : \mathbf{j} \leq \mathbf{m}(\mathbf{i}), \mathbf{i} \in \Lambda\}, \quad \mathbf{y}_{(\mathbf{j})} := (y_{(j_1)} \ y_{(j_2)} \ \dots \ y_{(j_M)}) \in \mathbf{\Gamma},$$

where $\mathbf{j} \leq \mathbf{m}(\mathbf{i})$ is understood componentwise; second, that the sparse grid collocation operator yields an approximation in $\mathcal{P}_{\mathbf{m}(\Lambda)}(\mathbf{\Gamma}; \mathcal{H})$,

$$S_\Lambda : C(\mathbf{\Gamma}; \mathcal{H}) \rightarrow \mathcal{P}_{\mathbf{m}(\Lambda)}(\mathbf{\Gamma}; \mathcal{H}), \quad \mathbf{m}(\Lambda) := \{\mathbf{j} \in \mathcal{F} : \mathbf{j} \leq \mathbf{m}(\mathbf{i}) \text{ for some } \mathbf{i} \in \Lambda\};$$

and third, together with the nestedness of the node sets, that S_Λ is *interpolatory*, i.e.,

$$S_\Lambda f(\mathbf{y}_{(\mathbf{i})}) = f(\mathbf{y}_{(\mathbf{i})}) \quad \forall \mathbf{y}_{(\mathbf{i})} \in \mathcal{Y}_\Lambda.$$

Remark 6. For finite and monotone multi-index sets Λ there exists $N \in \mathbb{N}$ multi-indices $\mathbf{i}_1, \dots, \mathbf{i}_N \in \Lambda$ such that

$$\Lambda = \bigcup_{n=1}^J \mathcal{R}_{\mathbf{i}_n}, \quad \mathcal{R}_{\mathbf{i}} := \{\mathbf{j} \in \mathcal{F} : \mathbf{j} \leq \mathbf{i}\},$$

i.e., the multiindices \mathbf{i}_n can be viewed as the corners of Λ . As an immediate consequence, we have

$$\mathcal{P}_{\mathbf{m}(\Lambda)}(\mathbf{\Gamma}; \mathcal{H}) = \bigoplus_{n=1}^N \mathcal{P}_{\mathbf{m}(\mathbf{i}_n)}(\mathbf{\Gamma}; \mathcal{H}).$$

Adaptive sparse collocation algorithms Two ways to construct monotone multi-index sets Λ for (hierarchical) sparse grid collocation are the classical algorithm introduced by Gerstner and Griebel in [GG03] (as well as numerous variations mentioned in the literature surveyed in the introduction) and the alternative algorithm introduced by Guignard and Nobile in [GN18]. Both can be seen as specific instances of the generic Algorithm 1. We describe the former here and the latter (or rather, a slight variation thereof) in the next section, together with a convergence analysis. To introduce these algorithms, we need to specify three “ingredients”: the candidate set \mathcal{C}_n , a *marking strategy* for determining marked sets $\mathcal{M}_n \subset \mathcal{C}_n$, and corresponding estimates $\eta_n(\mathbf{k})$ for the error contribution of indices in the candidate set. To this end, we require the following definitions (see also Figure 1):

- The *margin* $\text{Marg}(\Lambda) \subset \mathcal{F}$ of a multi-index set $\Lambda \subset \mathcal{F}$ is given by

$$\text{Marg}(\Lambda) := \{\mathbf{k} \in \mathcal{F} \setminus \Lambda : \mathbf{k} - \mathbf{e}_m \in \Lambda \text{ for some } m \in \mathbb{N}\}.$$

- The *reduced margin* $\text{R}(\Lambda) \subset \text{Marg}(\Lambda)$ of a subset $\Lambda \subset \mathcal{F}$ is given by

$$\text{R}(\Lambda) := \{\mathbf{k} \in \text{Marg}(\Lambda) : \mathbf{k} - \mathbf{e}_m \in \Lambda \text{ for all } m \in \mathbb{N}\}.$$

- The *monotone envelope* $E_\Lambda(\mathbf{k}) \subset \text{Marg}(\Lambda)$ of a multi-index $\mathbf{k} \in \text{Marg}(\Lambda)$:

$$E_\Lambda(\mathbf{k}) := \bigcap \{E \subset \text{Marg}(\Lambda) : \mathbf{k} \in E \text{ and } \Lambda \cup E \text{ is monotone}\}. \quad (12)$$

Note that $E_\Lambda(\mathbf{k}) \cup \Lambda$ is the smallest (in cardinality) monotone multi-index set containing $\Lambda \cup \{\mathbf{k}\}$ and that for $\mathbf{k} \in \text{R}(\Lambda)$ we have $E_\Lambda(\mathbf{k}) = \{\mathbf{k}\}$ by construction.

The adaptive procedure in [GG03] now chooses

- as candidate set \mathcal{C}_n the reduced margin of Λ_n , i.e. $\mathcal{C}_n = \text{R}(\Lambda_n)$;
- as estimators η_n , approximating the error contribution of $\mathbf{k} \in \mathcal{C}_n$ by the L^p -norm of the hierarchical surplus, i.e.,

$$\eta_n(\mathbf{k}) = \|\Delta_{\mathbf{k}} u\|_{L^p_\mu(\mathbf{\Gamma}; \mathcal{H})}, \quad \mathbf{k} \in \text{R}(\Lambda_n). \quad (13)$$

Note that this is merely an error *indicator* and not a proper *estimator*, i.e., no proof of the properties required by Theorem 3 is available. A large body of literature, however, provides numerical evidence that this error indicator is quite robust and gives good results in practice;

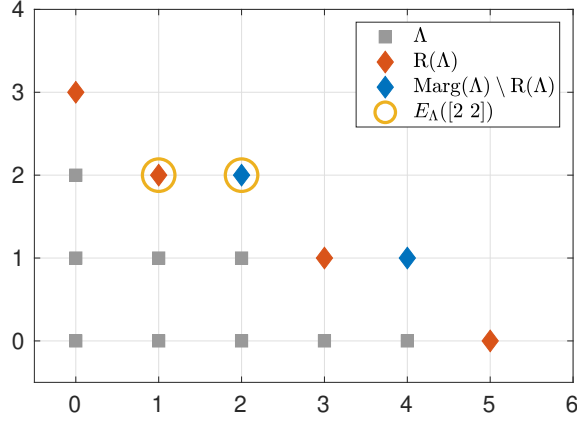


Figure 1: A multi-index set $\Lambda \subset \mathbb{N}_0^2$ (gray squares) and its margin $\text{Marg}(\Lambda)$ (colored diamonds): more specifically, the multi-indices of $\text{Marg}(\Lambda)$ that also belong to the reduced margin $R(\Lambda)$ are colored in red, whereas the remaining ones are colored in blue. Finally, we mark with yellow circles the indices of $\text{Marg}(\Lambda)$ that constitute $E_\Lambda([2, 2])$, i.e. the monotone envelope of $\mathbf{k} = [2, 2]$.

Algorithm 2 Adaptive sparse grid algorithm of Gerstner and Griebel [GG03]

- 1: $\Lambda_0 := \{\mathbf{0}\}$
- 2: $u_0 := S_{\Lambda_0} u$
- 3: **for** $n \in \mathbb{N}_0$ **do**
- 4: Compute reduced margin $R(\Lambda_n)$
- 5: Compute error indicators (reduced margin):

$$\eta_n(\mathbf{k}) = \|\Delta_{\mathbf{k}} u\|_{L_\mu^p(\Gamma; \mathcal{H})}, \quad \mathbf{k} \in R(\Lambda_n)$$

- 6: Choose $\mathbf{k}_n^* := \arg \max_{\mathbf{k} \in R(\Lambda_n)} \eta_n(\mathbf{k})$
 - 7: Set $\Lambda_{n+1} := \Lambda_n \cup \{\mathbf{k}_n^*\}$ and $u_{n+1} := S_{\Lambda_{n+1}} u$.
 - 8: **end for**
-

- as marking strategy, to select the index in the reduced margin which maximizes the value of η_n , i.e., $\mathcal{M}_n = \{\arg \max_{\mathbf{k} \in R(\Lambda_n)} \eta_n(\mathbf{k})\}$. An alternative strategy would be to use Dörfler marking and mark e.g. the 50% of the indices in the reduced margin with the largest η_n , cf. [Dör96].

Algorithm 2 summarizes the Gerstner–Griebel scheme as pseudocode. Note that, since S_Λ is interpolatory for \mathcal{Y}_n nested and Λ monotone, we can efficiently compute η_n in (13), and therefore $S_{\Lambda_{n+1}}$ based on S_{Λ_n} . For this, let $\mathbf{i} \in R(\Lambda_n)$ and $\Lambda_{n+1} = \Lambda_n \cup \{\mathbf{i}\}$. Then,

$$\Delta_{\mathbf{i}} u = \sum_{\mathbf{y}_{(j)} \in \mathcal{Y}_{\mathbf{i}} \setminus \mathcal{Y}_\Lambda} [u(\mathbf{y}_{(j)}) - (S_{\Lambda_n} u)(\mathbf{y}_{(j)})] h_j, \quad h_j(\mathbf{y}) := \prod_{m=1}^M h_{j_m}(y_m),$$

where the h_i are the univariate hierarchical Lagrange polynomials defined in (11) and the set of additional nodes $\mathcal{Y}_{\mathbf{i}}^+ := \mathcal{Y}_{\mathbf{i}} \setminus \mathcal{Y}_\Lambda$ is

$$\mathcal{Y}_{\mathbf{i}}^+ = \mathcal{Y}_{i_1}^+ \times \mathcal{Y}_{i_2}^+ \times \dots \times \mathcal{Y}_{i_M}^+, \quad \mathcal{Y}_{\mathbf{i}}^+ := \mathcal{Y}_{\mathbf{i}} \setminus \mathcal{Y}_{i-1} = \{\mathbf{y}_{(j)} : \mathbf{m}(i-1) + 1 \leq j \leq \mathbf{m}(i)\}.$$

The main shortcoming of this approach is that the computation of $\Delta_{\mathbf{i}} u$ requires solving the PDE to evaluate $u(\mathbf{y}_{(i)})$, and for this reason one may refer to this algorithm as *fully a posteriori*. Clearly, it would be a waste of computational resources to discard these additional PDE solutions: therefore, practical implementations of Algorithm 2 ultimately augment Λ to $\Lambda_{\text{end}} = \Lambda_n \cup R(\Lambda_n)$ at the last iteration and return $u_{\text{end}} = S_{\Lambda_{\text{end}}}$ instead of S_{Λ_n} . Nonetheless, this procedure is “suboptimal” in terms of computational effort. If the reduced margin is large, this operation can be expensive. Moreover, as previously mentioned, the choice of η_n in (13) is a heuristic and no convergence proof for the adaptive algorithm is available. To overcome this issue, we introduce and analyze in the next section another variation of Algorithm 1, for which we can prove convergence.

We close this section by pointing out that using a hierarchical basis is convenient but not necessary, and the standard (non-hierarchical) Lagrange basis can also be used to implement

Algorithm 2. To this end, one would need to draw on the so-called *combination technique* [GSZ92] for evaluating the detail operators $\Delta_i u$ as a linear combination of tensorized Lagrange interpolants,

$$\Delta_i u = \sum_{j \in \{0,1\}^M} (-1)^{|j|} (\mathcal{I}_{i_1-j_1} \otimes \mathcal{I}_{i_2-j_2} \otimes \cdots \otimes \mathcal{I}_{i_M-j_M}) u,$$

and to adjust the computation of $S_\Lambda u$ accordingly, see e.g. [NTTT16, GN18]; this has the advantage that non-nested sequences of node sets (such as zeros of orthogonal polynomials) can be used if desired, see e.g. [NTTT16, EST18].

4 Adaptive Sparse Collocation for the Diffusion Problem

We now turn attention to our above-mentioned slight variation of the adaptive algorithm by Guignard and Nobile from [GN18]; see Remark 8 for a discussion on the difference between the two versions. This algorithm is based on the following error estimator, for which reliability has been established in [GN18].

Proposition 7 ([GN18, Proposition 4.3]). *Let u denote the solution of the random elliptic PDE given in equation (1) with linear diffusion coefficient as in (2), and let $\Lambda \subset \mathcal{F}$ be a monotone subset such that the sparse grid collocation operator S_Λ as introduced in Section 3 is interpolatory. Then, for any $p \in [1, \infty]$ we have*

$$\|u - S_\Lambda u\|_{L_\mu^p(\Gamma; H_0^1(D))} \leq \frac{1}{a_{\min}} \sum_{\mathbf{k} \in \text{Marg}(\Lambda)} \|\Delta_{\mathbf{k}}(a \nabla S_\Lambda u)\|_{L_\mu^p(\Gamma; L^2(D))}.$$

This proposition suggests $\eta_n(\mathbf{k}) := \|\Delta_{\mathbf{k}}(a \nabla S_{\Lambda_n} u)\|_{L_\mu^p(\Gamma; L^2(D))}$ as an error estimator for adaptively constructing the sparse grid approximations $u_n = S_{\Lambda_n} u$ and also to consider the entire margins $\text{Marg}(\Lambda_n)$ as candidate sets. This yields Algorithm 3. Note here that the value $p \in [1, \infty]$ has to be chosen in advance and that $\mathcal{C}_n := \text{Marg}(\Lambda_n) \subset \mathcal{F}$ is, in fact, finite for finite M . Moreover, we highlight that Proposition 7 implies that Algorithm 3 satisfies the first assumption (reliable error estimator) of the abstract convergence result, stated in Theorem 3. Besides that, also the third assumption of Theorem 3 is satisfied by construction, i.e., by the marking strategy $\mathcal{M}_n := E_{\Lambda_n}(\mathbf{k}_n^*)$ (where $E_{\Lambda_n}(\mathbf{k}_n^*)$ is the monotone envelope of Λ_n , see Equation (12)) and the choice of \mathbf{k}_n^* , cf. Remark 4.

Algorithm 3 Adaptive sparse grid algorithm for the diffusion problem (1), variation of Guignard–Nobile in [GN18]

- 1: $\Lambda_0 := \{\mathbf{0}\}$
- 2: $u_0 := S_{\Lambda_0} u$
- 3: **for** $n \in \mathbb{N}_0$ **do**
- 4: Compute margin as candidate set $\mathcal{C}_n := \text{Marg}(\Lambda_n)$
- 5: Compute error estimators:

$$\eta_n(\mathbf{k}) := \|\Delta_{\mathbf{k}}(a \nabla u_n)\|_{L_\mu^p(\Gamma; L^2(D))}, \quad \mathbf{k} \in \text{Marg}(\Lambda_n) \quad (14)$$

- 6: choose $\mathbf{k}_n^* := \arg \max_{\mathbf{k} \in \mathcal{C}_n} \eta_n(\mathbf{k})$
 - 7: set $\mathcal{M}_n := E_{\Lambda_n}(\mathbf{k}_n^*)$
 - 8: set $\Lambda_{n+1} := \Lambda_n \cup \mathcal{M}_n$
 - 9: compute $u_{n+1} := S_{\Lambda_{n+1}} u$.
 - 10: **end for**
-

Remark 8 (Adaptive algorithm in [GN18]). *The difference between Algorithm 3 and its original version by Guignard and Nobile in [GN18] is that in [GN18] the following profit indicators are introduced instead of the error estimator $\eta_n(\mathbf{k})$ given in (14):*

$$\pi_n(\mathbf{k}) := \frac{\sum_{\mathbf{i} \in E_{\Lambda_n}(\mathbf{k})} \eta_n(\mathbf{i})}{\sum_{\mathbf{i} \in E_{\Lambda_n}(\mathbf{k})} W(\mathbf{i})}, \quad \mathbf{k} \in \text{Marg}(\Lambda_n), \quad (15)$$

with $W(\mathbf{i})$ denoting the work contribution of the multi-index \mathbf{i} , i.e., the number of new grid points in \mathcal{Y}_i^+ required to evaluate Δ_i which is given by

$$W(\mathbf{i}) := |\mathcal{Y}_i^+| = \prod_{m=1}^M (\mathbf{m}(i_m) - \mathbf{m}(i_m - 1)).$$

Then, \mathbf{k}_n^* is chosen as

$$\mathbf{k}_n^* := \arg \max_{\mathbf{k} \in \mathcal{C}_n} \pi_n(\mathbf{k}), \quad \mathcal{M}_n := E_{\Lambda_n}(\mathbf{k}_n^*). \quad (16)$$

In the case of linearly growing univariate node sets $\mathbf{m}(i) = i$ we have $W(i) \equiv 1$, i.e., $\pi_n(\mathbf{k}) = \frac{1}{|E_{\Lambda_n}(\mathbf{k})|} \sum_{i \in E_{\Lambda_n}(\mathbf{k})} \eta_n(i)$ corresponds to the average error estimator on the monotone envelope $E_{\Lambda_n}(\mathbf{k})$. We provide a more detailed discussion on both versions of the adaptive algorithm for the elliptic problem in Section 4.2 with a focus on computational aspects.

We now turn to our main result stating the convergence of Algorithm 3, under rather mild assumptions on the employed univariate interpolation nodes. Specifically, we assume an algebraic growth of the operator norm of the associated detail operators

$$\|\Delta_k\|_\infty := \sup_{0 \neq f \in C(\Gamma; \mathbb{R})} \frac{\|\Delta_k f\|_{C(\Gamma; \mathbb{R})}}{\|f\|_{C(\Gamma; \mathbb{R})}}, \quad k \in \mathbb{N}_0. \quad (17)$$

Theorem 9 (Convergence of Algorithm 3). *Given the assumptions of Theorem 1 and assuming there exist finite constants $0 \leq c, \theta < \infty$ such that*

$$\|\Delta_k\|_\infty \leq (1 + ck)^\theta \quad \forall k \in \mathbb{N}_0, \quad (18)$$

the approximations u_n constructed by Algorithm 3 satisfy

$$\lim_{n \rightarrow \infty} \|u - u_n\|_{L_\mu^p(\Gamma; H_0^1(D))} = 0.$$

We already established above that Algorithm 3 satisfies the first and third assumption of the abstract convergence theorem, i.e. Theorem 3. It thus remains to verify the second assumption. This turns out to be rather technical and is presented in detail in Section 5.

We now comment on the additional assumption (18) of Theorem 9 regarding the operator norms $\|\Delta_k\|_\infty$ of the univariate detail operators. Condition (18) is rather mild and satisfied, e.g., if the corresponding interpolation operators \mathcal{I}_k possess an at most algebraically increasing Lebesgue constant:

$$\|\mathcal{I}_k\|_\infty := \sup_{f: \|f\|_{C(\Gamma; \mathbb{R})} = 1} \|\mathcal{I}_k f\|_{C(\Gamma; \mathbb{R})} \leq c_1 + c_2 n^\theta \quad \forall k \geq 1, \quad (19)$$

for finite constants $0 \leq c_1, c_2, \theta < \infty$, since then with a finite $c = c(c_1, c_2, \theta) < \infty$

$$\|\Delta_k\|_\infty \leq \|\mathcal{I}_k\|_\infty + \|\mathcal{I}_{k-1}\|_\infty \leq 2c_1 + 2c_2 k^\theta \leq ck^\theta \quad \forall k \geq 1,$$

and $\Delta_0 = \mathcal{I}_0$, i.e., $\|\Delta_0\|_\infty = \|\mathcal{I}_0\|_\infty = 1$. Note that the algebraic growth bound (19) holds, for instance, for interpolation based on Leja and R-Leja nodes $y_{(j)} \in [-1, 1]$ introduced above, see [Chk13, Chk15] and references therein, where such bounds were proved for Leja and R-Leja nodes, respectively:

$$\|\mathcal{I}_k\|_\infty \leq 5k^2 \log k, \text{ for } k \geq 2, \quad \|\mathcal{I}_k\|_\infty \leq 2k, \text{ for } k \geq 1.$$

Moreover, for Clenshaw–Curtis nodes combined with the doubling rule $\mathbf{m}(k) = 2^k$, $k \geq 1$, we obtain by classical results [MP73, Bru78] that

$$\|\mathcal{I}_k\|_\infty \leq 1 + \frac{2}{\pi} \log(\mathbf{m}(k)) = 1 + \frac{2 \log 2}{\pi} k, \quad k \geq 1.$$

4.1 Extensions of Theorem 9

In this subsection we comment on two possible extensions of our convergence analysis.

Convergence of the adaptive algorithm by Guignard and Nobile in [GN18] As outlined in Remark 8, the adaptive algorithm proposed by Guignard and Nobile in [GN18] differs from Algorithm 3 only in the marking strategy or, to be more precise, by the choice of \mathbf{k}_n^* , see (16). Thus, in order to extend Theorem 9 to this algorithm it suffices to verify that the third assumption of Theorem 3 also holds for the marking strategy (16) w.r.t. to the error estimators η_n given in (14). We focus on the case of Leja nodes with a linear growth function $\mathbf{m}(i) \equiv i$ here, since the version with Clenshaw–Curtis nodes was analyzed in the recent work on convergence [FS20] mentioned in the introduction. If Leja points are considered, we can easily ensure convergence by a mild additional assumption: there exists a constant $0 < c < \infty$ such that for any monotone multi-index set Λ we have

$$\max_{\mathbf{k} \in \text{Marg}(\Lambda)} \eta_\Lambda(\mathbf{k}) \leq c \max_{\mathbf{k} \in \mathbf{R}(\Lambda)} \eta_\Lambda(\mathbf{k}), \quad \eta_\Lambda(\mathbf{k}) := \|\Delta_k(a \nabla S_\Lambda u)\|_{L_\mu^p(\Gamma; L^2(D))}, \quad (20)$$

i.e., the largest error estimator in the *full margin* can be bounded by the constant times the largest error estimator in the *reduced margin*. Indeed, by construction of the profits π_n in (15) and of the marking strategy in (16) we have for $\mathbf{m}(i) \equiv i$ that $\pi_n(\mathbf{k}) = \eta_n(\mathbf{k})$ if $\mathbf{k} \in \mathbf{R}(\Lambda_n)$ and

$$\max_{\mathbf{k} \in \mathbf{R}(\Lambda_n)} \eta_n(\mathbf{k}) = \max_{\mathbf{k} \in \mathbf{R}(\Lambda_n)} \pi_n(\mathbf{k}) \leq \frac{\sum_{i \in \mathcal{M}_n} \eta_n(i)}{\sum_{i \in \mathcal{M}_n} W(i)} \leq \sum_{i \in \mathcal{M}_n} \eta_n(i).$$

Hence, condition (20) then guarantees that the third assumption of Theorem 3 is also satisfied for the marking strategy (16). We consider (20) as a plausible assumption in practice, although pathological counterexamples may possibly be constructed.

Convergence of the Gerstner–Griebel algorithm The abstract convergence result, Theorem 3, as well as our techniques for proving Theorem 9 can also be exploited to show convergence of the adaptive algorithm by Gerstner and Griebel in [GG03], i.e. of Algorithm 2. To this end, we need of course to assume the reliability of the error indicators $\eta_n(\mathbf{k}) = \|\Delta_{\mathbf{k}} u\|_{L_{\mu}^p(\Gamma; \mathcal{H})}$. Since these hierarchical surpluses are not connected to the model problem (1), as is the case for the residual-based error estimators (14), we state the Theorem in a more general setting, i.e., we consider general Hilbert space-valued mappings $u: \Gamma \rightarrow \mathcal{H}$ and moreover, we do not restrict to solutions u that admit a Taylor expansion, but rather consider the more general case of a solution that admits an expansion over polynomials P_k with a certain growth of their maximum norm. Reliability is also not proved here but merely assumed, and must be checked on a case-by-case basis.

Theorem 10 (Convergence of Algorithm 2 by Gerstner and Griebel, [GG03]). *Let \mathcal{H} be a separable Hilbert space and let $u \in C(\Gamma; \mathcal{H})$ allow for a polynomial expansion (6) converging in $L_{\mu}^p(\Gamma; \mathcal{H})$ for a $p \in [1, \infty]$ where the corresponding univariate polynomials $P_k \in \mathcal{P}_k(\Gamma; \mathbb{R})$ satisfy*

$$\|P_k\|_{C(\Gamma; \mathbb{R})} \leq (1 + \tilde{c}k)^{\tilde{\theta}} \quad (21)$$

for finite constants $\tilde{c}, \tilde{\theta} \geq 0$. Further assume that

1. the coefficients $u_{\mathbf{k}} \in \mathcal{H}$, $\mathbf{k} \in \mathcal{F}$, of the polynomial expansion (6) satisfy

$$(\rho^{\mathbf{k}} \|u_{\mathbf{k}}\|_{\mathcal{H}})_{\mathbf{k} \in \mathcal{F}} \in \ell^2(\mathcal{F})$$

for a weight vector $\rho \in \mathbb{R}^M$ with $1 < \rho_m$ for all $m = 1, \dots, M$;

2. there exists a constant $C < \infty$ such that for any finite and monotone $\Lambda \subset \mathcal{F}$

$$\|u - S_{\Lambda} u\|_{L_{\mu}^p(\Gamma; \mathcal{H})} \leq C \sum_{\mathbf{k} \in \mathbf{R}(\Lambda)} \|\Delta_{\mathbf{k}} u\|_{L_{\mu}^p(\Gamma; \mathcal{H})}; \quad (22)$$

3. the univariate detail operators Δ_k satisfy (18) for finite constants $0 \leq c, \theta < \infty$.

Then we have for the approximations u_n constructed by Algorithm 2 that

$$\lim_{n \rightarrow \infty} \|u - u_n\|_{L_{\mu}^p(\Gamma; \mathcal{H})} = 0.$$

Note that the first item on the $u_{\mathbf{k}}$ is satisfied for the model problem by Theorem 1 and that for Taylor polynomials condition (21) holds with $\tilde{c} = \tilde{\theta} = 0$. This theorem provides an overview of the three most important "ingredients" for convergence of adaptive collocation: exponentially decaying coefficients $u_{\mathbf{k}}$, only algebraically growing norms of the $\Delta_{\mathbf{k}}$ and reliability of the employed error indicators. The proof of Theorem 10 is significantly easier than the proof of Theorem 9, because the error indicators do not depend on the current approximation. Nonetheless, proving Theorem 10 requires some auxiliary results stated in Section 5 and is therefore postponed to Section 5.2.

4.2 Computational Considerations

Having established the convergence of our variant of the algorithm by Guignard and Nobile, as stated in Algorithm 3, as well as of the Gerstner–Griebel adaptive sparse grid algorithm Algorithm 2 (GG algorithm for short in the following), we comment on the computational advantages and disadvantages of both:

1. The GG algorithm considers candidate indices in the *reduced margin* instead of the *full margin*. This makes treating problems with high-dimensional parameters somewhat easier with the GG algorithm, since the size of the full margin grows substantially faster than the reduced margin.

2. However, as already noted, the GG algorithm is *fully a posteriori*: evaluating the error indicators involves actually evaluating u (i.e., solving additional PDEs) on the new collocation points $\mathcal{Y}_n^+(\mathbf{k}) = \mathcal{Y}_k \setminus \mathcal{Y}_{\Lambda_n \cup \{\mathbf{k}\}}$ for each $\mathbf{k} \in \mathbf{R}(\Lambda_n)$, see (13) Algorithm 2. By contrast, Algorithm 3 computes its error estimator by evaluating *the current sparse grid interpolant* u_n at the new collocation points $\mathcal{Y}_n^+(\mathbf{k})$ for $\mathbf{k} \in \text{Marg}(\Lambda_n)$. This is a significant advantage of the error estimator-based algorithms (both the original version by Guignard and Nobile and our variant Algorithm 3) over the GG algorithm, in particular if solving the PDE for individual parameter values is computationally expensive (even though these additional PDE solves are not discarded but ultimately enter the final approximation returned by Algorithm 2, as already discussed in Section 3).
3. On the other hand, because the error estimators are based on the current approximation, they have to be recomputed in each step of Algorithm 3, i.e., in general $\eta_n(\mathbf{k}) \neq \eta_{n+1}(\mathbf{k})$ for any $\mathbf{k} \in \text{Marg}(\Lambda_n) \cap \text{Marg}(\Lambda_{n+1})$. This is not required by the GG algorithm. Thus, the evaluation of the sparse grid interpolant u_n should be implemented in a very efficient way, since this operation is repeated at each iteration for an increasingly large number of multi-indices in the margin. In this sense, the hierarchical representation of the sparse grid interpolant via hierarchical Lagrange polynomials and hierarchical surpluses is to be preferred to the classical combination technique representation [GSZ92], since the former usually yields a faster evaluation—at the price of a higher offline-cost due to the computation of the surpluses.
4. The hierarchical sparse grid representation as well as the error estimators in [GN18] for the diffusion problem require nested univariate node sets—for an efficient implementation and reliability, respectively. Instead, the GG algorithm also works with non-nested nodes, see e.g. [NTTT16, EST18, EST19]. This might be a rather minor point, since suitable nested node families in form of Leja or Clenshaw-Curtis nodes are available.

As an extensive numerical study of the error estimator-based adaptive scheme has been already carried out by Guignard and Nobile in [GN18], we present no further numerical experiments here. In their study, they observed for several numerical test examples of the diffusion problem (1) that the error estimator stated in Proposition 7 is sharp. These test examples included different dimensions of the physical domain ($d = 1, 2$) as well as different numbers M of parameter variables and different expansion functions a_m in the definition of the diffusion coefficient. Besides this, a second set of experiments in [GN18] compared the performance of the error estimator-based algorithm and the GG algorithm: both showed a similar performance w.r.t. the number of grid points in the corresponding adaptively constructed sparse grids \mathcal{Y}_{Λ_n} (recall that each sparse grid point corresponds to a PDE solve); however, if all PDE solves (i.e., also those necessary for evaluating the profits on the margin) are taken into account, than the GG algorithm performed significantly less effectively.

Although the algorithm by Guignard and Nobile in [GN18] slightly differs from Algorithm 3 as considered here, these differences are negligible for the numerical performance for the following reasons:

- The version of Algorithm 3 considered in [GN18] considers normalized profit indicators π_n for the indices \mathbf{k} , see (15). However, previous numerical evidence for the GG algorithm suggests that whether error indicators or profit indicators are used does not play a major role for the convergence, see e.g. [NTTT16]. Therefore, for the same reasons, one can expect Algorithm 3 to exhibit similar numerical behavior as the original adaptive algorithm by Guignard and Nobile in [GN18].
- Although the second set of results in [GN18] is for Clenshaw–Curtis collocation points only, it is well-known that in practice the performance of Leja and Clenshaw–Curtis points is quite similar for adaptive sparse collocation using the GG algorithm, see e.g. [NTT15]. Thus, it is again reasonable to assume that similar results to those reported in [GN18] also hold for Algorithm 3 using Leja nodes.
- The tests in [GN18] are performed with $p = \infty$ only, both for the evaluation of the error and for the computation of the error indicator. Our theory covers any $p \in [1, \infty]$, and we expect that GG and Algorithm 3 would behave similarly also for $p \neq \infty$.

5 Proofs of Theorems 9 and 10

We begin this section by stating four auxiliary results required for the subsequent proof of our main results, Theorems 9 and 10. First, we recall a statement on the operator norm of the tensorized

detail operators Δ_i given in (17).

Proposition 11 ([CCS14, Section 3]). *For the operator norm (17) of the tensorized detail operators*

$$\|\Delta_i\|_\infty = \sup_{0 \neq f \in C(\mathbf{r}; \mathbb{R})} \frac{\|\Delta_i f\|_{C(\mathbf{r}; \mathbb{R})}}{\|f\|_{C(\mathbf{r}; \mathbb{R})}}, \quad i \in \mathcal{F},$$

there holds

$$\|\Delta_i\|_\infty = \prod_{m=1}^M \|\Delta_{i_m}\|_\infty.$$

Next, we provide an estimate for the sparse grid collocation operator S_Λ applied to Taylor polynomials/multivariate monomials given an algebraically growing operator norm of the univariate detail operators. This result is similar to [EST18, Proposition 3.1].

Proposition 12. *Let there exist constants $1 < c < \infty$ and $\theta < \infty$ such that*

$$\|\Delta_i\|_\infty \leq (1 + ci)^\theta, \quad \forall i \in \mathbb{N}.$$

Then for the Taylor polynomials $T_{\mathbf{k}}(\mathbf{y}) := \mathbf{y}^{\mathbf{k}}$, $\mathbf{k} \in \mathcal{F}$, and $\mathbf{r} = [-1, 1]^M$ we have

$$\sup_{\Lambda \subseteq \mathcal{F}} \|S_\Lambda T_{\mathbf{k}}\|_{C(\mathbf{r}; \mathbb{R})} \leq \prod_{m=1}^M (1 + ck_m)^{1+\theta}, \quad \mathbf{k} \in \mathcal{F}.$$

Proof. First, notice that with \mathbf{m}^{-1} as in (8) and using (10) we have

$$\Delta_i T_{\mathbf{k}} = \prod_{m=1}^M \Delta_{i_m} T_{k_m} \equiv 0$$

if i_m is such that $\mathbf{m}(i_m - 1) \geq k_m$, i.e., if $i_m > \mathbf{m}^{-1}(k_m)$ for any m . Thus, with $\mathcal{R}_{\mathbf{k}} := \{\mathbf{j} \in \mathcal{F} : j_m \leq k_m \forall m = 1, \dots, M\}$, we obtain

$$\sup_{\Lambda \subseteq \mathcal{F}} \|S_\Lambda T_{\mathbf{k}}\|_{C(\mathbf{r}; \mathbb{R})} = \max_{\Lambda \subseteq \mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}} \|S_\Lambda T_{\mathbf{k}}\|_{C(\mathbf{r}; \mathbb{R})},$$

where $\mathbf{m}^{-1}(\mathbf{k}) = (\mathbf{m}^{-1}(k_1), \dots, \mathbf{m}^{-1}(k_M)) \in \mathbb{N}_0^M$. Moreover, the triangle inequality yields

$$\|S_\Lambda T_{\mathbf{k}}\|_{C(\mathbf{r}; \mathbb{R})} \leq \sum_{i \in \Lambda} \|\Delta_i T_{\mathbf{k}}\|_{C(\mathbf{r}; \mathbb{R})} \leq \sum_{i \in \Lambda} \|\Delta_i\|_\infty \|T_{\mathbf{k}}\|_{C(\mathbf{r}; \mathbb{R})} \leq \sum_{i \in \Lambda} \prod_{m=1}^M (1 + ci_m)^\theta.$$

Since we are considering Λ to be a subset of $\mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}$, we can further bound the last term as follows

$$\sum_{i \in \Lambda} \prod_{m=1}^M (1 + ci_m)^\theta \leq \sum_{i \in \mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}} \prod_{m=1}^M (1 + ck_m)^\theta \leq |\mathcal{R}_{\mathbf{k}}| \prod_{m=1}^M (1 + ck_m)^\theta = \prod_{m=1}^M (1 + ck_m)^{1+\theta},$$

since $|\mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}| \leq |\mathcal{R}_{\mathbf{k}}| = \prod_{m=1}^M (1 + k_m)$. □

Furthermore, we require a rather general result on the summability of sequences on \mathcal{F} .

Lemma 13 ([CM18, Lemmas 2 and 3]). *For any $0 < q < 1$, one has*

$$\boldsymbol{\rho} \in \mathbb{R}^M \text{ and } \min_{m=1, \dots, M} |\rho_m| > 1 \iff (\boldsymbol{\rho}^{-\mathbf{k}})_{\mathbf{k} \in \mathcal{F}} \in \ell^q(\mathcal{F}).$$

Moreover, for any $0 < q < 1$ and any algebraic factor

$$\beta(\mathbf{k}) := \prod_{m=1}^M (1 + ck_m)^\theta, \quad \mathbf{k} \in \mathcal{F},$$

with finite $c, \theta \geq 0$, one has

$$\boldsymbol{\rho} \in \mathbb{R}^M \text{ and } \min_{m=1, \dots, M} |\rho_m| > 1 \iff (\beta(\mathbf{k}) \boldsymbol{\rho}^{-\mathbf{k}})_{\mathbf{k} \in \mathcal{F}} \in \ell^q(\mathcal{F}).$$

Note that the original statement in [CM18, Lemmas 2 and 3] is for the case of countable sequences $\boldsymbol{\rho} = (\rho_m)_{m \in \mathbb{N}} \in \ell^q(\mathbb{N})$.

The last auxiliary result provides a simple estimate for the tails of converging series of the same form $(\beta(\mathbf{k}) \boldsymbol{\rho}^{-\mathbf{k}})_{\mathbf{k} \in \mathcal{F}}$ as considered in the previous lemma.

Proposition 14. *Let $\boldsymbol{\rho} \in \mathbb{R}^M$ be a vector of numbers $\rho_m > 1$, $m = 1, \dots, M$, and*

$$\beta(\mathbf{k}) := \prod_{m=1}^M (1 + ck_m)^\theta, \quad \mathbf{k} \in \mathcal{F},$$

an algebraic factor with finite $c, \theta \geq 0$. Then, we have for any $\mathbf{k} \in \mathcal{F}$

$$\sum_{j \geq \mathbf{k}} \beta(\mathbf{j}) \boldsymbol{\rho}^{-\mathbf{j}} \leq C \beta(\mathbf{k}) \boldsymbol{\rho}^{-\mathbf{k}}, \quad C := \sum_{\mathbf{k} \in \mathcal{F}} \beta(\mathbf{k}) \boldsymbol{\rho}^{-\mathbf{k}} < \infty \quad (23)$$

Proof. First, note that by Lemma 13 the constant C defined in (23) is indeed finite. By refactoring, we have

$$\sum_{j \geq \mathbf{k}} \beta(\mathbf{j}) \boldsymbol{\rho}^{-\mathbf{j}} = \sum_{j \geq \mathbf{k}} \prod_{m=1}^M (1 + cj_m)^\theta \rho_m^{-j_m} = \prod_{m=1}^M \left(\sum_{j_m \geq k_m} (1 + cj_m)^\theta \rho_m^{-j_m} \right).$$

We then obtain for each $m = 1, \dots, M$,

$$\begin{aligned} \sum_{j_m \geq k_m} (1 + cj_m)^\theta \rho_m^{-j_m} &= (1 + ck_m)^\theta \rho_m^{-k_m} \sum_{j=0}^{\infty} \left(\frac{1 + cj + ck_m}{1 + ck_m} \right)^\theta \rho_m^{-j} \\ &\leq (1 + ck_m)^\theta \rho_m^{-k_m} \sum_{j=0}^{\infty} (1 + cj)^\theta \rho_m^{-j}. \end{aligned}$$

Thus, the refactoring argument can be continued as

$$\begin{aligned} \sum_{j \geq \mathbf{k}} \beta(\mathbf{j}) \boldsymbol{\rho}^{-\mathbf{j}} &= \sum_{j \geq \mathbf{k}} \prod_{m=1}^M (1 + cj_m)^\theta \rho_m^{-j_m} \\ &\leq \prod_{m=1}^M \left((1 + ck_m)^\theta \rho_m^{-k_m} \sum_{j_m \geq 0} (1 + cj_m)^\theta \rho_m^{-j_m} \right) \\ &= \beta(\mathbf{k}) \boldsymbol{\rho}^{-\mathbf{k}} \sum_{j \geq 0} \prod_{m=1}^M (1 + cj_m)^\theta \rho_m^{-j_m} = C \beta(\mathbf{k}) \boldsymbol{\rho}^{-\mathbf{k}}, \end{aligned}$$

with C as in Equation (23). □

5.1 Proof of Theorem 9

Proof. We prove Theorem 9 by applying Theorem 3. To this end, we need to verify the three assumptions of Theorem 3. The first holds due to Proposition 7 and the third by construction, cf. Remark 4. Hence, it remains to verify the second assumption. To this end, we set

$$\widehat{\eta}_n(\mathbf{k}) := \begin{cases} \|\Delta_{\mathbf{k}}(a \nabla S_{\Lambda_n} u)\|_{L_\mu^p(\Gamma; L^2(D))}, & \mathbf{k} \in \Lambda_n \cup \mathcal{C}_n \\ 0, & \text{otherwise,} \end{cases} \quad (24)$$

and proceed in two steps (see also Remark 4):

1. We define the (formal) limit

$$u_\infty := \sum_{\mathbf{k} \in \Lambda_\infty} \Delta_{\mathbf{k}} u, \quad \Lambda_\infty := \bigcup_{n \in \mathbb{N}} \Lambda_n, \quad (25)$$

and verify in Lemma 15 below that $u_\infty \in C(\Gamma; H_0^1(D))$ as well as

$$\lim_{n \rightarrow \infty} \|u_\infty - u_n\|_{C(\Gamma; H_0^1(D))} = 0.$$

2. We then set

$$\eta_\infty(\mathbf{k}) := \begin{cases} \|\Delta_{\mathbf{k}}(a\nabla u_\infty)\|_{L^p_\mu(\Gamma; L^2(D))}, & \mathbf{k} \in \Lambda_\infty \cup \text{Marg}(\Lambda_\infty), \\ 0, & \text{otherwise,} \end{cases} \quad (26)$$

and show in Lemma 17 that

$$\lim_{n \rightarrow \infty} \|\eta_\infty - \hat{\eta}_n\|_{\ell^1} = 0,$$

which concludes the proof. \square

Lemma 15. *Given the assumptions of Theorem 9, the u_n , $n \in \mathbb{N}$ form a Cauchy sequence in $C(\Gamma; H_0^1(D))$. In particular, u_∞ given in (25) is its well-defined limit in $C(\Gamma; H_0^1(D))$.*

Proof. We abbreviate the norms in $C(\Gamma; H_0^1(D))$ and $C(\Gamma; \mathbb{R})$ by $\|\cdot\|_C$. Furthermore, let $\boldsymbol{\rho} \in \mathbb{R}^M$ be such that $1 < \rho_m < \alpha^{-1}$ as in equation (5) and let $T_{\mathbf{k}}$ and $t_{\mathbf{k}}$, $\mathbf{k} \in \mathcal{F}$, denote the multivariate Taylor polynomials and the corresponding Taylor coefficients of u , respectively. For $n, m \in \mathbb{N}$ with $n \leq m$ we obtain by the triangle and Cauchy–Schwarz inequalities

$$\begin{aligned} \|u_m - u_n\|_C &= \|S_{\Lambda_m \setminus \Lambda_n} u\|_C = \left\| \sum_{\mathbf{k} \in \mathcal{F}} t_{\mathbf{k}} S_{\Lambda_m \setminus \Lambda_n} T_{\mathbf{k}} \right\|_C \leq \sum_{\mathbf{k} \in \mathcal{F}} \|t_{\mathbf{k}}\|_{\mathcal{H}} \|S_{\Lambda_m \setminus \Lambda_n} T_{\mathbf{k}}\|_C \\ &\leq \left(\sum_{\mathbf{k} \in \mathcal{F}} \rho^{2\mathbf{k}} \|t_{\mathbf{k}}\|_{\mathcal{H}}^2 \right)^{1/2} \left(\sum_{\mathbf{k} \in \mathcal{F}} \rho^{-2\mathbf{k}} \|S_{\Lambda_m \setminus \Lambda_n} T_{\mathbf{k}}\|_C^2 \right)^{1/2}, \end{aligned}$$

where by Theorem 1

$$C_{u, \boldsymbol{\rho}} := \left(\sum_{\mathbf{k} \in \mathcal{F}} \rho^{2\mathbf{k}} \|t_{\mathbf{k}}\|_{\mathcal{H}}^2 \right)^{1/2} < \infty. \quad (27)$$

Since $\Delta_{\mathbf{i}} T_{\mathbf{k}} = 0$ if $i_m > \mathbf{m}^{-1}(k_m)$ for any m we have by Proposition 11 and the assumptions that

$$\begin{aligned} \|S_{\Lambda_m \setminus \Lambda_n} T_{\mathbf{k}}\|_C &\leq \sum_{\mathbf{i} \in \Lambda_m \setminus \Lambda_n} \|\Delta_{\mathbf{i}} T_{\mathbf{k}}\|_C \leq \sum_{\mathbf{i} \in \Lambda_\infty \setminus \Lambda_n} \|\Delta_{\mathbf{i}} T_{\mathbf{k}}\|_C \\ &= \sum_{\mathbf{i} \in (\Lambda_\infty \setminus \Lambda_n) \cap \mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}} \|\Delta_{\mathbf{i}} T_{\mathbf{k}}\|_C \\ &\leq g_n(\mathbf{k}) := \sum_{\mathbf{i} \in (\Lambda_\infty \setminus \Lambda_n) \cap \mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}} \prod_{m=1}^M (1 + ck_m)^\theta, \end{aligned}$$

where $\mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})} = \{\mathbf{i} \in \mathcal{F} : \mathbf{i} \leq \mathbf{m}^{-1}(\mathbf{k})\}$. Since for any of the finitely many $\mathbf{i} \in (\Lambda_\infty \setminus \Lambda_n) \cap \mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}$ there exists an $n_0 \in \mathbb{N}$ such that $\mathbf{i} \in \Lambda_n$ for all $n \geq n_0$, we obtain

$$\lim_{n \rightarrow \infty} g_n(\mathbf{k}) = \lim_{n \rightarrow \infty} g_n^2(\mathbf{k}) = 0 \quad \forall \mathbf{k} \in \mathcal{F}.$$

Moreover, we conclude as in the proof of Proposition 12

$$g_n(\mathbf{k}) \leq \sum_{\mathbf{i} \in \mathcal{R}_{\mathbf{m}^{-1}(\mathbf{k})}} \prod_{m=1}^M (1 + ck_m)^\theta \leq g(\mathbf{k}) := \prod_{m=1}^M (1 + ck_m)^{1+\theta}.$$

By Lemma 13 we have

$$\sum_{\mathbf{k} \in \mathcal{F}} \rho^{-2\mathbf{k}} g(\mathbf{k})^2 < \infty,$$

so that $g^2 : \mathcal{F} \rightarrow [0, \infty)$ serves as a summable dominating mapping of the $g_n^2 : \mathcal{F} \rightarrow [0, \infty)$ and we obtain by Lebesgue's dominated convergence theorem

$$\lim_{n \rightarrow \infty} \sum_{\mathbf{k} \in \mathcal{F}} \rho^{-2\mathbf{k}} g_n(\mathbf{k})^2 = 0.$$

Thus, since

$$\|u_m - u_n\|_C^2 \leq C_{u, \boldsymbol{\rho}}^2 \sum_{\mathbf{k} \in \mathcal{F}} \rho^{-2\mathbf{k}} g_n(\mathbf{k})^2 \quad \forall m \geq n,$$

we conclude that the approximations $u_n = \sum_{\mathbf{i} \in \Lambda_n} \Delta_{\mathbf{i}} u$ form a Cauchy sequence in the (complete) Banach space $C(\Gamma; H_0^1(D))$ with $u_\infty = \sum_{\mathbf{i} \in \Lambda_\infty} \Delta_{\mathbf{i}} u$ as its limit, since $\Lambda_n \uparrow \Lambda_\infty$. \square

For the second step of the proof of Theorem 9, we first state an important lemma concerning the decay of the error estimators.

Lemma 16. *Let the assumptions of Theorem 9 be satisfied and let $\Lambda \subset \mathcal{F}$ be an arbitrary monotone subset. Then there exists a constant $C = C(M, \rho, c, \theta, a) < \infty$ such that for*

$$\eta(\mathbf{k}, S_\Lambda u) := \|\Delta_{\mathbf{k}}(a \nabla S_\Lambda u)\|_{L_\mu^p(\Gamma; L^2(D))}, \quad \mathbf{k} \in \mathcal{F},$$

we have for any $\mathbf{k} \in \mathcal{F}$

$$\eta(\mathbf{k}, S_\Lambda u) \leq C g(\mathbf{k}), \quad g(\mathbf{k}) := \left(\prod_{m=1}^M (1 + ck_m)^{2\theta+1} \right) \rho^{-\mathbf{k}}.$$

Proof. Set $u_\Lambda := S_\Lambda u$. By linearity $\Delta_{\mathbf{k}}(a \nabla u_\Lambda)$ for $\mathbf{k} \in \mathcal{F}$ can be written as

$$\Delta_{\mathbf{k}}[a \nabla u_\Lambda] = \Delta_{\mathbf{k}} \left[a \sum_{\mathbf{i} \in \Lambda} \Delta_{\mathbf{i}} \nabla u \right] = \sum_{\mathbf{i} \in \Lambda} \Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} \nabla u].$$

Moreover, using the Taylor expansion of the solution u we deduce that

$$\Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} \nabla u] = \Delta_{\mathbf{k}} \left[a \Delta_{\mathbf{i}} \sum_{j \in \mathcal{F}} (\nabla t_j) T_j \right] = \sum_{j \in \mathcal{F}} (\nabla t_j) \Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} T_j]. \quad (28)$$

We observe that for certain combinations of \mathbf{i} , \mathbf{j} , and \mathbf{k} it holds $\Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} T_j] \equiv 0$. First of all,

$$\Delta_{\mathbf{i}} T_j = \prod_{m=1}^M (\Delta_{i_m} T_{j_m}) \equiv 0 \quad \text{if } \exists m: j_m \leq \mathbf{m}(i_m - 1),$$

since then $\Delta_{i_m} T_{j_m} \equiv 0$. Second, the function $a \Delta_{\mathbf{i}} T_j$ is a polynomial in \mathbf{y} belonging to the space

$$\mathcal{P}_{\mathbf{m}(\mathbf{i})+1} := \text{span} \{ \mathbf{y}^p : p_m \leq \mathbf{m}(i_m) + 1 \text{ for } m = 1, \dots, M \},$$

since a is affine in \mathbf{y} . Hence,

$$\Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} T_j] \equiv 0 \quad \text{if } \exists m: \mathbf{m}(i_m) + 1 \leq \mathbf{m}(k_m - 1),$$

We combine now both necessary conditions $\mathbf{j} \geq \mathbf{m}(\mathbf{i} - \mathbf{1}) + \mathbf{1}$ and $\mathbf{m}(\mathbf{i}) + \mathbf{1} \geq \mathbf{m}(\mathbf{k} - \mathbf{1}) + \mathbf{1}$ for $\Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} T_j] \neq 0$ to

$$\mathbf{j} \geq \mathbf{m}(\mathbf{k} - \mathbf{2}) + \mathbf{1} \geq \mathbf{k} - \mathbf{1},$$

where the last inequality follows due to $\mathbf{m}(k) \geq k$. Thus, introducing the notation $[\mathbf{k} - \mathbf{1}]_+ := (\max\{k_m - 1, 0\})_{m=1}^M$, the sum (28) reduces to

$$\Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} u] = \sum_{j \geq [\mathbf{k} - \mathbf{1}]_+} (\nabla t_j) \Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} T_j].$$

By interchanging the order of summation we obtain

$$\begin{aligned} \|\Delta_{\mathbf{k}}(a \nabla u_\Lambda)\|_{L_\mu^p(\Gamma; L^2(D))} &= \left\| \sum_{\mathbf{i} \in \Lambda} \Delta_{\mathbf{k}}(a \Delta_{\mathbf{i}} \nabla u_\Lambda) \right\|_{L_\mu^p(\Gamma; L^2(D))} \\ &= \left\| \sum_{\mathbf{i} \in \Lambda} \sum_{j \geq [\mathbf{k} - \mathbf{1}]_+} (\nabla t_j) \Delta_{\mathbf{k}}[a \Delta_{\mathbf{i}} T_j] \right\|_{L_\mu^p(\Gamma; L^2(D))} \\ &= \left\| \sum_{j \geq [\mathbf{k} - \mathbf{1}]_+} (\nabla t_j) \Delta_{\mathbf{k}}[a S_\Lambda T_j] \right\|_{L_\mu^p(\Gamma; L^2(D))}. \end{aligned}$$

We now set $\beta(\mathbf{k}) := \prod_{m=1}^M (1 + ck_m)^\theta$ as well as

$$a_{\max} := \sup_{\mathbf{y} \in \Gamma} \sup_{\mathbf{x} \in D} |a(\mathbf{x}, \mathbf{y})| < \infty. \quad (29)$$

By using the triangle inequality, Proposition 11 and Proposition 12 we deduce

$$\begin{aligned}
\|\Delta_{\mathbf{k}}(a\nabla u_\Lambda)\|_{L_\mu^p(\Gamma;L^2(D))} &= \left\| \sum_{j \geq [\mathbf{k}-1]_+} (\nabla t_j) \Delta_{\mathbf{k}} [a S_\Lambda T_j] \right\|_{L_\mu^p(\Gamma;L^2(D))} \\
&\leq \sum_{j \geq [\mathbf{k}-1]_+} \|(\nabla t_j)\|_{L^2(D)} \|\Delta_{\mathbf{k}} [a S_\Lambda T_j]\|_{C(\Gamma;\mathbb{R})} \\
&\leq \sum_{j \geq [\mathbf{k}-1]_+} \|t_j\|_{\mathcal{H}} \beta(\mathbf{k}) \|a S_\Lambda T_j\|_{C(\Gamma;\mathbb{R})} \\
&\leq \sum_{j \geq [\mathbf{k}-1]_+} \|t_j\|_{\mathcal{H}} \beta(\mathbf{k}) a_{\max} \|S_\Lambda T_j\|_{C(\Gamma;\mathbb{R})} \\
&\leq a_{\max} \beta(\mathbf{k}) \sum_{j \geq [\mathbf{k}-1]_+} \|t_j\|_{\mathcal{H}} \gamma(j),
\end{aligned}$$

where we set $\gamma(j) := \prod_{m=1}^M (1 + c j_m)^{1+\theta}$. By the Cauchy–Schwarz inequality we obtain

$$\sum_{j \geq [\mathbf{k}-1]_+} \|t_j\|_{\mathcal{H}} \gamma(j) \leq C_{u,\rho} \left(\sum_{j \geq [\mathbf{k}-1]_+} \rho^{-2j} \gamma(j)^2 \right)^{1/2},$$

with ρ as in Theorem 1 and $C_{u,\rho}$ as in (27). We can then apply Proposition 14 to bound $\sum_{j \geq [\mathbf{k}-1]_+} \rho^{-2j} \gamma(j)^2$. More specifically, Proposition 14 yields the existence of a constant $C_{\rho,c,\theta} < \infty$ such that it holds

$$\sum_{j \geq [\mathbf{k}-1]_+} \rho^{-2j} \gamma(j)^2 \leq C_{\rho,c,\theta} \rho^{-2[\mathbf{k}-1]_+} \gamma([\mathbf{k}-1]_+)^2 \leq C_{\rho,c,\theta} \left(\prod_{m=1}^M \rho_m^2 \right) \rho^{-2\mathbf{k}} \gamma(\mathbf{k})^2,$$

since γ is increasing and $\rho_m > 1$ for each m . Thus, for any $\mathbf{k} \in \mathcal{F}$ we get

$$\|\Delta_{\mathbf{k}}(a\nabla u_\Lambda)\|_{L_\mu^p(\Gamma;L^2(D))} \leq a_{\max} C_{u,\rho} \beta(\mathbf{k}) C_{\rho,c,\theta}^{1/2} \left(\prod_{m=1}^M \rho_m \right) \gamma(\mathbf{k}) \rho^{-\mathbf{k}}.$$

The statement follows with

$$C := a_{\max} C_{u,\rho} C_{\rho,c,\theta}^{1/2} \left(\prod_{m=1}^M \rho_m \right), \tag{30}$$

since $g(\mathbf{k}) = \beta(\mathbf{k})\gamma(\mathbf{k}) \rho^{-\mathbf{k}}$. \square

This bound of the error indicators is now used to proceed with the second step of the proof to verify the second assumption of Theorem 3.

Lemma 17. *Given the assumptions of Theorem 9 we have for η_∞ as in (26) and $\hat{\eta}_n$ as in (24) that*

$$\lim_{n \rightarrow \infty} \|\eta_\infty - \hat{\eta}_n\|_{\ell^1(\mathcal{F})} = 0.$$

Proof. We introduce the short-hand notation

$$\Lambda^+ := \Lambda \cup \text{Marg}(\Lambda), \quad \Lambda \subseteq \mathcal{F},$$

and notice that consequently $\Lambda_\infty^+ \subseteq \bigcup_{n \in \mathbb{N}} \Lambda_n^+$. Moreover, we have

$$|\eta_\infty(\mathbf{k}) - \hat{\eta}_n(\mathbf{k})| \leq \begin{cases} \|\Delta_{\mathbf{k}}(a\nabla(u_\infty - u_n))\|_{L_\mu^p(\Gamma;L^2(D))}, & \mathbf{k} \in \Lambda_\infty^+ \subset \Lambda_\infty^+, \\ \|\Delta_{\mathbf{k}}(a\nabla u_\infty)\|_{L_\mu^p(\Gamma;L^2(D))}, & \mathbf{k} \in \Lambda_\infty^+ \setminus \Lambda_n^+, \\ 0, & \mathbf{k} \in \mathcal{F} \setminus \Lambda_\infty^+. \end{cases}$$

Hence,

$$\|\eta_\infty - \hat{\eta}_n\|_{\ell^1(\mathcal{F})} \leq \underbrace{\sum_{\mathbf{k} \in \Lambda_\infty^+} \|\Delta_{\mathbf{k}}(a\nabla(u_\infty - u_n))\|_{L_\mu^p(\Gamma;L^2(D))}}_{\text{term I}} + \underbrace{\sum_{\mathbf{k} \in \Lambda_\infty^+ \setminus \Lambda_n^+} \|\Delta_{\mathbf{k}}(a\nabla u_\infty)\|_{L_\mu^p(\Gamma;L^2(D))}}_{\text{term II}}.$$

We would like to take the limit on both sides, and verify that the two terms on the right-hand side tend to zero, which we analyze separately in the following.

Term I Assuming for a moment that we can apply the dominated convergence theorem to exchange the sum and the limit, we would get

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \sum_{\mathbf{k} \in \Lambda_{\infty}^+} \|\Delta_{\mathbf{k}}(a \nabla(u_{\infty} - u_n))\|_{L_{\mu}^p(\Gamma; L^2(D))} \\
&= \sum_{\mathbf{k} \in \Lambda_{\infty}^+} \lim_{n \rightarrow \infty} \|\Delta_{\mathbf{k}}(a \nabla(u_{\infty} - u_n))\|_{L_{\mu}^p(\Gamma; L^2(D))} \quad \text{by dominated convergence} \\
&\leq \sum_{\mathbf{k} \in \Lambda_{\infty}^+} \lim_{n \rightarrow \infty} \beta(\mathbf{k}) \|a \nabla(u_{\infty} - u_n)\|_{C(\Gamma; L^2(D))} \quad \text{by Pr. 11, } \beta(\mathbf{k}) := \prod_{m=1}^M (1 + ck_m)^{\theta} \\
&\leq \sum_{\mathbf{k} \in \Lambda_{\infty}^+} \lim_{n \rightarrow \infty} \beta(\mathbf{k}) a_{\max} \|u_{\infty} - u_n\|_{C(\Gamma; H_0^1(D))} \quad \text{recalling the def. of } a_{\max} \text{ in (29)} \\
&= 0 \quad \text{by Lemma 15.}
\end{aligned}$$

In order to apply Lebesgue's dominated convergence, we need to check that there exists a function $g: \mathcal{F} \rightarrow [0, \infty)$ such that, for all $n \in \mathbb{N}$ and $\mathbf{k} \in \Lambda_{\infty}^+$,

$$\|\Delta_{\mathbf{k}}(a \nabla u_{\infty})\|_{L_{\mu}^p(\Gamma; L^2(D))} + \|\Delta_{\mathbf{k}}(a \nabla u_n)\|_{L_{\mu}^p(\Gamma; L^2(D))} \leq g(\mathbf{k}) \quad \text{and} \quad \sum_{\mathbf{k} \in \Lambda_{\infty}^+} g(\mathbf{k}) < \infty. \quad (31)$$

The bounding function g is obtained by Lemma 16: there exists a constant $C < \infty$ such that

$$\|\Delta_{\mathbf{k}}(a \nabla u_{\infty})\|_{L_{\mu}^p(\Gamma; L^2(D))} + \|\Delta_{\mathbf{k}}(a \nabla u_n)\|_{L_{\mu}^p(\Gamma; L^2(D))} \leq 2C g(\mathbf{k}),$$

with

$$g(\mathbf{k}) := \left(\prod_{m=1}^M (1 + ck_m)^{2\theta+1} \right) \rho^{-\mathbf{k}}.$$

The required summability of g is derived by Lemma 13, i.e.,

$$\sum_{\mathbf{k} \in \Lambda_{\infty}^+} 2C g(\mathbf{k}) \leq 2C \sum_{\mathbf{k} \in \mathcal{F}} \left(\prod_{m=1}^M (1 + ck_m)^{2\theta+1} \right) \rho^{-\mathbf{k}} < \infty.$$

Term II To verify that the limit of the second term is also zero, observe that the dominated convergence theorem in (31) implies

$$\sum_{\mathbf{k} \in \Lambda_{\infty}^+} \|\Delta_{\mathbf{k}}(a \nabla u_{\infty})\|_{L_{\mu}^p(\Gamma; L^2(D))} < \infty$$

Together with the fact that $\Lambda_{\infty}^+ \subseteq \bigcup_{n \in \mathbb{N}} \Lambda_n^+$, this implies the final result

$$\lim_{n \rightarrow \infty} \sum_{\mathbf{k} \in \Lambda_{\infty}^+ \setminus \Lambda_n^+} \|\Delta_{\mathbf{k}}(a \nabla u_{\infty})\|_{L_{\mu}^p(\Gamma; L^2(D))} = 0.$$

□

By Lemma 17, the three assumptions of Theorem 3 have been verified, proving convergence of the described adaptive algorithm.

5.2 Proof of Theorem 10

Proof. Again we prove the assertion by applying Theorem 3, i.e., verifying the three assumptions of Theorem 3. The first holds by assumption and the third by construction of Algorithm 2, cf. Remark 4. Thus, it remains again to verify the second assumption of Theorem 3. We set

$$\Lambda_n^+ := \Lambda_n \cup \mathcal{C}_n = \Lambda_n \cup \mathbf{R}(\Lambda_n)$$

as well as

$$\widehat{\eta}_n(\mathbf{k}) := \begin{cases} \|\Delta_{\mathbf{k}} u\|_{L_{\mu}^p(\Gamma; \mathcal{H})}, & \mathbf{k} \in \Lambda_n^+ \\ 0, & \text{otherwise,} \end{cases} \quad (32)$$

and define

$$\eta_{\infty}(\mathbf{k}) := \begin{cases} \|\Delta_{\mathbf{k}} u\|_{L_{\mu}^p(\Gamma; \mathcal{H})}, & \mathbf{k} \in \Lambda_{\infty}^+, \\ 0, & \text{otherwise,} \end{cases}, \quad \Lambda_{\infty}^+ := \bigcup_{n \in \mathbb{N}} \Lambda_n^+. \quad (33)$$

We verify in Lemma 18 below (which is similar to Lemmas 16 and 17) that

$$\lim_{n \rightarrow \infty} \|\eta_\infty - \hat{\eta}_n\|_{\ell^1} = 0,$$

which concludes the proof. \square

Lemma 18. *Let the assumptions of Theorem 10 be satisfied. Then, there exists a constant $C < \infty$ such that for any $\mathbf{k} \in \mathcal{F}$*

$$\|\Delta_{\mathbf{k}} u\|_{L^p_\mu(\Gamma; \mathcal{H})} \leq C g(\mathbf{k}), \quad g(\mathbf{k}) := \left(\prod_{m=1}^M (1 + \tilde{c}k_m)^{\tilde{\theta}} (1 + ck_m)^\theta \right) \rho^{-\mathbf{k}}. \quad (34)$$

Moreover, we have $(\eta_\infty(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}} \in \ell^1(\mathcal{F})$ for $\eta_\infty(\mathbf{k})$ as given in (33) and, therefore, for $\hat{\eta}_n$ as in (32)

$$\lim_{n \rightarrow \infty} \|\eta_\infty - \hat{\eta}_n\|_{\ell^1} = 0.$$

Proof. In the following we denote the norm in $L^p_\mu(\Gamma; \mathcal{H})$ and $C(\Gamma; \mathcal{H})$ simply by $\|\cdot\|_{L^p}$ and $\|\cdot\|_C$, respectively. By employing the polynomial expansion of u and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} \|\Delta_{\mathbf{k}} u\|_{L^p} &= \left\| \sum_{\mathbf{i} \in \mathcal{F}} u_{\mathbf{i}} \Delta_{\mathbf{k}} P_{\mathbf{i}} \right\|_{L^p} \leq \sum_{\mathbf{i} \in \mathcal{F}} \|u_{\mathbf{i}}\|_{\mathcal{H}} \|\Delta_{\mathbf{k}} P_{\mathbf{i}}\|_{L^p} \\ &\leq \left(\sum_{\mathbf{i} \in \mathcal{F}} \rho^{2\mathbf{i}} \|u_{\mathbf{i}}\|_{\mathcal{H}}^2 \right)^{1/2} \left(\sum_{\mathbf{i} \in \mathcal{F}} \rho^{-2\mathbf{i}} \|\Delta_{\mathbf{k}} P_{\mathbf{i}}\|_{L^p}^2 \right)^{1/2}, \end{aligned}$$

where $\rho \in \mathbb{R}^M$ is as assumed in Theorem 10. By assumption the first term is bounded by a constant

$$C_{u, \rho} := \left(\sum_{\mathbf{i} \in \mathcal{F}} \rho^{2\mathbf{i}} \|u_{\mathbf{i}}\|_{\mathcal{H}}^2 \right)^{1/2} < \infty.$$

Concerning the second term, we first note that

$$\Delta_{\mathbf{k}} P_{\mathbf{i}} = \prod_{m=1}^M \Delta_{k_m} P_{i_m} \equiv 0 \quad \text{if } \exists m: i_m \leq \mathbf{m}(k_m - 1).$$

Hence, we require $\mathbf{i} \geq \mathbf{m}(\mathbf{k} - \mathbf{1}) + \mathbf{1} \geq \mathbf{k}$ for $\Delta_{\mathbf{k}} P_{\mathbf{i}} \neq 0$ and therefore obtain by Proposition 11 and the assumption (21)

$$\begin{aligned} \sum_{\mathbf{k} \in \mathcal{F}} \rho^{-2\mathbf{k}} \|\Delta_{\mathbf{k}} P_{\mathbf{i}}\|_{L^p}^2 &= \sum_{\mathbf{i} \geq \mathbf{k}} \rho^{-2\mathbf{i}} \|\Delta_{\mathbf{k}} P_{\mathbf{i}}\|_{L^p}^2 \leq \sum_{\mathbf{i} \geq \mathbf{k}} \rho^{-2\mathbf{i}} \|\Delta_{\mathbf{k}} P_{\mathbf{i}}\|_C^2 \\ &\leq \sum_{\mathbf{i} \geq \mathbf{k}} \rho^{-2\mathbf{i}} \left(\prod_{m=1}^M (1 + ck_m)^\theta \right) \|P_{\mathbf{i}}\|_{C(\Gamma; \mathbb{R})}^2 \\ &\leq \gamma(\mathbf{k}) \sum_{\mathbf{i} \geq \mathbf{k}} \rho^{-2\mathbf{i}} \beta(\mathbf{i})^2, \end{aligned}$$

with

$$\beta(\mathbf{i}) := \prod_{m=1}^M (1 + \tilde{c}i_m)^{\tilde{\theta}}, \quad \gamma(\mathbf{k}) := \prod_{m=1}^M (1 + ck_m)^\theta.$$

Hence, by Proposition 14 we have for a finite constant C

$$\sum_{\mathbf{i} \geq \mathbf{k}} \rho^{-2\mathbf{i}} \beta(\mathbf{i})^2 \leq C \rho^{-2\mathbf{k}} \beta(\mathbf{k})^2$$

and, thus,

$$\|\Delta_{\mathbf{k}} u\|_{L^p} \leq C_{u, \rho} C^{1/2} \gamma(\mathbf{k}) \beta(\mathbf{k}) \rho^{-\mathbf{k}} \quad \mathbf{k} \in \mathcal{F},$$

which proves (34). Moreover, by Lemma 13 we know that $(g(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}} \in \ell^1(\mathcal{F})$, and hence, also $(\hat{\eta}_n(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}}, (\eta_\infty(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}} \in \ell^1(\mathcal{F})$, $n \in \mathbb{N}$. Finally, we have by definition of η_∞ and $\hat{\eta}_n$ that

$$\|\eta_\infty - \hat{\eta}_n\|_{\ell^1} = \sum_{\mathbf{k} \in \Lambda_\infty^+ \setminus \Lambda_n^+} \|\Delta_{\mathbf{k}} u\|_{L^p} \leq C_{u, \rho} C^{1/2} \sum_{\mathbf{k} \in \Lambda_\infty^+ \setminus \Lambda_n^+} g(\mathbf{k}).$$

The summability $(g(\mathbf{k}))_{\mathbf{k} \in \mathcal{F}} \in \ell^1(\mathcal{F})$ and $\Lambda_\infty^+ = \bigcup_{n \in \mathbb{N}} \Lambda_n^+$ then yield the desired result

$$\lim_{n \rightarrow \infty} \|\eta_\infty - \hat{\eta}_n\|_{\ell^1} \leq \lim_{n \rightarrow \infty} \sum_{\mathbf{k} \in \Lambda_\infty^+ \setminus \Lambda_n^+} g(\mathbf{k}) = 0.$$

□

6 Conclusions

We have proved convergence of an adaptive sparse collocation algorithm for approximating the solution of an elliptic PDE with a high-dimensional parameter $\mathbf{y} \in [-1, 1]^M$, applying the analysis technique from [BPRR19a], developed for the stochastic Galerkin FEM, to a slight variation of the algorithm proposed by Guignard and Nobile in [GN18]. In this sense, our work can be seen as an extension of [GN18], where a very close variant of the algorithm considered here was presented and analyzed numerically, but without convergence proof.

The algorithms we propose here and that in [GN18] are both modifications of the well-known dimension-adaptive sparse grid algorithm of Gerstner and Griebel in that they replace the hierarchical surplus error indicators with a rigorous residual-based error estimator. As a by-product of our analysis we also obtain a convergence proof for the Gerstner–Griebel algorithm applied to the same problem, under the assumption that the hierarchical surplus error indicator is also a reliable error estimator. The convergence proof is tailored to the specific problem, i.e., an elliptic PDE with parametric diffusion coefficient depending affinely on a finite number of parameters. Because the algorithm is based on a residual-based error estimator, the analysis is problem-specific and must be adapted for each new PDE as well as for different forms (e.g. nonlinear) of the random diffusion coefficient. However, we expect that a large part of the machinery proves valid or at least extensible in a straightforward way. Particularly, if reliable error estimators (for the approximation error w.r.t. the parameter variables) are available, only a stability condition of these estimators w.r.t. u_n needs to be established in order to verify the crucial second condition of the general convergence Theorem 3. Our analysis in Section 5.1 can serve as a blueprint for doing so.

Regarding possible extensions of this work, we point out that the convergence analysis we have presented proves convergence but does not provide a rate. This might be achieved by a saturation assumption following again the line of proof in [BPRR19a] for adaptive stochastic Galerkin FEM. Conversely, the extension of the specific model problem to the important case of the diffusion coefficient resulting from the parametrization of a log-normal random field is deemed to be more challenging. Another important yet challenging addition to our work would be to extend the convergence result to the infinite-dimensional case, i.e., to consider countably many parameters $M = \infty$ in the affine expansion of the diffusion coefficient (2). This would pose both theoretical and algorithmic challenges: on the theoretical side, our proof would need to be revisited since some constants are not bounded when $M \rightarrow \infty$ (in particular, the constant C in Lemma 16, cf. equation (30)). From the algorithmic point of view, having $M = \infty$ would lead to margin sets of infinite cardinality which is, of course, unfeasible. Under the assumption that $\|a_m\|_{L^\infty}$ in (2) are monotone decreasing (this assumption could be weakened), then a possible approach would be to implement a so-called “buffering” procedure, as discussed in [GN18] (see also [SS13, CCS14, NTTT16, EST18]): such an algorithm would start considering only the first $M_0 < \infty$ parameters, and any time a parameter is “activated” (i.e. a collocation point is added along that parameter dimension for the first time), the total number of considered parameters would increase by one, in such a way that there are always M_0 “non-activated” parameters.

A further interesting follow-up would be to carry out an extensive numerical study on a number of different PDEs for which finite element error estimators are available, and investigate numerically whether Algorithm 3 consistently displays good performance (i.e., similar to the GG algorithm) for all the PDEs considered. Both these numerical investigations exceed the scope of this work and are left for future research.

References

- [BCDVM] M. Bachmayr, A. Cohen, R. De Vore, and G. Migliorati, *Sparse polynomial approximation of parametric elliptic pdes. part ii: lognormal coefficients*, ESAIM: Mathematical Modelling and Numerical Analysis.
- [BCM17] M. Bachmayr, A. Cohen, and G. Migliorati, *Sparse polynomial approximation of parametric elliptic PDEs. part i: affine coefficients*, ESAIM: M2AN **51** (2017), no. 1, 321–339.

- [Bie11] Marcel Bieri, *A sparse composite collocation finite element method for elliptic sPDEs*, SIAM J. Numer. Anal. **49** (2011), no. 6, 2277–2301.
- [BNT07] Ivo Babuška, Fabio Nobile, and Raúl Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal. **45** (2007), no. 3, 1005–1034.
- [BPRR19a] A. Bespalov, D. Praetorius, L. Rocchi, and M. Ruggeri, *Convergence of adaptive stochastic Galerkin FEM*, SIAM J. Numer. Anal. **57** (2019), no. 5, 2359–2382.
- [BPRR19b] A. Bespalov, D. Praetorius, L. Rocchi, and M. Ruggeri, *Goal-oriented error estimation and adaptivity for elliptic PDEs with parametric or uncertain inputs*, Comput. Methods Appl. Mech. Engrg. **345** (2019), 951–982.
- [BPS14] A. Bespalov, C. E. Powell, and D. Silvester, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM J. Sci. Comput. **36** (2014), no. 2, A339–A363.
- [Bru78] L. Brutman, *On the Lebesgue function for polynomial interpolation*, SIAM J. Numer. Anal. **15** (1978), no. 4, 694–704.
- [BS09] Marcel Bieri and Christoph Schwab, *Sparse high order FEM for elliptic sPDEs*, Comput. Methods Appl. Mech. Engrg. **198** (2009), 1149–1170.
- [BS16] A. Bespalov and D. Silvester, *Efficient adaptive stochastic Galerkin methods for parametric operator equations*, SIAM J. Sci. Comput. **38** (2016), A2118–A2140.
- [BTNT12] Joakim Beck, Raul Tempone, Fabio Nobile, and Lorenzo Tamellini, *On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods*, M3AS **22** (2012), no. 9, 1250023 (33 pages).
- [CCS14] M. A. Chkifa, A. Cohen, and Ch. Schwab, *High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs*, Found. Comput. Math. **14** (2014), 601–633.
- [CDS10] A. Cohen, R. DeVore, and C. Schwab, *Convergence rates of best n -term Galerkin approximations for a class of elliptic sPDEs*, Foundations of Computational Mathematics **10** (2010), 615–646.
- [CDS11] A. Cohen, R. Devore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE’S*, Anal. Appl. (Singap.) **9** (2011), no. 1, 11–47.
- [Che18] Chen, P., *Sparse quadrature for high-dimensional integration with Gaussian measure*, ESAIM: M2AN **52** (2018), no. 2, 631–657.
- [Chk13] M. A. Chkifa, *On the Lebesgue constant of Leja sequences for the complex unit disk and of their real projection*, Journal of Approximation Theory **166** (2013), 176–200.
- [Chk15] M. A. Chkifa, *New bounds on the Lebesgue constants of Leja sequences on the unit disc and on R -Leja sequences*, Curves and Surfaces (J.-D. et al. Boissonnat, ed.), Springer International Publishing, 2015, pp. 109–128.
- [CM18] Albert Cohen and Giovanni Migliorati, *Multivariate approximation in downward closed polynomial spaces*, pp. 233–282, Springer International Publishing, Cham, 2018.
- [CPB19] A. J. Crowder, C. E. Powell, and A. Bespalov, *Efficient adaptive multilevel stochastic galerkin approximation using implicit a posteriori error estimation*, SIAM J. Sci. Comput. **41** (2019), A1681–A1705.
- [Dör96] Willy Dörfler, *A Convergent Adaptive Algorithm for Poisson’s Equation*, SIAM Journal on Numerical Analysis **33** (1996), no. 3, 1106–1124.
- [EGSZ14] M. Eigel, C. J. Gittelsohn, C. Schwab, and E. Zander, *Adaptive stochastic Galerkin FEM*, Comput. Methods Appl. Mech. Engrg. **270** (2014), 247–269.
- [EGSZ15] S. Eigel, C. J. Gittelsohn, Ch. Schwab, and E. Zander, *A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes*, ESAIM: M2AN **49** (2015), 1367–1398.

- [EM16] Martin Eigel and Christian Merdon, *Local equilibration error estimators for guaranteed error control in adaptive stochastic higher-order galerkin finite element methods*, SIAM/ASA Journal on Uncertainty Quantification **4** (2016), no. 1, 1372–1397.
- [EMPS20] Martin Eigel, Manuel Marschall, Max Pfeffer, and Reinhold Schneider, *Adaptive stochastic galerkin fem for lognormal coefficients in hierarchical tensor representations*, Numerische Mathematik (2020), 1–38.
- [EPS17] Martin Eigel, Max Pfeffer, and Reinhold Schneider, *Adaptive stochastic galerkin fem with hierarchical tensor representations*, Numerische Mathematik **136** (2017), no. 3, 765–803.
- [EST18] 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), no. 2, 877–905.
- [EST19] O. G. Ernst, B. Sprungk, and L. Tamellini, *On Expansions and Nodes for Sparse Grid Collocation of Lognormal Elliptic PDEs*, Arxiv e-prints (2019), no. 1906.01252, Accepted. Also available as IMATI report 19-02.
- [FGB⁺20] Ionuț-Gabriel Farcaș, Tobias Görler, Hans-Joachim Bungartz, Frank Jenko, and Tobias Neckel, *Sensitivity-driven adaptive sparse stochastic approximations in plasma microinstability analysis*, Journal of Computational Physics **410** (2020), 109394.
- [FS20] M. Feischl and A. Scaglioni, *Convergence of adaptive stochastic collocation with finite elements*, 2020, Available as arXiv:2008.12591.
- [GG03] Thomas Gerstner and Michael Griebel, *Dimension-adaptive tensor-product quadrature*, Computing **71** (2003), 65–87.
- [GK09] Michael Griebel and Stephan Knapek, *Optimized general sparse grid approximation spaces for operator equations*, Math. Comp. **78** (2009), no. 268, 2223–2257.
- [GN18] D. Guignard and F. Nobile, *A posteriori error estimation for the stochastic collocation finite element method*, SIAM J. Numer. Anal. **56** (2018), no. 5, 3121–3143.
- [GSZ92] M. Griebel, M. Schneider, and C. Zenger, *A combination technique for the solution of sparse grid problems*, Iterative Methods in Linear Algebra (P. de Groen and R. Beauwens, eds.), IMACS, Elsevier, North Holland, 1992, pp. 263–281.
- [Heg03] M. Hegland, *Adaptive sparse grids*, Proc. of 10th Computational Techniques and Applications Conference CTAC-2001 (K. Burrage and Roger B. Sidje, eds.), vol. 44, April 2003, pp. C335–C353.
- [HS14] Viet Ha Hoang and Christoph Schwab, *N-term Wiener chaos approximation rates for elliptic PDEs with lognormal Gaussian random inputs*, Mathematical Models and Methods in Applied Sciences **24** (2014), no. 4, 797–826.
- [Kli06] A. Klimke, *Uncertainty modeling using fuzzy arithmetic and sparse grids*, Ph.D. thesis, Universität Stuttgart, Shaker Verlag, Aachen, 2006.
- [LSS19] Jens Lang, Robert Scheichl, and David Silvester, *A fully adaptive multilevel stochastic collocation strategy for solving elliptic PDEs with random data*, 2019, Available as arXiv:1902.03409.
- [MH03] Lionel Mathelin and M. Hussaini, *A stochastic collocation algorithm for uncertainty analysis*, Technical Report NASA/CR-2003-212153, NASA Langley Research Center, 2003.
- [MP73] J. H. McCabe and G. M. Phillips, *On a certain class of Lebesgue constants*, BIT **13** (1973), 694–704.
- [MZ09] Xiang Ma and Nicholas Zabaras, *An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations*, J. Comp. Phys. **228** (2009), 3084–3113.

- [NTT15] F. Nobile, L. Tamellini, and R. Tempone, *Comparison of Clenshaw-Curtis and Leja Quasi-Optimal Sparse Grids for the Approximation of Random PDEs*, Spectral and High Order Methods for Partial Differential Equations - ICOSAHOM '14 (R. M. Kirby, M. Berzins, and J. S. Hesthaven, eds.), Lecture Notes in Computational Science and Engineering, vol. 106, Springer International Publishing, 2015, pp. 475–482.
- [NTTT16] Fabio Nobile, Lorenzo Tamellini, Francesco Tesei, and Raúl Tempone, *An adaptive sparse grid algorithm for elliptic PDEs with lognormal diffusion coefficient*, Sparse Grids and Applications-Stuttgart 2014 (J. Garcke and D. Pflüger, eds.), Springer, Cham, 2016, pp. 191–220.
- [NTW08a] Fabio Nobile, Raúl Tempone, and Clayton G. Webster, *An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM J. Numer. Anal. **46** (2008), no. 5, 2411–2442.
- [NTW08b] ———, *A sparse grid stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal. **46** (2008), no. 5, 2309–2345.
- [Pet03] K. Petras, *Smolyak cubature of given polynomial degree with few nodes for increasing dimension*, Numerische Mathematik **93** (2003), 729–753.
- [SJ14] B. Schieche and Lang J., *Adjoint error estimation for stochastic collocation methods*, Sparse Grids and Applications - Munich 2012 (Cham) (J. Garcke and D. Pflüger, eds.), Lecture Notes in Computational Science and Engineering, vol. 97, Springer, 2014, pp. 271–293.
- [SS13] Claudia Schillings and Christoph Schwab, *Sparse, adaptive Smolyak quadratures for Bayesian inverse problems*, Inverse Probl. **29** (2013), no. 6, 065011.
- [XH05] Dongbin Xiu and Jan S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput. **27** (2005), no. 3, 1118–1139.
- [ZS20] Zech, Jakob and Schwab, Christoph, *Convergence rates of high dimensional smolyak quadrature*, ESAIM: M2AN **54** (2020), no. 4, 1259–1307.