

ORNL REPORT

ORNL/TM-2014/468
Unlimited Release
Printed October 2014

Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients

H. Tran, C. G. Webster and G. Zhang

Prepared by
Department of Computational and Applied Mathematics
Computer Science and Mathematics Division
Oak Ridge National Laboratory
One Bethel Valley Road, Oak Ridge, Tennessee 37831

The Oak Ridge National Laboratory is operated by UT-Battelle, LLC,
for the United States Department of Energy under Contract DE-AC05-00OR22725.
Approved for public release; further dissemination unlimited.

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via the U.S. Department of Energy (DOE) Information Bridge.

Web site <http://www.osti.gov/bridge>

Reports produced before January 1, 1996, may be purchased by members of the public from the following source.

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Oak Ridge, TN 37831

Telephone 703-605-6000 (1-800-553-6847)

TDD 703-487-4639

Fax 703-605-6900

E-mail info@ntis.gov

Web site <http://www.ntis.gov/support/ordernowabout.htm>

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and International Nuclear Information System (INIS) representatives from the following source.

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone 865-576-8401

Fax 865-576-5728

E-mail reports@osti.gov

Web site <http://www.osti.gov/contact.html>

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.



ANALYSIS OF QUASI-OPTIMAL POLYNOMIAL APPROXIMATIONS FOR PARAMETERIZED PDES WITH DETERMINISTIC AND STOCHASTIC COEFFICIENTS

Hoang A. Tran ^{*} Clayton G. Webster [†] Guannan Zhang [‡]

Abstract. In this work, we present a generalized methodology for analyzing the convergence of quasi-optimal Taylor and Legendre approximations, applicable to a wide class of parameterized elliptic PDEs with both deterministic and stochastic inputs. Such methods construct an index set Λ_M that corresponds to the “best M -terms” based on sharp estimates of the polynomial coefficients. In particular, we consider several cases of N dimensional affine and non-affine diffusion coefficients, and prove analytic dependence of the PDE solution map $\mathbf{z} \mapsto u(\mathbf{z})$ in a polydisc or polyellipse of the complex plane \mathbb{C}^N respectively. The framework we propose for analyzing asymptotic truncation errors of quasi-optimal methods is based on an extension of the underlying multi-index set into a continuous domain, and then an approximation of the cardinality (number of integer multi-indices) by its Lebesgue measure. Several types of isotropic and anisotropic (weighted) multi-index sets are explored, and rigorous proofs reveal sharp asymptotic error estimates in which we achieve sub-exponential convergence rates (of the form $M \exp(-(\kappa M)^{1/N})$, with κ a constant depending on the shape and size of multi-index sets) with respect to the total number of degrees of freedom. Through several theoretical examples, we explicitly derive the constant κ and use the resulting sharp bounds to illustrate the effectiveness of Legendre over Taylor approximations, as well as compare our rates of convergence with current published results. Finally, computational evidence complements the theory and shows the advantage of our generalized methodology compared to previously developed estimates.

1. Introduction. This paper focuses on a relevant model boundary value problem, involving the simultaneous solution of a family of equations, parameterized by a vector $\mathbf{y} = (y_1, \dots, y_N) \in \Gamma = \prod_{i=1}^N \Gamma_i \subset \mathbb{R}^N$, on a bounded Lipschitz domain $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. In particular, we consider a differential operator \mathcal{L} defined on D , and let $a(x, \mathbf{y})$, with $x \in D$ and $\mathbf{y} \in \Gamma$, represent the input coefficient associated with the operator \mathcal{L} . The forcing term $f = f(x) \in L^2(D)$ is assumed to be a fixed function of $x \in D$. We concentrate on the following parameterized boundary value problem: for all $\mathbf{y} \in \Gamma$, find $u(\cdot, \mathbf{y}) : \bar{D} \rightarrow \mathbb{R}$, such that the following equation holds

$$\mathcal{L}(a(\cdot, \mathbf{y})) [u(\cdot, \mathbf{y})] = f(\cdot) \quad \text{in } D, \tag{1.1}$$

subject to suitable (possibly parameterized) boundary conditions. We require a and f to be chosen such that system (1.1) is well-posed in a Banach space, with unique solution u , such that, when suppressing the explicit dependence on x , the map $\mathbf{y} \mapsto u(\mathbf{y})$ is defined from the parameter domain Γ into the solution space $V(D)$.

Problems such as (1.1) arise in contexts of both deterministic and stochastic modeling. In the deterministic setting, the parameter vector \mathbf{y} is known or controlled by the user, and a typical goal is to study the dependence of u on these parameters, e.g., optimizing an output of the equation with respect to \mathbf{y} (see [10,30] for more details). On the other hand, stochastic modeling is motivated by many engineering and science problems in which the input data

^{*}Department of Computational and Applied Mathematics, Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, MS-6164, Oak Ridge, TN 37831-6164 (tranha@ornl.gov).

[†]Department of Computational and Applied Mathematics, Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, MS-6164, Oak Ridge, TN 37831-6164 (webstercg@ornl.gov).

[‡]Department of Computational and Applied Mathematics, Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, MS-6164, Oak Ridge, TN 37831-6164 (zhangg@ornl.gov).

is not known exactly. A quantification of the effect of the input uncertainties on the output of simulations is necessary to obtain a reliable prediction of the physical system. A natural way to incorporate the presence of input uncertainties into the governing model (1.1) is to consider the parameters $\{y_n(\omega)\}_{n=1}^N$ as random variables and $\mathbf{y}(\omega) : \Omega \rightarrow \Gamma$ a random vector, where $\omega \in \Omega$ and Ω is the set of outcomes. In this setting, we assume the components of \mathbf{y} have a joint probability density function (PDF) $\varrho : \Gamma \rightarrow \mathbb{R}_+$, with $\varrho \in L^\infty(\Gamma)$ known directly through, e.g., truncations of correlated random fields [20, 28, 29, 36], such that the probability space is equivalent to $(\Gamma, \mathcal{B}(\Gamma), \varrho(\mathbf{y})d\mathbf{y})$, where $\mathcal{B}(\Gamma)$ denotes the Borel σ -algebra on Γ and $\varrho(\mathbf{y})d\mathbf{y}$ is the probability measure of \mathbf{y} .

Monte Carlo (MC) methods (see, e.g., [18]) are the most popular approaches for approximating high-dimensional integrals, such as expectation or two-point correlation, based on independent realizations $u(\mathbf{y}_k)$, $k = 1, \dots, M$, of the solution to (1.1); approximations of the expectation or other QoIs are obtained by averaging over the corresponding realizations of that quantity. The resulting numerical error is proportional to $M^{-1/2}$, thus, achieving convergence rates independent of dimension N , but requiring a very large number of samples to achieve reasonably small errors. Moreover, MC methods do not have the ability to simultaneously approximate the solution map $\mathbf{y} \mapsto u(\mathbf{y})$, since they are quadrature techniques and do not exploit the fact in many scenarios, the solutions smoothly depend on the coefficient a . Taking this smooth dependence into account, several global polynomial approximation techniques, for instance, *intrusive* Galerkin methods [3, 35] and *non-intrusive* collocation methods [2, 32], have been proposed, often featuring much faster convergence rates.

Let $\mathcal{S} = \{\boldsymbol{\nu} = (\nu_i)_{1 \leq i \leq N} : \nu_i \in \mathbb{N}\}$. Global polynomial approximation methods seek to build an approximation u_Λ to the solution u of the form:

$$u_\Lambda(x, \mathbf{y}) = \sum_{\boldsymbol{\nu} \in \Lambda} c_\nu(x) \Psi_\nu(\mathbf{y}), \quad (1.2)$$

for a finite multi-index set $\Lambda \subset \mathcal{S}$, where Ψ_ν is a multivariate polynomial in $\text{span}\{\mathbf{y}^\mu : \boldsymbol{\mu} \leq \boldsymbol{\nu}\}$ for $\boldsymbol{\nu} \in \Lambda$ and $c_\nu \in V(D)$ is the coefficient to be computed, both of which are method specific. Here, for two vectors $\boldsymbol{\nu}, \boldsymbol{\mu} \in \mathcal{S}$, we say $\boldsymbol{\mu} \leq \boldsymbol{\nu}$ if and only if $\mu_i \leq \nu_i$ for all $1 \leq i \leq N$. Also, given $\boldsymbol{\alpha} = (\alpha_i)_{1 \leq i \leq N}$ a vector of real numbers, we define $\boldsymbol{\alpha}^\nu = \prod_{1 \leq i \leq N} \alpha_i^{\nu_i}$ with the convention $0^0 := 1$. We will often suppress the dependence on x and use the notations $u(\mathbf{y}) := u(\cdot, \mathbf{y})$ and $a(\mathbf{y}) := a(\cdot, \mathbf{y})$ without loss of generality. In this paper, we are interested in solving (1.1) using a class of polynomial approximations based on the Taylor and Legendre expansions of solution u . The polynomial basis considered herein is thus given by the monomials $\Psi_\nu(\mathbf{y}) = \mathbf{y}^\nu$ (in the former case) and the tensorized Legendre polynomials $\Psi_\nu(\mathbf{y}) = L_\nu(\mathbf{y})$ (in the latter case).

The evaluation of u_Λ in (1.2) requires the computation of $\#\Lambda$ coefficients $c_\nu(x) \in V(D)$, where $\#\Lambda$ is the cardinality of Λ . A naive choice of Λ and their corresponding polynomial spaces $\mathbb{P}_\Lambda(\Gamma) = \text{span}\{\Psi_\nu(\mathbf{y}), \boldsymbol{\nu} \in \Lambda\}$, for instance, tensor product polynomial spaces, could lead to an infeasible computational cost, especially when the dimension of the parameter domain is high. It is important to be able to construct the set of the most effective indices for the approximation (1.2), which provides maximum accuracy for a given cardinality. In other words, given a fixed $M \in \mathbb{N}$, one searches for a set Λ_M which minimizes

the error $u - \sum_{\nu \in \Lambda} c_\nu \Psi_\nu$ among all index sets $\Lambda \subset \mathcal{S}$ of cardinality M . This practice has been known as *best M -term approximations*.

The literature on the best M -term Taylor and Galerkin approximations has been growing fast recently, among them we refer to [6, 7, 9, 11–14, 22, 23, 25]. In the benchmark work [14], the analytic dependence of the solutions of parametric elliptic PDEs on the parameters was proved under mild assumptions on the input coefficients, and convergence analysis of the best M -term Taylor and Legendre approximations was established subsequently. Consider, for example, the expansion of u on $\Gamma = [-1, 1]^N$ by a family of L^∞ normalized polynomials, i.e., $\|\Psi_\nu\|_{L^\infty(\Gamma)} = 1$. Application of the triangle inequality yields

$$\sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\nu \in \Lambda_M} c_\nu \Psi_\nu(\mathbf{y}) \right\|_{V(D)} \leq \sum_{\nu \in \Lambda_M^c} \|c_\nu\|_{V(D)},$$

which suggests determining the optimal index set Λ_M is achieved by choosing the set of indices ν corresponding to M largest $\|c_\nu\|_{V(D)}$. Here, Λ_M^c denotes the complement of Λ_M in \mathcal{S} . In [14], the error of such approximation was estimated due to Stechkin inequality (see, e.g., [16]) such that

$$\sum_{\nu \in \Lambda_M^c} \|c_\nu\|_{V(D)} \leq \|(\|c_\nu\|_{V(D)})_{\nu \in \mathcal{S}}\|_{\ell^p(\mathcal{S})} M^{1-\frac{1}{p}}, \quad (1.3)$$

where p is some number in $(0, 1)$ such that $(\|c_\nu\|_{V(D)})_{\nu \in \mathcal{S}}$ is ℓ^p -summable. It should be noted that the convergence rate (1.3) does not depend on the dimension of the parameter domain Γ (which is possibly countably infinite therein). This error estimate, however, has some limitations. First, sharp, explicit evaluation of the coefficient $\|(\|c_\nu\|_{V(D)})_{\nu \in \mathcal{S}}\|_{\ell^p(\mathcal{S})}$ is inaccessible in general (thus so is the total estimate). Secondly, (1.3) often occurs with infinitely many values of p and stronger rates, corresponding to smaller p , are also attached to bigger coefficients. For a specific range of M , the effective rate of convergence is unclear. In implementation, finding the best index set and polynomial space is an infeasible task, since this requires computation of all of the c_ν . As a strategy to circumvent this challenge, adaptive algorithms which generate the index set in a near optimal, greedy procedure were developed in [11]. This method gives the optimal rates; however, it comes with a high cost of exploring the polynomial space, which may be daunting in high-dimensional problems.

Instead of building the index set based on exact values of polynomial coefficients c_ν , an attractive alternative approach (referred to as *quasi-optimal approximation* throughout this paper) is to establish sharp upper bounds of c_ν (by a priori or a posteriori methods), and then construct Λ_M corresponding to M largest such bounds. For this strategy, the main computational work for the selection of the (near) best terms reduces to determining sharp coefficient estimates, which is expected to be significantly cheaper than exact calculations. Quasi-optimal polynomial approximation has been performed for some parametric elliptic models with optimistic results: while the upper bounds of $\|c_\nu\|_{V(D)}$ (denoted from now by $B(\nu)$) were computed with a negligible cost, the method was comparably as accurate as best M -term approach, as shown in [6, 7]. The first rigorous numerical analysis of quasi-optimal approximation was presented in [6] for $B(\nu) = \rho^{-\nu}$ with ρ being a vector $(\rho_i)_{1 \leq i \leq N}$ with

$\rho_i > 1 \forall i$. In that work, the asymptotic sub-exponential convergence rate was proved based on optimizing the Stechkin estimation. Briefly, the analysis applied Stechkin inequality to yield

$$\sum_{\boldsymbol{\nu} \in \Lambda_M^c} B(\boldsymbol{\nu}) \leq \|B(\boldsymbol{\nu})\|_{\ell^p(\mathcal{S})} M^{1-\frac{1}{p}}, \quad (1.4)$$

then took advantage of the formula of $B(\boldsymbol{\nu})$ to compute $p \in (0, 1)$, depending on M , which minimizes $\|B(\boldsymbol{\nu})\|_{\ell^p(\mathcal{S})} M^{1-\frac{1}{p}}$.

Although known as an essential tool to study the convergence rate of best M -term approximations, Stechkin inequality is probably less efficient for quasi-optimal methods. As a generic estimate, it does not fully exploit the available information of the decay of coefficient bounds. In such a setting, a direct estimate of $\sum_{\boldsymbol{\nu} \in \Lambda_M^c} B(\boldsymbol{\nu})$ may be viable and advantageous to provide a sharper result. In addition, the process of solving the minimization problem $p^* = \operatorname{argmin}_{p \in (0,1)} \|B(\boldsymbol{\nu})\|_{\ell^p(\mathcal{S})} M^{1-\frac{1}{p}}$ needs to be tailored to $B(\boldsymbol{\nu})$, making this approach not ideal for generalization. Currently, this minimization approach has been limited for some quite simple types of upper bounds. In many scenarios, the sharp estimates of the coefficients may involve complicated bounds which are not even explicitly computable, such as those proposed in [14]. The extension of this approach to such cases seems to be impossible.

In this work, we present a generalized methodology for convergence analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients. We particularly focus on elliptic equations where the input coefficient depends affinely and non-affinely on the parameters (see Section 2). However, since our error analysis only depends on the upper bounds of polynomial coefficients, we expect that the methods and results presented herein can be applied to other, more general model problems with finite parametric dimension, including nonlinear elliptic PDEs, initial value problems and parabolic equations [12,22,23,25]. Our approach seeks a direct estimate of $\sum_{\boldsymbol{\nu} \in \Lambda_M^c} B(\boldsymbol{\nu})$ without using the Stechkin inequality. It involves a partition of $B(\Lambda_M^c)$ into a family of small positive real intervals $(\mathcal{I}_j)_{j \in \mathcal{J}}$ and the corresponding splitting of Λ_M^c into disjoint subsets \mathcal{Q}_j of indices $\boldsymbol{\nu}$, such that $B(\boldsymbol{\nu}) \in \mathcal{I}_j$. Under this process, the truncation error can be bounded as

$$\sum_{\boldsymbol{\nu} \in \Lambda_M^c} B(\boldsymbol{\nu}) = \sum_{j \in \mathcal{J}} \sum_{\boldsymbol{\nu} \in \mathcal{Q}_j} B(\boldsymbol{\nu}) \leq \sum_{j \in \mathcal{J}} \#(\mathcal{Q}_j) \cdot \max(\mathcal{I}_j),$$

and therefore, the quality of the error estimate mainly depends on the approximation of cardinality of \mathcal{Q}_j . To tackle this problem, we develop a strategy which extends \mathcal{Q}_j into continuous domain and, through relating the number of N -dimensional lattice points to continuous volume (Lebesgue measure), establishes a sharp estimate of the cardinality $\#(\mathcal{Q}_j)$ up to any prescribed accuracy. This development includes the utilization and extension of several results on lattice point enumeration; for a survey we refer to [8,21]. Under some weak assumptions on $B(\boldsymbol{\nu})$ (which are satisfied by all existing coefficient estimates we are aware of), we achieve an asymptotic sub-exponential convergence rate of truncation error of the form $M \exp(-(\kappa M)^{1/N})$, where κ is a constant depending on the shape and size of

quasi-optimal index sets. Through several examples, we explicitly derive κ and demonstrate the optimality of our estimate both theoretically (by proving a lower bound) and computationally (via comparison with exact calculation of truncation error). The advantage of our analysis framework is therefore twofold. First, it applies to a general class of quasi-optimal approximations; and second, it yields sharp estimates of the asymptotic convergence rates.

Our paper is organized as follows. In Section 2, we describe the elliptic equations with parameterized input coefficient and necessary mathematical notations which are used throughout the paper. In Section 3, we prove the analyticity of the solution u with respect to parameter and derive coefficient estimates of Taylor and Legendre expansions of u . The advantage of Legendre over Taylor expansions will also be discussed. We give a convergence analysis for a general class of multi-indexed series $\sum_{\nu \in \mathcal{S}} B(\nu)$ in Section 4. Section 5 is devoted to further discussions on the error lower bound, as well as the pre-asymptotic estimate in a simplified case. By means of these results, asymptotic error estimate of several quasi-optimal polynomial approximations is presented in Section 6.

2. Problem setting. We consider solving simultaneously the following parameterized linear, elliptic PDE:

$$\begin{cases} -\nabla \cdot (a(x, \mathbf{y}) \nabla u(x, \mathbf{y})) &= f(x), \quad \forall (x, \mathbf{y}) \in D \times \Gamma, \\ u(x, \mathbf{y}) &= 0, \quad \forall (x, \mathbf{y}) \in \partial D \times \Gamma, \end{cases} \quad (2.1)$$

on a bounded Lipschitz domain $D \subset \mathbb{R}^d$, with the coefficient $a(\cdot, \mathbf{y})$ defined on $\Gamma = \prod_{i=1}^N \Gamma_i \subset \mathbb{R}^N$, with $\Gamma_i = [-1, 1]$, $\forall i \in \{1, \dots, N\}$. We require the following assumption:

Assumption 1 (Continuity and coercivity). *There exist constants $0 < a_{\min} \leq a_{\max}$ such that for all $x \in \bar{D}$ and $\mathbf{y} \in \Gamma$*

$$a_{\min} \leq a(x, \mathbf{y}) \leq a_{\max}.$$

The Lax-Milgram lemma ensures the existence and uniqueness of solution u in $V(D) \otimes L^2_{\varrho}(\Gamma)$, where $V(D) = H_0^1(D)$ and $L^2_{\varrho}(\Gamma)$ is the space of square integrable functions on Γ with respect to the measure $\varrho(\mathbf{y}) d\mathbf{y}$ with $\varrho(\mathbf{y}) = \prod_{i=1}^N \varrho_i(y_i)$, $\varrho_i = \frac{1}{2}$, $\forall \mathbf{y} \in \Gamma$. This setting represents parametric elliptic models as well as stochastic models with bounded support random coefficient. We denote $V^*(D) = H^{-1}(D)$ and, without loss of generality, assume $a_{\min} = 1$ in this work.

The corresponding weak formulation for (2.1) is written as follows: find $u(x, \mathbf{y}) \in V(D) \otimes L^2_{\varrho}(\Gamma)$ such that

$$\begin{aligned} & \int_{\Gamma} \int_D a(x, \mathbf{y}) \nabla u(x, \mathbf{y}) \cdot \nabla v(x, \mathbf{y}) dx d\mathbf{y} \\ &= \int_{\Gamma} \int_D f(x) v(x, \mathbf{y}) dx d\mathbf{y} \quad \forall v \in V(D) \otimes L^2_{\varrho}(\Gamma). \end{aligned} \quad (2.2)$$

Following the arguments in [14], we derive the convergence of Taylor and Legendre approximations based on the analyticity of the solution on complex domains. Here, the convergence is proved under the affine parameter dependence of diffusion coefficients for the Taylor series, but we relax this assumption for the Legendre series. More specifically,

we only assume a holomorphic extension $a(x, \mathbf{z})$ of $a(x, \mathbf{y})$ for the complex variable $\mathbf{z} = (z_1, \dots, z_N)^\top$:

Assumption 2 (Holomorphic parameter dependence). *The complex continuation of a , represented as the map $a : \mathbb{C}^N \rightarrow L^\infty(D)$, is a $L^\infty(D)$ -valued holomorphic function on \mathbb{C}^N .*

This condition is easily fulfilled with $a(x, \mathbf{y})$ consisting of polynomials, exponential, sine and cosine functions of the variables y_1, \dots, y_N . Below, we give some examples of diffusion coefficients which can be accommodated in our framework. The rigorous proofs and discussion on the advantage of Legendre over Taylor approximations will be postponed to the next section.

Example 1. For the input coefficient depending affinely on the parameters, i.e.,

$$a(x, \mathbf{y}) = a_0(x) + \sum_{i=1}^N y_i \psi_i(x), \quad x \in \bar{D}, \mathbf{y} \in \Gamma,$$

where $a_0 \in L^\infty(D)$, $(\psi_i)_{1 \leq i \leq N} \subset L^\infty(D)$ such that a satisfies Assumption 1; both Taylor and Legendre series approximations of $u(\mathbf{y})$ to (2.1) converge.

Example 2. Consider the input coefficient defined as

$$a(x, \mathbf{y}) = a_0(x) + \left(\sum_{i=1}^N y_i \psi_i(x) \right)^2, \quad x \in \bar{D}, \mathbf{y} \in \Gamma,$$

with $a_0 \in L^\infty(D)$, $a_0(x) \geq a_{\min} > 0 \forall x \in \bar{D}$ and $(\psi_i)_{1 \leq i \leq N} \subset L^\infty(D)$. It is easy to see that $a(x, \mathbf{y})$ satisfies Assumptions 1–2. Thus, the Legendre series approximation of $u(\mathbf{y})$ to (2.1) converges for this model.

Example 3. Consider the input coefficient defined as

$$a(x, \mathbf{y}) = a_0(x) + \exp \left(\sum_{i=1}^N y_i \psi_i(x) \right), \quad x \in \bar{D}, \mathbf{y} \in \Gamma,$$

with $a_0 \in L^\infty(D)$, $a_0(x) \geq 0 \forall x \in \bar{D}$ and $(\psi_i)_{1 \leq i \leq N} \subset L^\infty(D)$. We have $a(x, \mathbf{y})$ satisfies Assumptions 1–2 and Legendre series approximation of $u(\mathbf{y})$ to (2.1) converges.

Another framework for establishing convergence of Legendre series was presented in [12] and applied to a large variety of parametric PDEs (non-elliptic, infinite dimensional noise and non-affine dependence on parameters). This approach imposes analyticity assumptions on the solution, which requires nontrivial validation in practice. Instead, in this work, we focus on elliptic equations which allows us to derive concise, minimal assumptions on the input coefficient (as seen above), under which the convergence of Legendre approximations holds straightforwardly. It is also worth recalling that quasi-optimal error estimates only depend on the sharp upper bound of the polynomial coefficients. Therefore, while not studied herein, PDE models covered by [12, 22, 23, 25], bringing about same types of coefficient bounds as those considered in Section 6, can be treated by our quasi-optimal analysis.

3. Analyticity of the solutions and estimates of the polynomial coefficients. Loosely speaking, the coefficients of Taylor and Legendre expansions can be estimated via three steps:

1. Extending the uniform ellipticity of a from Γ to certain polydiscs/polyellipses in \mathbb{C}^N ;
2. Proving the analyticity of the solution on those extended domains, and;
3. Estimating the expansion coefficients using the analyticity properties and Cauchy's integral formula.

We will discuss each step in detail in the next subsections. By $\Re(z)$ and $\Im(z)$, we denote the real and imaginary part of a complex number z .

3.1. Complex uniform ellipticity. The convergence of Taylor approximations is proved using the uniform ellipticity of the input coefficient in polydiscs containing Γ , based on complex analysis argument.

Definition 1. For $0 < \delta < a_{\min}$ and $\boldsymbol{\rho}$ denoting the vector $(\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \ \forall i$, we say $a(x, \mathbf{y})$ satisfies $(\delta, \boldsymbol{\rho})$ -polydisc uniform ellipticity assumption (referred to as **DUE**($\delta, \boldsymbol{\rho}$)) if there holds

$$\Re(a(x, \mathbf{z})) \geq \delta$$

for all $x \in \overline{D}$ and all $\mathbf{z} = (z_i)_{1 \leq i \leq N}$ contained in the polydisc

$$\mathcal{O}_{\boldsymbol{\rho}} = \bigotimes_{1 \leq i \leq N} \{z_i \in \mathbb{C} : |z_i| \leq \rho_i\}.$$

At the same time, Legendre expansions require the uniform ellipticity in smaller complex domains: the polyellipses.

Definition 2. For $0 < \delta < a_{\min}$ and $\boldsymbol{\rho}$ denoting the vector $(\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \ \forall i$, we say $a(x, \mathbf{y})$ satisfies $(\delta, \boldsymbol{\rho})$ -polyellipse uniform ellipticity assumption (referred to as **EUE**($\delta, \boldsymbol{\rho}$)) if there holds

$$\Re(a(x, \mathbf{z})) \geq \delta$$

for all $x \in \overline{D}$ and all $\mathbf{z} = (z_i)_{1 \leq i \leq N}$ contained in the polyellipse

$$\mathcal{E}_{\boldsymbol{\rho}} = \bigotimes_{1 \leq i \leq N} \left\{ z_i \in \mathbb{C} : \Re(z_i) = \frac{\rho_i + \rho_i^{-1}}{2} \cos \phi, \Im(z_i) = \frac{\rho_i - \rho_i^{-1}}{2} \sin \phi, \phi \in [0, 2\pi) \right\}.$$

A close look at **DUE** and **EUE** reveals the advantage of Legendre over Taylor expansions. The polyellipses $\mathcal{E}_{\boldsymbol{\rho}}$ extend the real domain Γ in a continuous manner, so that if $\boldsymbol{\rho}$ tends toward $\mathbf{1}$, $\mathcal{E}_{\boldsymbol{\rho}}$ shrinks to Γ . Thus, it is hopeful that the uniform ellipticity property of $a(x, \mathbf{y})$ in Γ (Assumption 1) can carry over to some polyellipses $\mathcal{E}_{\boldsymbol{\rho}}$ (at least with $\boldsymbol{\rho}$ close to $\mathbf{1}$). In fact, we prove that **EUE** property is a consequence of Assumptions 1 and 2.

Lemma 1. Let $a : \Gamma \rightarrow L^\infty(D)$ be a continuous function satisfying Assumptions 1 and 2. Then, for all $\delta < a_{\min}$, there exists a vector $\boldsymbol{\rho} = (\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \ \forall i$ such that **EUE**($\delta, \boldsymbol{\rho}$) holds.

On the other hand, **DUE** always requires an extension of the coercive property in Γ to the unit polydisc $\mathcal{O}_{\mathbf{1}}$, to say the least, which is not possible generally. For illustration, the sets of \mathbf{z} such that $\Re(a(x, \mathbf{z})) \geq \delta$ for all $x \in \overline{D}$ with some fixed $\delta > 0$ (referred to as the

domains of uniform ellipticity) are plotted in Figure 1 for some typical 1-dimensional parametric coefficients. The maximal ellipses and discs contained in these domains are shown. We observe that for the affine coefficient, the set spans unrestrictedly along the imaginary axis, and discs covering Γ can easily be placed inside. It highlights the success of Taylor approximations for parameterized models which depend affinely on the parameters, [14, 22, 23]. This property however no longer holds for non-affine, yet holomorphic diffusion coefficients. Taylor approximations for these cases can be treated by a real analysis approach, but under additional strong constraints, see [7].

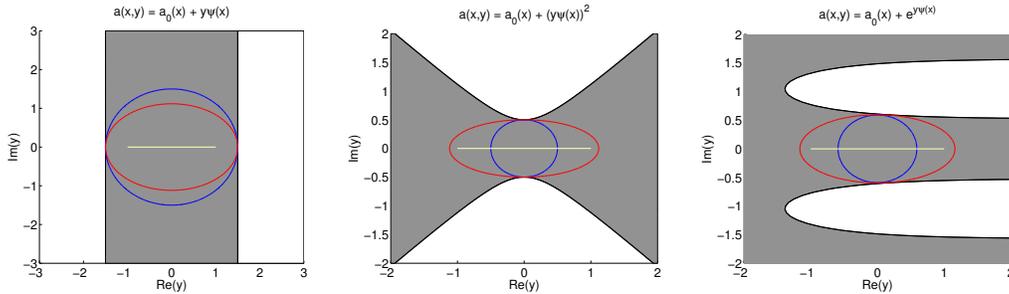


Figure 1: Domains of uniform ellipticity for some 1- d parametric input coefficients (indicated in gray). The yellow lines represent the interval $[-1, 1]$. The blue and red curves are the maximal discs and ellipses which can be contained in those domains, respectively.

We close this subsection with a proof of Lemma 1.

Proof. (of Lemma 1). Since $a(\mathbf{z})$ is holomorphic in \mathbb{C}^N , we have $\Re(a(\mathbf{z}))$ is a continuous mapping. By Heine-Cantor theorem, $\Re(a(\mathbf{z}))$ is uniformly continuous in any compact subset of \mathbb{C}^N . Fixing a $0 < \delta < a_{\min}$, without loss of generality, we can choose $\xi > 0$ such that $\forall \mathbf{z} \in \mathbb{C}^N, \forall \mathbf{z}' \in \Gamma$ satisfying $\|\mathbf{z} - \mathbf{z}'\| \leq \xi$, there holds

$$\|\Re(a(\mathbf{z})) - \Re(a(\mathbf{z}'))\|_{L^\infty(D)} \leq a_{\min} - \delta.$$

This implies

$$\Re(a(x, \mathbf{z})) \geq \delta - a_{\min} + \Re(a(x, \mathbf{z}')) \geq \delta,$$

for all $x \in \overline{D}$, $\mathbf{z} \in \mathbb{C}^N$ such that $\|\mathbf{z} - \mathbf{z}'\| \leq \xi$ with some $\mathbf{z}' \in \Gamma$. Denoting $\Gamma_\xi = \{\mathbf{z} \in \mathbb{C}^N : \text{dist}(\mathbf{z}, \Gamma) \leq \xi\}$, we proceed to prove there exists $\boldsymbol{\rho} = (\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \forall i$ such that the polyellipse $\mathcal{E}_{\boldsymbol{\rho}}$ is included in Γ_ξ .

First, consider the “polyrectangle”

$$\Xi = \bigotimes_{1 \leq i \leq N} \left\{ z_i \in \mathbb{C} : |\Re(z_i)| \leq 1 + \frac{\xi}{\sqrt{2N}}, |\Im(z_i)| \leq \frac{\xi}{\sqrt{2N}} \right\},$$

we will show that $\Xi \subset \Gamma_\xi$. Indeed, for every $\mathbf{z} \in \Xi$, choose $\mathbf{z}' = (z'_i)_{1 \leq i \leq N}$ as follows: if $|\Re(z_i)| \leq 1$, then $z'_i = \Re(z_i)$; otherwise, $z'_i = \text{sgn}(\Re(z_i))$. It is easy to see that $\mathbf{z}' \in \Gamma$.

Furthermore, for all $i \in \{1, \dots, N\}$,

$$|\Re(z_i) - \Re(z'_i)| \leq \frac{\xi}{\sqrt{2N}}, \quad |\Im(z_i) - \Im(z'_i)| \leq \frac{\xi}{\sqrt{2N}}.$$

Thus, $|z_i - z'_i| \leq \frac{\xi}{\sqrt{N}}$ and we have

$$\|\mathbf{z} - \mathbf{z}'\| \leq \left(\sum_{i=1}^N |z_i - z'_i|^2 \right)^{1/2} \leq \xi,$$

This gives $\mathbf{z} \in \Gamma_\xi$ and $\Xi \subset \Gamma_\xi$.

It remains to find $\boldsymbol{\rho}$ satisfying $\mathcal{E}_\rho \subset \Xi$. To make this hold, we only need to select $\boldsymbol{\rho}$ such that the lengths of axes of each ellipse are less than the lengths of corresponding sizes of the rectangle, i.e.,

$$\frac{\rho_i + \rho_i^{-1}}{2} \leq 1 + \frac{\xi}{\sqrt{2N}}, \quad \text{and} \quad \frac{\rho_i - \rho_i^{-1}}{2} \leq \frac{\xi}{\sqrt{2N}}.$$

The choice of $\rho_i = \frac{\xi}{\sqrt{2N}} + \sqrt{\frac{\xi^2}{2N} + 1} > 1$ fulfills this condition, with which $\mathcal{E}_\rho \subset \Xi \subset \Gamma_\xi$. There follows

$$\Re(a(x, \mathbf{z})) \geq \delta$$

for all $x \in \bar{D}$, $\mathbf{z} \in \mathcal{E}_\rho$ and a satisfies **EUE**($\delta, \boldsymbol{\rho}$), as desired. \square

3.2. Analyticity of the solutions with respect to the parameters. If **DUE/EUE** holds, according to the Lax-Milgram theorem, $u(\mathbf{z}) \in V(D)$ is defined and uniformly bounded in certain polydiscs \mathcal{O}_ρ /polyellipses \mathcal{E}_ρ containing Γ . Exploiting this fact and the analyticity of $a(\mathbf{z})$ in \mathbb{C}^N , we establish the analyticity of the map $\mathbf{z} \mapsto u(\mathbf{z})$. The results given in this section generalize those in Section 2.1 of [14] to the cases of smooth, non-affine diffusion coefficients. However, we only consider model (2.1) with finite dimensional parameters. First, we state a stability result whose proof can be found in [14].

Lemma 2. *If $u(\mathbf{z})$ and $u(\tilde{\mathbf{z}})$ are weak solutions to (2.1) with coefficients $a(\mathbf{z})$ and $a(\tilde{\mathbf{z}})$, respectively, in $L^\infty(D)$ and there exists $\delta > 0$ such that*

$$\Re(a(x, \mathbf{z})) \geq \delta, \quad \Re(a(x, \tilde{\mathbf{z}})) \geq \delta,$$

for all $x \in \bar{D}$ then

$$\|u(\mathbf{z})\|_{V(D)} \leq \frac{\|f\|_{V^*(D)}}{\delta}, \tag{3.1}$$

$$\text{and} \quad \|u(\mathbf{z}) - u(\tilde{\mathbf{z}})\|_{V(D)} \leq \frac{\|f\|_{V^*(D)}}{\delta^2} \|a(\mathbf{z}) - a(\tilde{\mathbf{z}})\|_{L^\infty(D)}. \tag{3.2}$$

The analyticity of u with respect to the parameters is proved in the following theorem.

Theorem 1. *Assume that the coefficient $a(x, \mathbf{y})$ satisfies Assumptions 1–2. If **DUE**($\delta, \boldsymbol{\rho}$)*

(**EUE**($\delta, \boldsymbol{\rho}$) correspondingly) holds for some $0 < \delta < a_{\min}$ and $\boldsymbol{\rho} = (\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \ \forall i$, then the function $\mathbf{z} \mapsto u(\mathbf{z})$ is holomorphic in an open neighborhood of the polydisc $\mathcal{O}_{\boldsymbol{\rho}}$ (the polyellipse $\mathcal{E}_{\boldsymbol{\rho}}$ correspondingly).

Proof. We will prove this theorem for $a(x, \mathbf{y})$ satisfying **DUE**($\delta, \boldsymbol{\rho}$). The other case should follow similarly. Defining

$$\mathcal{A} = \left\{ \mathbf{z} \in \mathbb{C}^N : \Re(a(x, \mathbf{z})) > \frac{\delta}{2} \text{ for all } x \in \overline{D} \right\},$$

the proof consists of two steps showing that

1. $\text{int}(\mathcal{A})$ is an open neighborhood of the polydisc $\mathcal{O}_{\boldsymbol{\rho}}$, and;
2. the map $\mathbf{z} \mapsto u(\mathbf{z})$ is holomorphic in $\text{int}(\mathcal{A})$.

Here, $\text{int}(\mathcal{A})$ is the interior of \mathcal{A} .

First, let us choose an arbitrary element $\tilde{\mathbf{z}}$ in $\mathcal{O}_{\boldsymbol{\rho}}$. For $B(\mathbf{z}, r)$, we denote the open ball radius r centered at \mathbf{z} in \mathbb{C}^N . Observing that the map $\mathbf{z} \mapsto a(\mathbf{z})$ is holomorphic in \mathbb{C}^N , we have $\mathbf{z} \mapsto \Re(a(\mathbf{z}))$ is a continuous function in \mathbb{C}^N . There exists $r_{\tilde{\mathbf{z}}} > 0$ depending on $\tilde{\mathbf{z}}$ such that for all $\mathbf{z} \in B(\tilde{\mathbf{z}}, r_{\tilde{\mathbf{z}}})$,

$$\|\Re(a(\tilde{\mathbf{z}})) - \Re(a(\mathbf{z}))\|_{L^\infty(D)} < \frac{\delta}{2}.$$

This gives

$$\Re(a(x, \mathbf{z})) > \Re(a(x, \tilde{\mathbf{z}})) - \frac{\delta}{2} \geq \frac{\delta}{2}, \quad \forall x \in \overline{D}, \mathbf{z} \in B(\tilde{\mathbf{z}}, r_{\tilde{\mathbf{z}}}),$$

and $B(\tilde{\mathbf{z}}, r_{\tilde{\mathbf{z}}}) \subset \mathcal{A}$ for all $\tilde{\mathbf{z}} \in \mathcal{O}_{\boldsymbol{\rho}}$. We obtain $\tilde{\mathbf{z}} \in \text{int}(\mathcal{A})$ for all $\tilde{\mathbf{z}} \in \mathcal{O}_{\boldsymbol{\rho}}$, which concludes Step 1.

We proceed to show that $\mathbf{z} \mapsto u(\mathbf{z})$ is holomorphic in $\text{int}(\mathcal{A})$. Notice that $\forall \mathbf{z} \in \text{int}(\mathcal{A})$,

$$\frac{\delta}{2} < \Re(a(x, \mathbf{z})) \leq |a(x, \mathbf{z})| \leq \|a(\mathbf{z})\|_{L^\infty(D)}, \quad \forall x \in \overline{D},$$

so, $u(\mathbf{z})$ is well-defined in $\text{int}(\mathcal{A})$.

Now, fixing $i \in \{1, \dots, N\}$ and $\mathbf{z} \in \text{int}(\mathcal{A})$, for $h \in \mathbb{C} \setminus \{0\}$ we consider the different quotient

$$w_h(\mathbf{z}) := \frac{u(\mathbf{z} + h\mathbf{e}_i) - u(\mathbf{z})}{h} \in V(D),$$

where \mathbf{e}_i denotes the Kronecker sequence with 1 at index i and 0 at other indices. For h sufficiently small, $\mathbf{z} + h\mathbf{e}_i \in \mathcal{A}$ since $\text{int}(\mathcal{A})$ is an open set. We need to prove that $\lim_{h \rightarrow 0} w_h(\mathbf{z})$ exists, with which Hartogs' theorem would imply the analyticity of $u(\mathbf{z})$.

For all $v \in V(D)$,

$$\int_D a(\mathbf{z} + h\mathbf{e}_i) \nabla u(\mathbf{z} + h\mathbf{e}_i) \cdot \nabla v \, dx = \int_D a(\mathbf{z}) \nabla u(\mathbf{z}) \cdot \nabla v \, dx.$$

Subtracting $\int_D a(\mathbf{z}) \nabla u(\mathbf{z} + h\mathbf{e}_i) \cdot \nabla v \, dx$ and dividing both sides by h give

$$\int_D \left(\frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} \right) \nabla u(\mathbf{z} + h\mathbf{e}_i) \cdot \nabla v \, dx = - \int_D a(\mathbf{z}) \nabla w_h(\mathbf{z}) \cdot \nabla v \, dx. \quad (3.3)$$

Since $a(\mathbf{z})$ is holomorphic in \mathbb{C}^N , we can define

$$\partial_i a(\mathbf{z}) = \lim_{h \rightarrow 0} \frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} \in L^\infty(D).$$

Next, we will prove that

$$\sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \left| \int_D \left(\frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} \right) \nabla u(\mathbf{z} + h\mathbf{e}_i) \cdot \nabla v \, dx - \int_D \partial_i a(\mathbf{z}) \nabla u(\mathbf{z}) \cdot \nabla v \, dx \right| \quad (3.4)$$

converges towards 0 as $h \rightarrow 0$. Indeed,

$$\begin{aligned} & \sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \left| \int_D \left(\frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} \right) \nabla u(\mathbf{z} + h\mathbf{e}_i) \cdot \nabla v \, dx - \int_D \partial_i a(\mathbf{z}) \nabla u(\mathbf{z}) \cdot \nabla v \, dx \right| \\ & \leq \sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \int_D \left| \frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} - \partial_i a(\mathbf{z}) \right| |\nabla u(\mathbf{z} + h\mathbf{e}_i)| |\nabla v| \, dx \\ & \quad + \sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \int_D |\partial_i a(\mathbf{z})| |\nabla u(\mathbf{z} + h\mathbf{e}_i) - \nabla u(\mathbf{z})| |\nabla v| \, dx \\ & \leq \sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \left\| \frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} - \partial_i a(\mathbf{z}) \right\|_{L^\infty(D)} \|u(\mathbf{z} + h\mathbf{e}_i)\|_{V(D)} \|v\|_{V(D)} \\ & \quad + \sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \|\partial_i a(\mathbf{z})\|_{L^\infty(D)} \|u(\mathbf{z} + h\mathbf{e}_i) - u(\mathbf{z})\|_{V(D)} \|v\|_{V(D)} \\ & \leq \left\| \frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} - \partial_i a(\mathbf{z}) \right\|_{L^\infty(D)} \|u(\mathbf{z} + h\mathbf{e}_i)\|_{V(D)} \\ & \quad + \|\partial_i a(\mathbf{z})\|_{L^\infty(D)} \|u(\mathbf{z} + h\mathbf{e}_i) - u(\mathbf{z})\|_{V(D)}. \end{aligned} \quad (3.5)$$

The first term of the expression (3.5) converges towards 0 as $h \rightarrow 0$ since

$$\begin{aligned} & \left\| \frac{a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})}{h} - \partial_i a(\mathbf{z}) \right\|_{L^\infty(D)} \rightarrow 0 \text{ (from the definition of } \partial_i a(\mathbf{z}) \text{),} \\ & \|u(\mathbf{z} + h\mathbf{e}_i)\|_{V(D)} \leq \frac{2\|f\|_{V^*(D)}}{\delta} \text{ (from the fact that } \mathbf{z} + h\mathbf{e}_i \in \mathcal{A} \text{ and (3.1)).} \end{aligned}$$

On the other hand, since $\mathbf{z}, \mathbf{z} + h\mathbf{e}_i \in \mathcal{A}$ and from (3.2), it gives

$$\|u(\mathbf{z} + h\mathbf{e}_i) - u(\mathbf{z})\|_{V(D)} \leq \frac{4\|f\|_{V^*(D)}}{\delta^2} \|a(\mathbf{z} + h\mathbf{e}_i) - a(\mathbf{z})\|_{L^\infty(D)}.$$

Note that $a(\mathbf{z})$ is continuous in \mathbf{z} , we have $\|u(\mathbf{z} + h\mathbf{e}_i) - u(\mathbf{z})\|_{V(D)} \rightarrow 0$ as $h \rightarrow 0$ and so does the second term of (3.5). This concludes the convergence towards 0 of (3.4).

Let $w_0(\mathbf{z}) \in V(D)$ be the solution to

$$\int_D a(\mathbf{z}) \nabla w_0(\mathbf{z}) \cdot \nabla v \, dx = - \int_D \partial_i a(\mathbf{z}) \nabla u(\mathbf{z}) \cdot \nabla v \, dx, \quad \forall v \in V(D). \quad (3.6)$$

Such $w_0(\mathbf{z})$ exists since $L(v) := - \int_D \partial_i a(\mathbf{z}) \nabla u(\mathbf{z}) \cdot \nabla v \, dx$ is a continuous and linear functional in $V(D)$.

Substituting (3.3) and (3.6) to (3.4), we have

$$\sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \left| \int_D a(\mathbf{z}) \nabla w_h(\mathbf{z}) \cdot \nabla v \, dx - \int_D a(\mathbf{z}) \nabla w_0(\mathbf{z}) \cdot \nabla v \, dx \right| \rightarrow 0 \text{ as } h \rightarrow 0.$$

For h such that $w_h(\mathbf{z}) \neq w_0(\mathbf{z})$, denote the complex conjugate of $w_h(\mathbf{z}) - w_0(\mathbf{z})$ by $\overline{w_h(\mathbf{z}) - w_0(\mathbf{z})}$. We observe that

$$\begin{aligned} \frac{\delta}{2} \|w_h(\mathbf{z}) - w_0(\mathbf{z})\|_{V(D)} &\leq \left| \int_D a(\mathbf{z}) \nabla (w_h(\mathbf{z}) - w_0(\mathbf{z})) \cdot \frac{\nabla \overline{(w_h(\mathbf{z}) - w_0(\mathbf{z}))}}{\|w_h(\mathbf{z}) - w_0(\mathbf{z})\|_{V(D)}} \, dx \right| \\ &\leq \sup_{\substack{v \in V(D) \\ \|v\|_{V(D)} \leq 1}} \left| \int_D a(\mathbf{z}) \nabla (w_h(\mathbf{z}) - w_0(\mathbf{z})) \cdot \nabla v \, dx \right|. \end{aligned}$$

There follows $\lim_{h \rightarrow 0} w_h(\mathbf{z}) = w_0(\mathbf{z})$ in $V(D)$, as desired. \square

3.3. Estimates of the polynomial coefficients. Under the analyticity properties established in Theorem 1, the convergence of Taylor and Legendre expansions of the solutions, as well as estimates of the expansion coefficients, are well-studied, e.g., in [6, 12, 14]. In this subsection, we revisit those results in the context of finite-dimensional, possibly non-affine parametric coefficients. Recall that $\mathcal{S} = \{\boldsymbol{\nu} = (\nu_i)_{1 \leq i \leq N} : \nu_i \in \mathbb{N}\}$. For all $\boldsymbol{\nu} \in \mathcal{S}$, we introduce the multivariate notations $|\boldsymbol{\nu}| := \sum_{1 \leq i \leq N} \nu_i$, $\boldsymbol{\nu}! := \prod_{1 \leq i \leq N} \nu_i!$ and define the partial derivative $\partial^{\boldsymbol{\nu}} u := \frac{\partial^{|\boldsymbol{\nu}|} u}{\partial \nu_1 z_1 \dots \partial \nu_N z_N}$. The Taylor series of $u(\mathbf{y})$ reads

$$u(\mathbf{y}) = \sum_{\boldsymbol{\nu} \in \mathcal{S}} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}}, \quad (3.7)$$

where the coefficients $t_{\boldsymbol{\nu}} \in V(D)$ are defined as

$$t_{\boldsymbol{\nu}} := \frac{1}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}} u(0), \quad \boldsymbol{\nu} \in \mathcal{S}.$$

The convergence of the Taylor expansion in Γ and estimates of $\|t_{\boldsymbol{\nu}}\|_{V(D)}$ are given in the following result.

Proposition 1. *Assume that the coefficient $a(x, \mathbf{y})$ satisfies Assumptions 1–2. If $\mathbf{DUE}(\delta, \boldsymbol{\rho})$ holds for some $0 < \delta < a_{\min}$ and $\boldsymbol{\rho} = (\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \, \forall i$ then the Taylor series*

$\sum_{\nu \in \mathcal{S}} t_\nu \mathbf{y}^\nu$ converges uniformly towards $u(\mathbf{y})$ in Γ . Furthermore, we have the estimate

$$\|t_\nu\|_{V(D)} \leq \frac{\|f\|_{V^*(D)}}{\delta} \boldsymbol{\rho}^{-\nu}. \quad (3.8)$$

Proof. We can actually prove that the complex power series $\sum_{\nu \in \mathcal{S}} t_\nu \mathbf{z}^\nu$ converges uniformly towards $u(\mathbf{z})$ in $\mathcal{O}_\rho \supset \Gamma$. Indeed, from the proof of Theorem 1, $\mathbf{z} \mapsto u(\mathbf{z})$ is a holomorphic function in an open neighborhood $\text{int}(\mathcal{A})$ of the polydisc \mathcal{O}_ρ . By standard results on analyticity of Hilbert-valued functions, see, e.g., Section 2.1 in [24], this implies the Taylor series of $u(\mathbf{z})$ is uniformly summable in \mathcal{O}_ρ .

The estimate (3.8) of $\|t_\nu\|_{V(D)}$ can be obtained by a Cauchy's integral formula argument as in Lemma 2.4 in [14]. \square

On the other hand, the tensorized Legendre series of $u(\mathbf{y})$ is defined as

$$u(\mathbf{y}) = \sum_{\nu \in \mathcal{S}} u_\nu P_\nu(\mathbf{y}), \quad (3.9)$$

where $P_\nu(\mathbf{y}) = \prod_{i=1}^N P_{\nu_i}(y_i)$, with P_{ν_i} denoting the monodimensional Legendre polynomials of degree ν_i according to L^∞ normalization $\|P_{\nu_i}\|_{L^\infty([-1,1])} = P_{\nu_i}(1) = 1$.

A second type of Legendre expansion, which employs the L^2 normalized version of P_ν , is also considered. Denote the multivariate polynomials $L_\nu(\mathbf{y}) = \prod_{i=1}^N L_{\nu_i}(y_i)$, with $L_{\nu_i}(y_i)$ given by

$$L_{\nu_i}(y_i) := \sqrt{2\nu_i + 1} P_{\nu_i}(y_i).$$

The Legendre series in this case can be written as

$$u(\mathbf{y}) = \sum_{\nu \in \mathcal{S}} v_\nu L_\nu(\mathbf{y}). \quad (3.10)$$

We note that the coefficients $u_\nu, v_\nu \in V(D)$ are defined by

$$v_\nu = \int_\Gamma u(\mathbf{y}) L_\nu(\mathbf{y}) \varrho(\mathbf{y}) d\mathbf{y} \quad \text{and} \quad u_\nu = v_\nu \left(\prod_{i=1}^N (2\nu_i + 1) \right)^{1/2}. \quad (3.11)$$

The following proposition establishes estimates of $\|u_\nu\|_{V(D)}$, $\|v_\nu\|_{V(D)}$ and the convergence of the Legendre expansions of u in Γ .

Proposition 2. *Assume that the coefficient $a(x, \mathbf{y})$ satisfies Assumptions 1–2. If $\mathbf{EUE}(\delta, \boldsymbol{\rho})$ holds for some $0 < \delta < a_{\min}$ and $\boldsymbol{\rho} = (\rho_i)_{1 \leq i \leq N}$ with $\rho_i > 1 \ \forall i$ then we have the estimates*

$$\|u_\nu\|_{V(D)} \leq C_{\boldsymbol{\rho}, \delta} \boldsymbol{\rho}^{-\nu} \prod_{i=1}^N (2\nu_i + 1), \quad \|v_\nu\|_{V(D)} \leq C_{\boldsymbol{\rho}, \delta} \boldsymbol{\rho}^{-\nu} \prod_{i=1}^N \sqrt{2\nu_i + 1}, \quad (3.12)$$

where $C_{\boldsymbol{\rho}, \delta} = \frac{\|f\|_{V^*(D)}}{\delta} \prod_{i=1}^N \frac{\ell(\mathcal{E}_{\rho_i})}{4(\rho_i - 1)}$ with $\ell(\mathcal{E}_{\rho_i})$ denoting the perimeter of the ellipse \mathcal{E}_{ρ_i} .

Consequently, the Legendre series $\sum_{\nu \in \mathcal{S}} u_\nu P_\nu$ and $\sum_{\nu \in \mathcal{S}} v_\nu L_\nu$ converge towards u in $L^\infty(\Gamma, V(D))$. The series $\sum_{\nu \in \mathcal{S}} v_\nu L_\nu$ also converges towards u in $V(D) \otimes L^2_q(\Gamma)$.

Proof. From Theorem 1, we have $\mathbf{z} \mapsto u(\mathbf{z})$ is a holomorphic function in an open neighborhood $\text{int}(\mathcal{A})$ of the polyellipse \mathcal{E}_ρ . Using this fact, the proof of estimates (3.12) is similar to those of Lemma 4.2 in [14] and Proposition 8 in [6], based on an application of Cauchy's integral formula, and we shall not repeat it here.

Next, we note that for any finite set $\Lambda \subset \mathcal{S}$,

$$\begin{aligned} \left\| u - \sum_{\nu \in \Lambda} u_\nu P_\nu \right\|_{L^\infty(\Gamma, V(D))} &= \sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\nu \in \Lambda} u_\nu P_\nu(\mathbf{y}) \right\|_{V(D)} \leq \sum_{\nu \notin \Lambda} \|u_\nu\|_{V(D)}, \\ \left\| u - \sum_{\nu \in \Lambda} v_\nu L_\nu \right\|_{V(D) \otimes L^2_\rho(\Gamma)} &= \left(\sum_{\nu \notin \Lambda} \|v_\nu\|_{V(D)}^2 \right)^{1/2}. \end{aligned}$$

Therefore, to obtain the convergence of the Legendre expansions $\sum_{\nu \in \mathcal{S}} u_\nu P_\nu$ and $\sum_{\nu \in \mathcal{S}} v_\nu L_\nu$ in $L^\infty(\Gamma, V(D))$ and $V(D) \otimes L^2_\rho(\Gamma)$ respectively, it is enough to prove the ℓ^1 summability of $(\|u_\nu\|_{V(D)})_{\nu \in \mathcal{S}}$ and ℓ^2 summability of $(\|v_\nu\|_{V(D)})_{\nu \in \mathcal{S}}$. These can be done using the estimates in (3.12) and the argument in Theorem 4.1 in [14]. \square

Under Assumptions 1-2, we remark that **EUE** and, if adding affine dependence on parameters, **DUE** normally hold for infinitely many couples of (δ, ρ) . We call the set of all (δ, ρ) such that **EUE** (δ, ρ) /**DUE** (δ, ρ) is fulfilled the admissible set and denote it by **Ad** for both cases. For a fixed $\nu \in \mathcal{S}$, the best coefficient bounds given by Propositions 1 and 2 will be

$$\|t_\nu\|_{V(D)} \leq \inf_{(\delta, \rho) \in \text{Ad}} \frac{\|f\|_{V^*(D)}}{\delta} \rho^{-\nu}, \quad \|u_\nu\|_{V(D)} \leq \inf_{(\delta, \rho) \in \text{Ad}} C_{\rho, \delta} \rho^{-\nu} \prod_{i=1}^N (2\nu_i + 1). \quad (3.13)$$

Finding an efficient computation of these infimums and algorithm to construct the corresponding quasi-optimal index sets is an open question. In the specific case where the basis functions ψ_i have non-overlapping supports, however, the vectors ρ solving the minimization problems in (3.13) can be found easily. In this case, the best a priori estimates retrieve the forms (3.8) and (3.12). Recent studies have shown that although these theoretical bounds are not sharp, they construct correct quasi-optimal polynomial spaces, see [6].

4. Asymptotic convergence analysis for a general class of multi-indexed series. In this section, we introduce a new, generalized approach to estimating the asymptotic convergence of a class of multi-indexed series, which is relevant to quasi-optimal approximation settings and accommodates most types of Taylor and Legendre coefficient bounds established in current literature. Consider a multi-indexed sequence (of coefficient estimates) written in the form $(e^{-b(\nu)})_{\nu \in \mathcal{S}}$. Recalling that Λ_M is the set of indices ν corresponding to the M largest $e^{-b(\nu)}$ and Λ_M^c denotes the complement of Λ_M in \mathcal{S} , we are interested in finding a sharp convergence rate of $\sum_{\nu \in \Lambda_M^c} e^{-b(\nu)}$ with respect to M . It is enough to analyze this sum with Λ_M^c being the sets of all ν such that $e^{-b(\nu)} < e^{-J}$ with some $J \in \mathbb{N}$.

Our method can be summarized as follows. First, we split Λ_M^c into a family $(\mathcal{Q}_j)_{j \in \mathbb{N}, j \geq J}$ of disjoint subsets of \mathcal{S} based on values of $e^{-b(\nu)}$, where \mathcal{Q}_j contains ν satisfying $e^{-j-1} \leq$

$e^{-b(\boldsymbol{\nu})} < e^{-j}$, so that the truncation error can be bounded as

$$\sum_{\boldsymbol{\nu} \in \Lambda_M^c} e^{-b(\boldsymbol{\nu})} = \sum_{j \geq J} \sum_{\boldsymbol{\nu} \in \mathcal{Q}_j} e^{-b(\boldsymbol{\nu})} \leq \sum_{j \geq J} \#(\mathcal{Q}_j) \cdot e^{-j}. \quad (4.1)$$

Obviously, finding a sharp approximation of $\#(\mathcal{Q}_j)$ is central to estimate (4.1). We define the *superlevel sets* \mathcal{P}_j of N -dimensional real points

$$\mathcal{P}_j := \{\boldsymbol{\nu} \in [0, \infty)^N : e^{-b(\boldsymbol{\nu})} \geq e^{-j}\} = \{\boldsymbol{\nu} \in [0, \infty)^N : b(\boldsymbol{\nu}) \leq j\}, \quad (4.2)$$

and, with notice that $\#(\mathcal{Q}_j) = \#(\mathcal{P}_{j+1} \cap \mathbb{Z}^N) - \#(\mathcal{P}_j \cap \mathbb{Z}^N)$, seek to count points with integer coordinates in \mathcal{P}_j . An appealing approach to solving this problem is to study the interplay between $\#(\mathcal{P}_j \cap \mathbb{Z}^N)$ and the continuous volume (Lebesgue measure) of \mathcal{P}_j . We first employ the following well-known result in measure theory, reflecting the intuitive fact that for a geometric body \mathcal{P} in \mathbb{R}^N , the volume of \mathcal{P} , denoted by $|\mathcal{P}|$, can be approximated by the number of shrunken integer points inside \mathcal{P} , see, e.g., Section 7.2 in [21] and Section 1.1 in [34].

Lemma 3. *Suppose $\mathcal{P} \subset \mathbb{R}^N$ is a bounded Jordan measurable set. For $j \in \mathbb{N}$, $j > 0$, there holds*

$$|\mathcal{P}| = \lim_{j \rightarrow \infty} \frac{1}{j^N} \cdot \#(\mathcal{P} \cap \frac{1}{j} \mathbb{Z}^N) = \lim_{j \rightarrow \infty} \frac{1}{j^N} \cdot \#(j\mathcal{P} \cap \mathbb{Z}^N). \quad (4.3)$$

Concerning our goal of estimating (4.1), Lemma 3 has an interesting consequence: If $b(\boldsymbol{\nu})$ is defined such that $\frac{1}{j}\mathcal{P}_j = \mathcal{P}$, $\forall j \in \mathbb{N}$, with some $\mathcal{P} \subset \mathbb{R}^N$, one obtains a simple asymptotic formula for $\#(\mathcal{P}_j \cap \mathbb{Z}^N)$:

$$\#(\mathcal{P}_j \cap \mathbb{Z}^N) \simeq j^N |\mathcal{P}|. \quad (4.4)$$

Such approximation is powerful since, loosely speaking, it would allow replacing $\#(\mathcal{Q}_j)$ by $((j+1)^N - j^N)|\mathcal{P}|$ and reduce (4.1) to a much easier, yet equivalent problem of estimating the truncation error via

$$\sum_{\boldsymbol{\nu} \in \Lambda_M^c} e^{-b(\boldsymbol{\nu})} \lesssim \sum_{j \geq J} |\mathcal{P}| ((j+1)^N - j^N) e^{-j}. \quad (4.5)$$

The property that the sets $\frac{1}{j}\mathcal{P}_j$ are unchanged over $j \in \mathbb{N}$ is, however, restrictive, corresponding to only a few types of coefficient upper bounds, for instance, $b(\boldsymbol{\nu})$ is linear in $\boldsymbol{\nu}$. For this approach of estimation to be considered useful in general quasi-optimal approximation setting, this condition needs to be relaxed.

For the technicality, we now extend definition (4.2) to equip the superlevel sets with real indices: for $\tau \in (0, \infty)$, define

$$\mathcal{P}_\tau := \{\boldsymbol{\nu} \in [0, \infty)^N : e^{-b(\boldsymbol{\nu})} \geq e^{-\tau}\} = \{\boldsymbol{\nu} \in [0, \infty)^N : b(\boldsymbol{\nu}) \leq \tau\}. \quad (4.6)$$

Note that the assertion of Lemma 3 still holds if replacing $j \in \mathbb{N}$ by $\tau \in (0, \infty)$. We establish, in Lemma 4 below, formula (4.4) under some weaker assumptions on $(\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$:

- i) \mathcal{P}_τ is Jordan measurable for countably infinite $\tau \in (0, \infty)$,
- ii) The chain $(\frac{1}{\tau}\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$ is either ascending or descending towards a Jordan measurable limiting set $\mathcal{P} \subset \mathbb{R}^N$ with $0 < |\mathcal{P}| < \infty$.

As we shall see later, these properties are satisfied by most existing polynomial coefficient estimates.

Lemma 4. *Suppose $(\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$ is a family of bounded Lebesgue measurable sets in \mathbb{R}^N satisfying either*

$$\frac{1}{\tau_1}\mathcal{P}_{\tau_1} \subset \frac{1}{\tau_2}\mathcal{P}_{\tau_2}, \quad \forall \tau_1 \geq \tau_2 > 0, \quad (4.7)$$

$$\text{or} \quad \frac{1}{\tau_1}\mathcal{P}_{\tau_1} \supset \frac{1}{\tau_2}\mathcal{P}_{\tau_2}, \quad \forall \tau_1 \geq \tau_2 > 0. \quad (4.8)$$

Denote $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} \frac{1}{\tau}\mathcal{P}_\tau$ if (4.7) holds and $\mathcal{P} = \bigcup_{\tau \in \mathbb{R}^+} \frac{1}{\tau}\mathcal{P}_\tau$ for the other case. If \mathcal{P} is bounded Jordan measurable, $|\mathcal{P}| > 0$, and there exists a sequence $(\tau_j)_{j \in \mathbb{N}}$ with $\tau_j \rightarrow \infty$ such that \mathcal{P}_{τ_j} is Jordan measurable for all j , there follows

$$|\mathcal{P}| = \lim_{\tau \rightarrow \infty} \frac{1}{\tau^N} \cdot \#(\mathcal{P}_\tau \cap \mathbb{Z}^N). \quad (4.9)$$

Proof. We will give a proof with $(\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$ satisfying (4.7). The other case can be shown analogously. Let ε be an arbitrary positive number. By Lemma 3,

$$\frac{1}{\tau^N} \cdot \#(\tau\mathcal{P} \cap \mathbb{Z}^N) \rightarrow |\mathcal{P}| \quad \text{as } \tau \rightarrow \infty.$$

Since $\mathcal{P} \subset \frac{1}{\tau}\mathcal{P}_\tau \forall \tau$, we can choose $T_1 > 0$ such that $\forall \tau > T_1$,

$$|\mathcal{P}| - \varepsilon \leq \frac{1}{\tau^N} \cdot \#(\tau\mathcal{P} \cap \mathbb{Z}^N) \leq \frac{1}{\tau^N} \cdot \#(\mathcal{P}_\tau \cap \mathbb{Z}^N). \quad (4.10)$$

On the other hand, from $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} \frac{1}{\tau}\mathcal{P}_\tau$, it yields $|\mathcal{P}| = \lim_{\tau \rightarrow \infty} \left| \frac{1}{\tau}\mathcal{P}_\tau \right|$. Let us pick an $L > 0$ so that \mathcal{P}_L is Jordan measurable and $\left| \frac{1}{L}\mathcal{P}_L \right| \leq |\mathcal{P}| + \frac{\varepsilon}{2}$. By Lemma 3,

$$\frac{1}{\tau^N} \cdot \# \left(\frac{\tau}{L}\mathcal{P}_L \cap \mathbb{Z}^N \right) \rightarrow \left| \frac{\mathcal{P}_L}{L} \right| \quad \text{as } \tau \rightarrow \infty.$$

There exists $T_2 > L$ satisfying $\forall \tau > T_2$,

$$\frac{1}{\tau^N} \cdot \# \left(\frac{\tau}{L}\mathcal{P}_L \cap \mathbb{Z}^N \right) \leq \left| \frac{\mathcal{P}_L}{L} \right| + \frac{\varepsilon}{2} \leq |\mathcal{P}| + \varepsilon.$$

Since $\tau > L$, we have $\mathcal{P}_\tau \subset \frac{\tau}{L}\mathcal{P}_L$, which gives

$$\frac{1}{\tau^N} \cdot \#(\mathcal{P}_\tau \cap \mathbb{Z}^N) \leq \frac{1}{\tau^N} \cdot \# \left(\frac{\tau}{L}\mathcal{P}_L \cap \mathbb{Z}^N \right) \leq |\mathcal{P}| + \varepsilon. \quad (4.11)$$

Combining (4.10) and (4.11) proves (4.9). \square

Lemma 4 provides us with an asymptotic formula of the form (4.4) to approximate the number of integer points inside \mathcal{P}_τ , under some conditions on $(\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$. Given a coefficient upper bound $e^{-b(\boldsymbol{\nu})}$, it is desirable to derive properties of $b(\boldsymbol{\nu})$ such that its corresponding superlevel sets $(\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$ fulfill these conditions. For all $\boldsymbol{\nu} \in [0, \infty)^N$, define the map $H_{\boldsymbol{\nu}} : (0, \infty) \rightarrow \mathbb{R}$ as

$$H_{\boldsymbol{\nu}}(\tau) = \frac{1}{\tau} b(\tau \boldsymbol{\nu}), \forall \tau \in (0, \infty).$$

We proceed to state and validate the following assumptions on $b(\boldsymbol{\nu})$.

Assumption 3. *The map $b : [0, \infty)^N \rightarrow \mathbb{R}$ satisfies*

1. $b(\mathbf{0}) = 0$ and b is continuous in $[0, \infty)^N$,
2. $H_{\boldsymbol{\nu}}$ is either increasing in $(0, \infty)$ for all $\boldsymbol{\nu} \in [0, \infty)^N$ or decreasing in $(0, \infty)$ for all $\boldsymbol{\nu} \in [0, \infty)^N$,
3. $b(\boldsymbol{\nu}) \in \Theta(|\boldsymbol{\nu}|)$. In other words, there exists $0 < c < C$ such that $c|\boldsymbol{\nu}| < b(\boldsymbol{\nu}) < C|\boldsymbol{\nu}|$ as $\boldsymbol{\nu} \rightarrow \infty$.

Lemma 5. *Assume that $b : [0, \infty)^N \rightarrow \mathbb{R}$ satisfies Assumption 3. For $\tau \in (0, \infty)$, denote $\mathcal{P}_\tau = \{\boldsymbol{\nu} \in [0, \infty)^N : b(\boldsymbol{\nu}) \leq \tau\}$. Let*

$$\mathcal{P} = \begin{cases} \bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau) & \text{if } H_{\boldsymbol{\nu}} \text{ is increasing } \forall \boldsymbol{\nu} \in [0, \infty)^N, \\ \bigcup_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau) & \text{if } H_{\boldsymbol{\nu}} \text{ is decreasing } \forall \boldsymbol{\nu} \in [0, \infty)^N. \end{cases} \quad (4.12)$$

Then, $0 < |\mathcal{P}| < \infty$. If \mathcal{P} is Jordan measurable, there holds

$$|\mathcal{P}| = \lim_{\tau \rightarrow \infty} \frac{1}{\tau^N} \cdot \#(\mathcal{P}_\tau \cap \mathbb{Z}^N). \quad (4.13)$$

Proof. From the continuity of b in $[0, \infty)^N$ (Assumption 3.1), \mathcal{P}_τ is Jordan measurable for all except a countable number of values of τ (see [19]).

Next, from Assumption 3.2, if $H_{\boldsymbol{\nu}}$ is increasing for all $\boldsymbol{\nu}$, one has $\frac{1}{\tau_2} b(\tau_2 \boldsymbol{\nu}) \leq \frac{1}{\tau_1} b(\tau_1 \boldsymbol{\nu}) \leq 1$, $\forall \tau_1 \geq \tau_2 > 0$, $\forall \boldsymbol{\nu} \in [0, \infty)^N$, which implies

$$\frac{1}{\tau_1} \mathcal{P}_{\tau_1} \subset \frac{1}{\tau_2} \mathcal{P}_{\tau_2}, \forall \tau_1 \geq \tau_2 > 0.$$

Since $b(\boldsymbol{\nu})$ converges towards $+\infty$ as $\boldsymbol{\nu} \rightarrow \infty$, \mathcal{P}_τ is bounded for every $\tau \in (0, \infty)$. It is trivial that $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau)$ is bounded. Let $\boldsymbol{\nu} \notin \mathcal{P}$, we have $\boldsymbol{\nu} \notin \frac{1}{\tau} \mathcal{P}_\tau$ for τ large enough. Combining with Assumption 3.3 yields $C\tau|\boldsymbol{\nu}| > b(\tau \boldsymbol{\nu}) > \tau$. Thus, $B(\mathbf{0}, 1/C) \subset \mathcal{P}$ and $|\mathcal{P}| > 0$.

If, on the other hand, $H_{\boldsymbol{\nu}}$ is decreasing for all $\boldsymbol{\nu}$, then $\frac{1}{\tau_1} b(\tau_1 \boldsymbol{\nu}) \leq \frac{1}{\tau_2} b(\tau_2 \boldsymbol{\nu}) \leq 1$, $\forall \tau_1 \geq \tau_2 > 0$, $\forall \boldsymbol{\nu} \in [0, \infty)^N$, which gives

$$\frac{1}{\tau_1} \mathcal{P}_{\tau_1} \supset \frac{1}{\tau_2} \mathcal{P}_{\tau_2}, \forall \tau_1 \geq \tau_2 > 0.$$

Since $\mathcal{P} = \bigcup_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau)$, it is trivial that $|\mathcal{P}| > 0$. Furthermore, for any $\nu \in \mathcal{P}$, $b(\tau\nu) \leq \tau$ with τ large enough. Combining with Assumption 3.3 that $b(\tau\nu) > c\tau|\nu|$, this implies $|\nu| < \frac{1}{c}$. Thus, $\mathcal{P} \subset B(\mathbf{0}, 1/c)$ and $|\mathcal{P}| < \infty$.

If \mathcal{P} Jordan measurable, since the family $(\mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$ has been proved to satisfy the conditions of Lemma 4, we can apply this to get (4.13). \square

As seen in the proof, the continuity of b (Assumption 3.1) assures that the superlevel sets \mathcal{P}_τ are “well-behaved” (Jordan measurable). Meanwhile, the monotonicity of H_ν (Assumption 3.2) leads to the ascending (or descending) property of the chain $(\frac{1}{\tau} \mathcal{P}_\tau)_{\tau \in \mathbb{R}^+}$. To guarantee the limiting set \mathcal{P} is bounded and not null, we assume $c|\nu| < b(\nu) < C|\nu|$ for some $0 < c < C$ (Assumption 3.3), so that $B(\mathbf{0}, 1/C) \subset \mathcal{P} \subset B(\mathbf{0}, 1/c)$. It should be noted that c and C are generic constants, which are only utilized to represent the boundedness of \mathcal{P} and do not affect our convergence rate, thus a specification of c and C is not necessary. In the subsequent analysis, we applies (4.13) to derive an error estimate, only depending on \mathcal{P} and the parameter dimension N , of the form $M \exp(-(M/|\mathcal{P}|)^{1/N})$. This rate is consistent with the proven sub-exponential convergence $M \exp(-(\kappa M)^{1/N})$ for some simple coefficient upper bounds [6]. Nevertheless, our analysis completely exploits information on the size and shape of the (possibly complicated) index sets in the asymptotic regime via the introduction of \mathcal{P} and, as a result, acquires the optimal value of κ .

It is worth remarking that Lemma 4 requires \mathcal{P} to be Jordan measurable. Indeed, we show here a simple counterexample in which \mathcal{P} is not Jordan measurable and (4.13) fails to hold. Consider the integer-indexed collection of Jordan measurable sets $(\mathcal{P}_j)_{j \in \mathbb{N}}$ defined by

$$\mathcal{P}_j = j \left([0, 1] \setminus \{p/q : p, q \in \mathbb{Z}, 0 \leq p \leq q \leq j\} \right).$$

Observing that $(\frac{1}{j} \mathcal{P}_j)_{j \in \mathbb{N}}$ is descending towards $\mathcal{P} = [0, 1] \setminus \mathbb{Q}$, which is not Jordan measurable. We have $\#(\mathcal{P}_j \cap \mathbb{Z}^N) = 0 \ \forall j$ while $|\mathcal{P}| = 1$, contradictory to (4.13). The conditions on Jordan measurability of \mathcal{P} is, however, not restrictive in the context of quasi-optimal methods, since the shapes of limiting sets are often not very fractal. Indeed, all examples investigated herein show the convexity of \mathcal{P} , which trivially implies its Jordan measurability, as required.

The mathematical evidence that Assumption 3 is satisfied by published Taylor and Legendre coefficient estimates will be presented in Section 6. Four examples of upper bounds $e^{-b(\nu)}$ will be considered, including $\rho^{-\nu}$ (as in (3.8)), $\inf_{(\delta, \rho) \in \mathcal{A}d} (\frac{\rho^{-\nu}}{\delta})$ (as in (3.13)), $\rho^{-\nu} \prod_{i=1}^N \sqrt{2\nu_i + 1}$ (as in (3.12)), and $\frac{|\nu|!}{\nu!} \alpha^\nu$ (as in [7, 13]). For now, with Lemma 4 giving an approximation for $\#(\mathcal{P}_j \cap \mathbb{Z}^N)$, it remains to study the estimation problem (4.5). We proceed to prove the following supporting result.

Lemma 6. *For any $N, J, L \in \mathbb{N}$, if $J \geq \max \left\{ \frac{1}{e^{1/N} - 1}, \frac{L}{e^{(L-1)/N} - 1} \right\}$, it gives*

$$\sum_{j \geq J} j^N e^{-j} \leq L J^N e^{-J} \frac{e}{e-1}. \quad (4.14)$$

Particularly,

$$\sum_{j \geq J} j^N e^{-j} \leq 2J^N e^{-J} \frac{e}{e-1}, \quad \forall J \geq \frac{2}{e^{1/N} - 1}, \quad (4.15)$$

$$\sum_{j \geq J} j^N e^{-j} \leq (N+1)J^N e^{-J} \frac{e}{e-1}, \quad \forall J \geq \frac{1}{e^{1/N} - 1}, N \geq 4. \quad (4.16)$$

Proof. We have

$$\frac{1}{J^N} \sum_{j \geq J} j^N e^{-j} = \sum_{k \geq 0} \sum_{\ell=0}^{L-1} \left[\left(1 + \frac{Lk + \ell}{J} \right)^N e^{-J-Lk-\ell} \right]. \quad (4.17)$$

We prove that for every $k \geq 0$, $0 \leq \ell \leq L-1$,

$$\left(1 + \frac{Lk + \ell}{J} \right)^N e^{-J-Lk-\ell} \leq e^{-J-k}. \quad (4.18)$$

Consider $\ell = 0$. If $k = 0$, (4.18) holds trivially. If $k > 0$, it is equivalent to

$$\left(1 + \frac{Lk}{J} \right)^N \leq e^{(L-1)k}, \quad \text{or} \quad J \geq \frac{Lk}{e^{(L-1)k/N} - 1},$$

which is true since $J \geq \frac{L}{e^{(L-1)/N} - 1}$.

Now, for $\ell > 0$, observe that

$$\begin{aligned} \left(1 + \frac{Lk + \ell}{J} \right)^N &= \left(1 + \frac{Lk}{J} \right)^N \left(1 + \frac{\ell}{Lk + J} \right)^N \\ &\leq e^{(L-1)k} \left(1 + \frac{\ell}{J} \right)^N \leq e^{(L-1)k + \ell}, \end{aligned}$$

since $J \geq \frac{1}{e^{1/N} - 1}$ and (4.18) follows.

Combining (4.17) and (4.18) gives

$$\frac{1}{J^N} \sum_{j \geq J} j^N e^{-j} \leq L \sum_{j \geq J} e^{-j} = L e^{-J} \frac{e}{e-1},$$

which yields (4.14).

(4.15) can be obtained from (4.14) with $L = 2$. For (4.16), applying (4.14) with $L = N + 1$, we only need to verify $\frac{1}{e^{1/N} - 1} \geq \frac{N+1}{e-1}$. We have

$$\frac{e-1}{e^{1/N} - 1} = \sum_{i=0}^{N-1} e^{i/N} \geq N + 1,$$

since $e^{(N-1)/N} \geq 1.5$ and $e^{(N-2)/N} \geq 1.5$ for $N \geq 4$, and $e^{i/N} \geq 1$ for all $0 \leq i \leq N-3$,

proving (4.16). □

It is easy to see that $\sum_{j \geq J} j^N e^{-j}$ is also bounded from below by

$$\sum_{j \geq J} j^N e^{-j} \geq J^N \sum_{j \geq J} e^{-j} \geq J^N e^{-J} \frac{e}{e-1}, \quad \forall J \in \mathbb{N},$$

verifying the sharpness of estimate (4.14). This sub-exponential convergence rate, however, is effective with $J \geq \frac{1}{e^{1/N}-1} \simeq N$. Since $J^N e^{-J}$ is increasing with respect to J for $J < N$, this seems not an appropriate rate to describe the decay of $\sum_{j \geq J} j^N e^{-j}$ in the pre-asymptotic regime.

We are now ready to analyze the asymptotic truncation error of the general multi-indexed series $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ relevant to quasi-optimal Taylor and Legendre approximations. The main result of this section is stated and proved below.

Theorem 2. *Consider the multi-indexed series $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ with $b : [0, \infty)^N \rightarrow \mathbb{R}$ satisfying Assumption 3. For $\tau \in (0, \infty)$, denote $\mathcal{P}_\tau = \{\nu \in [0, \infty)^N : b(\nu) \leq \tau\}$ and Λ_M the set of indices corresponding to M largest $e^{-b(\nu)}$. Define $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau)$ or $\mathcal{P} = \bigcup_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau)$ as in (4.12). If \mathcal{P} is Jordan measurable, for any $\varepsilon > 0$, there exists $M_\varepsilon > 0$ depending on ε such that*

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \leq C_u(\varepsilon) M \exp\left(-\left(\frac{M}{|\mathcal{P}|(1+\varepsilon)}\right)^{1/N}\right) \quad (4.19)$$

for all $M > M_\varepsilon$. Here, $C_u(\varepsilon) = (4e + 4\varepsilon e - 2) \frac{e}{e-1}$.

Proof. We apply Lemma 5 to get

$$|\mathcal{P}| = \lim_{\tau \rightarrow \infty} \frac{1}{\tau^N} \cdot \#(\mathcal{P}_\tau \cap \mathbb{Z}^N).$$

For a fixed $\varepsilon > 0$, there exists $\Delta_\varepsilon > 0$ such that for all integer $j > \Delta_\varepsilon$,

$$\begin{aligned} -\varepsilon |\mathcal{P}| &\leq |\mathcal{P}| - \frac{1}{j^N} \cdot \#(\mathcal{P}_j \cap \mathbb{Z}^N) \leq \frac{1}{2} |\mathcal{P}|, \\ \text{i.e., } \frac{1}{2} j^N |\mathcal{P}| &\leq \#(\mathcal{P}_j \cap \mathbb{Z}^N) \leq j^N |\mathcal{P}| (1 + \varepsilon). \end{aligned} \quad (4.20)$$

To analyze the asymptotic convergence of $\sum_{\nu \notin \Lambda_M} e^{-b(\nu)}$, it is sufficient to consider this sum with $\Lambda_M = \mathcal{P}_J \cap \mathbb{Z}^N$, $J \in \mathbb{N}$. First, observe that for all integer $J > \Delta_\varepsilon$ and $J \geq \frac{1}{e^{1/N}-1}$, from (4.20),

$$\begin{aligned} \sum_{\nu \notin \mathcal{P}_J \cap \mathbb{Z}^N} e^{-b(\nu)} &\leq \sum_{j \geq J} (\#(\mathcal{P}_{j+1} \cap \mathbb{Z}^N) - \#(\mathcal{P}_j \cap \mathbb{Z}^N)) e^{-j} \\ &\leq \sum_{j \geq J} \left[(j+1)^N |\mathcal{P}| (1 + \varepsilon) - \frac{1}{2} j^N |\mathcal{P}| \right] e^{-j} \end{aligned}$$

$$\begin{aligned}
 &\leq |\mathcal{P}| \sum_{j \geq J} \left[(j+1)^N - j^N \right] e^{-j} + |\mathcal{P}| \sum_{j \geq J} \left[\varepsilon(j+1)^N + \frac{1}{2}j^N \right] e^{-j} \\
 &\leq (e-1)|\mathcal{P}| \sum_{j \geq J} j^N e^{-j} + \left(\varepsilon e + \frac{1}{2} \right) |\mathcal{P}| \sum_{j \geq J} j^N e^{-j},
 \end{aligned}$$

the last estimate coming from $(j+1)^N < ej^N$.

Apply Lemma 6 with $L = 2$ and $J \geq \frac{2}{e^{1/N}-1}$, we have

$$\sum_{\nu \notin \mathcal{P}_J \cap \mathbb{Z}^N} e^{-b(\nu)} \leq (2e + 2\varepsilon e - 1) |\mathcal{P}| J^N e^{-J} \frac{e}{e-1}. \quad (4.21)$$

Now, we need to write (4.21) in term of $M = \#\Lambda_M = \#(\mathcal{P}_J \cap \mathbb{Z}^N)$. From (4.20), it is easy to see that

$$\frac{1}{2} J^N |\mathcal{P}| \leq M \leq J^N |\mathcal{P}| (1 + \varepsilon). \quad (4.22)$$

Combining (4.21)–(4.22) gives

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \leq C_u(\varepsilon) M \exp \left(- \left(\frac{M}{|\mathcal{P}|(1 + \varepsilon)} \right)^{1/N} \right),$$

where $C_u(\varepsilon) = (4e + 4\varepsilon e - 2) \frac{e}{e-1}$, as desired. \square

Remark 1 (Theoretical minimum cardinality M_ε). *The error estimate (4.19) holds with*

$$M > M_\varepsilon := \#(\mathcal{P}_{J_\varepsilon} \cap \mathbb{Z}^N), \text{ where } J_\varepsilon = \max \left\{ \frac{2}{e^{1/N}-1}, \Delta_\varepsilon \right\}. \quad (4.23)$$

It is shown in (4.20) that Δ_ε is decreasing with respect to ε . Thus, a stronger convergence rate, corresponding to smaller ε , would be realized at larger cardinality M . An evaluation of Δ_ε is not accessible to us in general, making explicit computation (or mathematical formula) of minimum cardinality M_ε not feasible. However, in the settings where \mathcal{P} is a rational convex polytope, Δ_ε can be acquired computationally. The interplay between ε and M_ε will be investigated through several examples within such settings in Section 5.

In any case, (4.19) requires $J \geq \frac{2}{e^{1/N}-1}$. This condition can be relaxed with a slightly weaker estimate. Indeed, applying (4.16) instead of (4.15) in the proof of Theorem 2, one gets

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \leq \frac{N+1}{2} C_u(\varepsilon) M \exp \left(- \left(\frac{M}{|\mathcal{P}|(1 + \varepsilon)} \right)^{1/N} \right), \quad (4.24)$$

given

$$M > M'_\varepsilon := \#(\mathcal{P}_{J'_\varepsilon} \cap \mathbb{Z}^N), \text{ where } J'_\varepsilon = \max \left\{ \frac{1}{e^{1/N}-1}, \Delta_\varepsilon \right\}. \quad (4.25)$$

Remark 2 (An extension of Theorem 2). *The convergence estimate (4.19) does not apply for $|\mathcal{P}| = 0$ or \mathcal{P} unbounded ($b(\boldsymbol{\nu}) \notin \Theta(|\boldsymbol{\nu}|)$). With minor modifications in the above analysis, our results can be extended to a wider class of $b(\boldsymbol{\nu})$ where Assumption 3.3 ($b(\boldsymbol{\nu}) \in \Theta(|\boldsymbol{\nu}|)$) is replaced by the condition that $b(\boldsymbol{\nu}) \in \Theta(|\boldsymbol{\nu}|^\beta)$ (i.e., there exist constants $0 < c < C$ such that $c|\boldsymbol{\nu}|^\beta < b(\boldsymbol{\nu}) < C|\boldsymbol{\nu}|^\beta$ as $\boldsymbol{\nu} \rightarrow \infty$) with some fixed $\beta > 0$. In such cases, it gives*

$$\sum_{\boldsymbol{\nu} \notin \Lambda_M} e^{-b(\boldsymbol{\nu})} \leq C_u(\varepsilon) M \exp\left(-\left(\frac{M}{|\mathcal{P}|(1+\varepsilon)}\right)^{\beta/N}\right)$$

as $M \rightarrow \infty$. Here, $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} \left(\frac{1}{\tau^{1/\beta}} \mathcal{P}_\tau\right)$ or $\mathcal{P} = \bigcup_{\tau \in \mathbb{R}^+} \left(\frac{1}{\tau^{1/\beta}} \mathcal{P}_\tau\right)$ (depending on whether $\frac{1}{\tau^{1/\beta}} \mathcal{P}_\tau$ is descending or ascending).

5. The optimality of our proposed estimation and pre-asymptotic error analysis: a simplified case. In this section, we consider the particular case in which

- i) \mathcal{P} is a rational convex polytope,
- ii) $\mathcal{P}_\tau = \tau \mathcal{P}$ for all $\tau \in (0, \infty)$.

This setting, arising from the multi-indexed sequence $(e^{-b(\boldsymbol{\nu})})_{\boldsymbol{\nu} \in \mathcal{S}}$ with

$$b(\boldsymbol{\nu}) = \sup_{\lambda \in \mathbf{A}} \left(\sum_{i=1}^N \lambda_i \nu_i \right), \text{ where } \mathbf{A} \text{ is a finite subset of } (\mathbb{Q}^+)^N, \quad (5.1)$$

is appropriate for Taylor coefficient estimate of the form (3.8) and to some extent, (3.13) (Details will be discussed in Section 6). The advantage here is that the number of integer points $\#(\mathcal{P}_j \cap \mathbb{Z}^N)$ can be represented by a computable *Ehrhart quasi-polynomial* of degree N in j (see [8], Chapter 3 and [33], Chapter 4). In other words, there exist a period q and polynomials E_0, \dots, E_{q-1} of degree N with leading coefficient $|\mathcal{P}|$ such that $\#(\mathcal{P}_j \cap \mathbb{Z}^N) = E_i(j)$ if $j \equiv i \pmod{q}$. We exploit this property for two tasks: first, to establish a lower bound of $\sum_{\boldsymbol{\nu} \notin \Lambda_M} e^{-b(\boldsymbol{\nu})}$ and verify the sharpness of estimate (4.19); second, to calculate the minimum cardinalities for (4.19) and (4.24) to hold (via the computations of the Ehrhart quasi-polynomials) and study the relation between them and the convergence rate. To circumvent the constraints on M_ε and M'_ε , an estimate of the truncation errors in the pre-asymptotic regime will be derived.

5.1. Lower bound of the truncation errors. We begin this section with an additional assumption on b , which is fulfilled by $b(\boldsymbol{\nu})$ defined in (5.1).

Assumption 4 (Monotonically increasing). $b : [0, \infty)^N \rightarrow \mathbb{R}$ satisfies: $\forall \boldsymbol{\nu}, \boldsymbol{\mu} \in [0, \infty)^N$, if $\boldsymbol{\nu} \leq \boldsymbol{\mu}$, then $b(\boldsymbol{\nu}) \leq b(\boldsymbol{\mu})$.

Given this monotone property, the number of integer points inside a superlevel set \mathcal{P}_τ is always larger than its Lebesgue measure. This observation is verified in the following lemma.

Lemma 7. Assume that $b : [0, \infty)^N \rightarrow \mathbb{R}$ is continuous and satisfies Assumption 4. For $\tau \in (0, \infty)$, denote $\mathcal{P}_\tau = \{\boldsymbol{\nu} \in [0, \infty)^N : b(\boldsymbol{\nu}) \leq \tau\}$. We have

$$\#(\mathcal{P}_\tau \cap \mathbb{Z}^N) \geq |\mathcal{P}_\tau|, \quad \forall \tau > 0.$$

Proof. We consider a partition of $[0, \infty)^N$ by the family of cells $(I_\nu)_{\nu \in \mathcal{S}}$ defined as

$$I_\nu = \bigotimes_{1 \leq i \leq N} [\nu_i, \nu_i + 1).$$

Denoting $\mathcal{S}^* = \{\nu \in \mathcal{S} : \mathcal{P}_\tau \cap I_\nu \neq \emptyset\}$. If $\nu \in \mathcal{S}^*$, by definition, there exists $\mu \in I_\nu$ such that $b(\mu) \leq \tau$. Since $\nu \leq \mu$ and b satisfies Assumption 4, it gives $b(\nu) \leq b(\mu) \leq \tau$. We have $\nu \in \mathcal{P}_\tau \cap \mathbb{Z}^N$, which implies $\mathcal{S}^* \subset \mathcal{P}_\tau \cap \mathbb{Z}^N$ and $\#\mathcal{S}^* \leq \#\mathcal{P}_\tau \cap \mathbb{Z}^N$.

On the other hand, there holds

$$|\mathcal{P}_\tau| = \sum_{\nu \in \mathcal{S}} |\mathcal{P}_\tau \cap I_\nu| \leq \sum_{\nu \in \mathcal{S}^*} |I_\nu| = \#\mathcal{S}^*.$$

We obtain $|\mathcal{P}_\tau| \leq \#\mathcal{S}^* \leq \#\mathcal{P}_\tau \cap \mathbb{Z}^N$, as desired. \square

Now, we proceed to establish a lower bound for the truncation errors of series $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ with $b(\nu)$ having the form (5.1).

Theorem 3. *Consider the multi-indexed series $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ with $b(\nu)$ given by (5.1). There exists a constant $M^* > 0$ such that*

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \geq C_\ell M^{1-\frac{1}{N}} \exp\left(-\left(\frac{M}{|\mathcal{P}|}\right)^{1/N}\right) \quad (5.2)$$

for all $M > M^*$. Here, \mathcal{P} is defined as in (4.12), $C_\ell = \frac{1}{2} \left(\frac{2}{3}\right)^{1-\frac{1}{N}} \frac{N|\mathcal{P}|^{\frac{1}{N}} q}{e^q - 1}$ where q is the period of Ehrhart quasi-polynomial of \mathcal{P} .

Proof. It is easy to see that b satisfies Assumption 3. Particularly, $b(\tau\nu) = \tau b(\nu)$, H_ν is constant and $\frac{1}{\tau}\mathcal{P}_\tau = \mathcal{P} = \{\nu \in [0, \infty)^N : b(\nu) \leq 1\}$ for all $\tau \in (0, \infty)$, $\nu \in [0, \infty)^N$. By definition of b , \mathcal{P} is a rational convex polytope. We can find $q \in \mathbb{N}$ and an N -order polynomial E with leading coefficient $|\mathcal{P}|$ such that

$$\#(\mathcal{P}_{jq} \cap \mathbb{Z}^N) = E(jq), \quad \forall j \in \mathbb{N}. \quad (5.3)$$

For $\Lambda_M = \mathcal{P}_{Jq} \cap \mathbb{Z}^N$, it gives

$$\begin{aligned} \sum_{\nu \notin \Lambda_M} e^{-b(\nu)} &\geq \sum_{j \geq J} (\#(\mathcal{P}_{(j+1)q} \cap \mathbb{Z}^N) - \#(\mathcal{P}_{jq} \cap \mathbb{Z}^N)) e^{-(j+1)q} \\ &= \sum_{j \geq J} (E(jq+q) - E(jq)) e^{-(j+1)q} \end{aligned} \quad (5.4)$$

Denoting $E(t) = |\mathcal{P}|t^N + \sum_{i=0}^{N-1} c_i t^i$, $\forall t \in \mathbb{R}$, we have

$$E(jq+q) - E(jq) \geq q|\mathcal{P}|N(jq)^{N-1} - q \sum_{i=0}^{N-1} |c_i| i(jq+q)^{i-1}. \quad (5.5)$$

There exists $\Upsilon_1 > 0$ satisfying

$$\sum_{i=0}^{N-1} |c_i| i(jq + q)^{i-1} \leq \frac{1}{2} |\mathcal{P}| N(jq)^{N-1}, \quad \forall j \in \mathbb{N}, j > \Upsilon_1. \quad (5.6)$$

Combining (5.4)–(5.6) yields for $J > \Upsilon_1$,

$$\begin{aligned} \sum_{\nu \notin \Lambda_M} e^{-b(\nu)} &\geq \frac{1}{2} q |\mathcal{P}| N \sum_{j \geq J} (jq)^{N-1} e^{-(j+1)q} \\ &\geq \frac{1}{2} N q^N J^{N-1} |\mathcal{P}| \sum_{j \geq J} e^{-(j+1)q} = \frac{1}{2} N q (qJ)^{N-1} |\mathcal{P}| \frac{e^{-qJ}}{e^q - 1}. \end{aligned} \quad (5.7)$$

We need to write this estimate in term of the cardinality M . First, notice that b satisfies Assumption 4, there holds

$$|\mathcal{P}| (Jq)^N = |\mathcal{P}_{Jq}| \leq \#(\mathcal{P}_{Jq} \cap \mathbb{Z}^N). \quad (5.8)$$

Applying Theorem 2, it gives $|\mathcal{P}| = \lim_{j \rightarrow \infty} \frac{1}{(jq)^N} \cdot \#(\mathcal{P}_{jq} \cap \mathbb{Z}^N)$. We can choose $\Upsilon_2 > 0$ such that for all $j \in \mathbb{N}$, $j > \Upsilon_2$,

$$-\frac{1}{2} |\mathcal{P}| \leq |\mathcal{P}| - \frac{1}{(jq)^N} \cdot \#(\mathcal{P}_{jq} \cap \mathbb{Z}^N). \quad (5.9)$$

Since $M = \#(\mathcal{P}_{Jq} \cap \mathbb{Z}^N)$, from (5.8) and (5.9), one has

$$|\mathcal{P}| (Jq)^N \leq M \leq \frac{3}{2} |\mathcal{P}| (Jq)^N \quad \text{for } J > \Upsilon_2. \quad (5.10)$$

Combining (5.7) and (5.10) gives

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \geq C_\ell M^{1-\frac{1}{N}} \exp\left(-\left(\frac{M}{|\mathcal{P}|}\right)^{1/N}\right),$$

where $C_\ell = \frac{1}{2} \left(\frac{2}{3}\right)^{1-\frac{1}{N}} \frac{N|\mathcal{P}|^{\frac{1}{N}} q}{e^q - 1}$. The proof is now complete. \square

Theorem 2 and Proposition 3 reveal that for $b(\nu)$ given by (5.1), the asymptotic truncation error of $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ can be bounded from below and above as

$$C_\ell M^{1-\frac{1}{N}} \exp\left(-\left(\frac{M}{|\mathcal{P}|}\right)^{1/N}\right) \leq \sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \leq C_u(\varepsilon) M \exp\left(-\left(\frac{M}{|\mathcal{P}|(1+\varepsilon)}\right)^{1/N}\right),$$

where C_ℓ and $C_u(\varepsilon)$ are mild constants in comparison with the total bounds. The optimality of our estimation is verified in these cases.

5.2. Asymptotic minimum cardinalities and their relation with the convergence rate. In this section, we will apply Ehrhart (quasi-)polynomial to investigate the mini-

mum cardinality for our asymptotic convergence rate to hold. Recall that for any $\varepsilon > 0$, the upper estimates (4.19) and (4.24) occur with $J > J_\varepsilon = \max\left\{\frac{2}{e^{1/N}-1}, \Delta_\varepsilon\right\}$ and $J > J'_\varepsilon = \max\left\{\frac{1}{e^{1/N}-1}, \Delta_\varepsilon\right\}$, respectively. The first constraints in both conditions are straightforward and we focus on quantifying Δ_ε . From (4.20), Δ_ε is the positive real number such that

$$\frac{1}{2}j^N|\mathcal{P}| \leq \#(\mathcal{P}_j \cap \mathbb{Z}^N) \leq j^N|\mathcal{P}|(1 + \varepsilon), \quad \forall j \in \mathbb{N}, j > \Delta_\varepsilon. \quad (5.11)$$

In case \mathcal{P} is a rational convex polytope and $\mathcal{P}_\tau = \tau\mathcal{P} \quad \forall \tau \in (0, \infty)$, we can ignore the left inequality of (5.11), which by Lemma 7 is true for all $j \in \mathbb{N}$. There exists a (quasi-)polynomial

$$E^*(j) = |\mathcal{P}|j^N + \sum_{i=0}^{N-1} c_i^*(j)j^i, \quad (5.12)$$

with $c_i^* : \mathbb{N} \rightarrow \mathbb{Q}$ being a periodic function with integer period q such that

$$\#(\mathcal{P}_j \cap \mathbb{Z}^N) = E^*(j), \quad \forall j \in \mathbb{N}, \quad (5.13)$$

see [8], Chapter 3 and [33], Chapter 4. Replacing (5.13) to (5.11), Δ_ε can be characterized as the largest among the solutions of

$$\varepsilon|\mathcal{P}|j^N - \sum_{i=0}^{N-1} c_i^*(j)j^i = 0.$$

The numerical computation of formula of Ehrhart polynomial $E^*(j)$ can be done efficiently [15], allowing us to quantify Δ_ε and the theoretical minimum cardinality $M_\varepsilon (= E^*(J_\varepsilon))$ accurately. We present a brief study on the relation between M_ε and ε for some polytopes, including:

- (P.1): $b(\boldsymbol{\nu}) = \sum_{i=1}^4 \nu_i$ ($N = 4$),
- (P.2): $b(\boldsymbol{\nu}) = \nu_1 + \nu_2 + 2\nu_3 + 4\nu_4$ ($N = 4$),
- (P.3): $b(\boldsymbol{\nu}) = \sum_{i=1}^8 \nu_i$ ($N = 8$),
- (P.4): $b(\boldsymbol{\nu}) = \sum_{i=1}^8 \frac{\nu_i}{2^{i-3}}$ ($N = 8$),
- (P.5): $b(\boldsymbol{\nu}) = \sup\left\{\frac{1}{2}\sum_{i=1}^8 \nu_i, \frac{5}{16}\sum_{i=1}^8 \nu_i + \frac{5}{16}\nu_j : 1 \leq j \leq 8\right\}$ ($N = 8$),
- (P.6): $b(\boldsymbol{\nu}) = \sup\left\{\frac{1}{5}\sum_{i=1}^8 \nu_i, \frac{1}{8}\sum_{i=1}^8 \nu_i + \frac{1}{8}\nu_j : 1 \leq j \leq 8\right\}$ ($N = 8$).

(P.1)-(P.4) correspond to 4- and 8-simplices with different levels of anisotropy. The lengths of edges connecting the origin and other vertices are equal for (P.1) and (P.3) and slightly vary for (P.2), while (P.4) is quite a skinny simplex. On the other hand, (P.5) is a truncated, enlarged version of (P.3) where the vertices are at $\frac{1}{5}$ of the way along the axis edges and $\frac{2}{5}$ along other edges, resulting in a polytope with 65 vertices. (P.6) in turn is obtained through an enlargement of (P.5).

Figure 2 shows the variation of M_ε and M'_ε as well as the *rate adjusting parameter* $1/(1 + \varepsilon)^{1/N}$ in the estimates (4.19) and (4.24) with respect to ε . The other parameter $C_u(\varepsilon)$ is negligible except for ε very large and not plotted here. The formulas of Ehrhart polynomials are calculated using the software package `LattE` [5]. First, we observe that choosing a smaller ε gives a stronger convergence, yet M_ε must also be increased. The good news is that while the best convergence $M \exp\left(-\left(\frac{M}{|\mathcal{P}|}\right)^{1/N}\right)$ is realized only as $\varepsilon \rightarrow 0$, ε need not to be small to obtain a strong rate, especially in high dimension. For instance, $\varepsilon = 1.0$ gives the rate $\sim M \exp\left(-0.83 \left(\frac{M}{|\mathcal{P}|}\right)^{1/N}\right)$ with $N = 4$ and $\sim M \exp\left(-0.92 \left(\frac{M}{|\mathcal{P}|}\right)^{1/N}\right)$ with $N = 8$.

Not surprisingly, M_ε and M'_ε is shown to be larger for higher dimension. For a fixed N , the anisotropy of the polytopes significantly impacts M_ε and M'_ε : these values are close for (P.3) and (P.5), which possess different shapes and scales but span equally in coordinate axes, and much larger for (P.4), the simplex with skinny shape. Generally, increasing ε alleviates the restrictions on M_ε and M'_ε , as this will reduce Δ_ε . The strategy is, however, ineffective once $\frac{2}{e^{1/N}-1}$ (or $\frac{1}{e^{1/N}-1}$) exceeds Δ_ε and dominates (4.23) and (4.25), at which point, these conditions can no further be relaxed. Thus, while M_ε and M'_ε are almost not affected by the scale of polytopes with ε close to 0, their lower bounds (imposed by $J \gtrsim \frac{1}{e^{1/N}-1} \simeq N$) are more restrictive for large polytopes; in such cases, mild constraints on M_ε and M'_ε may be unattainable. This fact is illustrated by a comparison of two similar polytopes (P.5) and (P.6) in Figure 2: M_ε and M'_ε eventually stop to decay in both cases, but the bound is higher for (P.6), the polytope with larger scale.

In short, our asymptotic convergence analysis applies to the range $J \geq N$. In the next part, we propose an alternative estimate of truncation errors, which is effective in the pre-asymptotic regime $J < N$. In Section 6.3, we will show in some examples that the actual condition on M_ε for (4.19) to hold can be much milder than the theoretical minimum cardinality posed by Theorem 2 and investigated here.

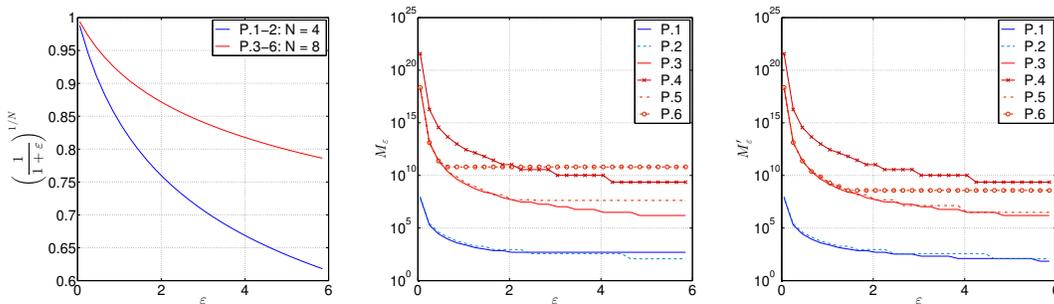


Figure 2: The variation of the rate adjusting parameter and theoretical minimum cardinalities M_ε and M'_ε for the upper estimate (4.19) with respect to ε .

5.3. A pre-asymptotic estimate of truncation errors. To acquire an estimation of $\sum_{\nu \notin \Lambda_M} e^{-b(\nu)}$ in pre-asymptotic regime, following the arguments in Theorem 2, non-asymptotic bounds of $\#(\mathcal{P}_j \cap \mathbb{Z}^N)$ and $\sum_{j \geq J} j^N e^{-j}$ need to be established. An upper bound of

$\#(\mathcal{P}_j \cap \mathbb{Z}^N)$ is derived in the following lemma.

Lemma 8. *Let $b : [0, \infty)^N \rightarrow \mathbb{R}$ be continuous and satisfy Assumption 4. Assuming that $b(\tau \boldsymbol{\nu}) = \tau b(\boldsymbol{\nu})$ for all $\tau \in (0, \infty)$, $\boldsymbol{\nu} \in [0, \infty)^N$. For $\tau \in (0, \infty)$, denote $\mathcal{P}_\tau = \{\boldsymbol{\nu} \in [0, \infty)^N : b(\boldsymbol{\nu}) \leq \tau\}$. There follows*

$$\#(\mathcal{P}_j \cap \mathbb{Z}^N) \leq j^N \cdot \#(\mathcal{P} \cap \mathbb{Z}^N), \quad \forall j \in \mathbb{N},$$

where $\mathcal{P} = \{\boldsymbol{\nu} \in [0, \infty)^N : b(\boldsymbol{\nu}) \leq 1\}$ ($= \frac{1}{\tau} \mathcal{P}_\tau$ for all τ).

Proof. Since $b(\tau \boldsymbol{\nu}) \equiv \tau b(\boldsymbol{\nu})$, we have $\mathcal{P}_\tau = \tau \mathcal{P}$, $\forall \tau > 0$, thus,

$$\#(\mathcal{P}_j \cap \mathbb{Z}^N) = \#(j \mathcal{P} \cap \mathbb{Z}^N) = \# \left(\mathcal{P} \cap \frac{1}{j} \mathbb{Z}^N \right), \quad \forall j \in \mathbb{N}.$$

Given $\boldsymbol{\mu} \in \left(\mathcal{P} \cap \frac{1}{j} \mathbb{Z}^N \right)$, $\boldsymbol{\mu}$ can be written uniquely in the form

$$\boldsymbol{\mu} = \boldsymbol{\nu} + \bigotimes_{i=1}^N \frac{r_i}{j},$$

where $\boldsymbol{\nu} \in \mathcal{S}$ and $r_i \in \mathbb{Z}$, $0 \leq r_i \leq j-1$, $\forall 1 \leq i \leq N$.

Since $\boldsymbol{\nu} \leq \boldsymbol{\mu}$ and b satisfies Assumption 4, it gives $b(\boldsymbol{\nu}) \leq b(\boldsymbol{\mu}) \leq 1$ and, consequently, $\boldsymbol{\nu} \in \mathcal{P} \cap \mathbb{Z}^N$. We have

$$\begin{aligned} \# \left(\mathcal{P} \cap \frac{1}{j} \mathbb{Z}^N \right) &\leq \# \left\{ \boldsymbol{\nu} + \bigotimes_{i=1}^N \frac{r_i}{j} : \boldsymbol{\nu} \in \mathcal{P} \cap \mathbb{Z}^N, r_i \in \mathbb{Z}, 0 \leq r_i \leq j-1, \forall 1 \leq i \leq N \right\} \\ &= j^N \cdot \#(\mathcal{P} \cap \mathbb{Z}^N), \end{aligned}$$

as desired. \square

Next, we give a non-asymptotic estimate of $\sum_{j \geq J} j^N e^{-j}$ based on tight approximation of $\sum_{j \leq J-1} j^N e^{-j}$ for $J \leq N+1$. Indeed, since $\tau \mapsto \tau^N e^{-\tau}$ is increasing in $[0, N]$, we have

$$\sum_{j=1}^{J-1} j^N e^{-j} \geq \int_0^{J-1} \tau^N e^{-\tau} d\tau. \quad (5.14)$$

Applying Theorem 4.1, [31], yields

$$\int_0^{J-1} \tau^N e^{-\tau} d\tau \geq \frac{(J-1)^{N+1}}{N+1} \exp \left(-\frac{(J-1)(N+1)}{N+2} \right). \quad (5.15)$$

Combining (5.14) and (5.15), it gives

$$\begin{aligned} \sum_{j \geq J} j^N e^{-j} &= \sum_{j=1}^{\infty} j^N e^{-j} - \sum_{j=1}^{J-1} j^N e^{-j} \\ &\leq \sum_{j=1}^{\infty} j^N e^{-j} - \frac{(J-1)^{N+1}}{N+1} \exp\left(-\frac{(J-1)(N+1)}{N+2}\right). \end{aligned} \quad (5.16)$$

A mathematical formula of the sum $\sum_{j=1}^{\infty} j^N e^{-j}$ is not accessible. However, it is independent of J and can be written in term of the well-studied *polylogarithm functions*

$$Li_s(z) = \sum_{j=1}^{\infty} \frac{z^j}{j^s}, \quad \text{for } z \in \mathbb{C}, |z| < 1, s \in \mathbb{R}. \quad (5.17)$$

see [26,27]. Combining (5.16) and (5.17), we have proved the following Lemma:

Lemma 9. *For any $N, J \in \mathbb{N}$, if $J \leq N+1$, it gives*

$$\sum_{j \geq J} j^N e^{-j} \leq Li_{-N}(1/e) - \frac{(J-1)^{N+1}}{N+1} \exp\left(-\frac{(J-1)(N+1)}{N+2}\right). \quad (5.18)$$

In Figure 3, we compare the performance of the asymptotic bound (4.16) and the pre-asymptotic bound (5.18) in estimating the truncation error of $\sum_{j=1}^{\infty} j^N e^{-j}$ for $N = 20$. The pre-asymptotic estimate shows an excellent agreement with true value for small J ; however, it cannot capture the error decay when J is big. The asymptotic bound, on the other hand, successfully predicts the convergence rate of $\sum_{j=1}^{\infty} j^N e^{-j}$, but is not effective with small J . We are now in the position to prove a pre-asymptotic estimation of $\sum_{\nu \notin \Lambda_M} e^{-b(\nu)}$.

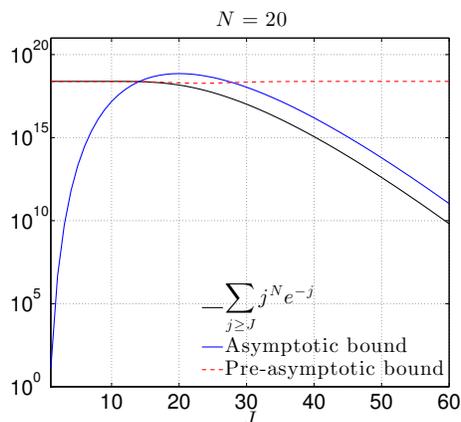


Figure 3: A comparison of the asymptotic bound (4.16) and the pre-asymptotic bound (5.18) in estimating $\sum_{j \geq J} j^N e^{-j}$ for $N = 20$.

Theorem 4. *Consider the multi-indexed series $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ with $b(\nu)$ being continuous and satisfying Assumption 4. Assuming that $b(\tau\nu) = \tau b(\nu)$ for all $\tau \in (0, \infty), \nu \in [0, \infty)^N$.*

For $\tau \in (0, \infty)$, denote $\mathcal{P}_\tau = \{\nu \in [0, \infty)^N : b(\nu) \leq \tau\}$ and Λ_M the set of indices corresponding to M largest $e^{-b(\nu)}$. Define \mathcal{P} as in (4.12).

For $M \in \mathbb{N}$, if $M \leq \#(\mathcal{P}_N \cap \mathbb{Z}^N)$, there holds

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \leq e\sigma \left[Li_{-N}(1/e) - \frac{1}{N+1} \left(\frac{M}{\sigma} \right)^{\frac{N+1}{N}} \exp \left(-\frac{M^{\frac{1}{N}}(N+1)}{\sigma^{\frac{1}{N}}(N+2)} \right) \right]. \quad (5.19)$$

Here, $\sigma = \#(\mathcal{P} \cap \mathbb{Z}^N)$ and Li denotes the polylogarithm function.

Proof. Applying Lemma 8, it gives

$$\#(\mathcal{P}_j \cap \mathbb{Z}^N) \leq j^N \cdot \#(\mathcal{P} \cap \mathbb{Z}^N), \quad \forall j \in \mathbb{N}. \quad (5.20)$$

To estimate $\sum_{\nu \notin \Lambda_M} e^{-b(\nu)}$, it is sufficient to consider this sum with $\Lambda_M = \mathcal{P}_J \cap \mathbb{Z}^N$, $J \in \mathbb{N}$, $J \leq N$. We have

$$\begin{aligned} \sum_{\nu \notin \mathcal{P}_J \cap \mathbb{Z}^N} e^{-b(\nu)} &\leq \sum_{j \geq J} (\#(\mathcal{P}_{j+1} \cap \mathbb{Z}^N) - \#(\mathcal{P}_j \cap \mathbb{Z}^N)) e^{-j} \\ &\leq \#(\mathcal{P} \cap \mathbb{Z}^N) \sum_{j \geq J} (j+1)^N e^{-j} \\ &\leq e \cdot \#(\mathcal{P} \cap \mathbb{Z}^N) \left[Li_{-N}(1/e) - \frac{J^{N+1}}{N+1} \exp \left(-\frac{J(N+1)}{N+2} \right) \right], \end{aligned}$$

by applying Lemma 9.

From (5.20), it gives $J \geq \left(\frac{M}{\#(\mathcal{P} \cap \mathbb{Z}^N)} \right)^{1/N}$. It is easy to see that the mapping $j \mapsto \frac{j^{N+1}}{N+1} \exp \left(-\frac{j(N+1)}{N+2} \right)$ is increasing in $[0, N]$. There follows

$$\sum_{\nu \notin \Lambda_M} e^{-b(\nu)} \leq e\sigma \left[Li_{-N}(1/e) - \frac{1}{N+1} \left(\frac{M}{\sigma} \right)^{\frac{N+1}{N}} \exp \left(-\frac{M^{\frac{1}{N}}(N+1)}{\sigma^{\frac{1}{N}}(N+2)} \right) \right],$$

where $\sigma = \#(\mathcal{P} \cap \mathbb{Z}^N)$, implying the assertion (5.19). \square

Remark 3. *The pre-asymptotic analysis presented above does not employ Ehrhart polynomials, hence applies to a wider class of b than those given by (5.1). In this subsection, b only needs to be continuous, satisfy Assumption 4 and $b(\tau\nu) = \tau b(\nu)$ for all $\tau \in (0, \infty)$, $\nu \in [0, \infty)^N$.*

6. Asymptotic convergence rates of quasi-optimal approximations. As we have seen so far, the error of a quasi-optimal polynomial approximation can be estimated by a series of coefficient upper bounds based on which the polynomial spaces are constructed. For the upper bounds developed in recent publications, we will verify that such series fall into the class of multi-indexed series analyzed in Section 4, allowing an application of our framework to those settings. In fact, as we shall see, in all considered cases, the coefficient estimates, written as $e^{-b(\nu)}$, satisfies Assumption 3 and $\sum_{\nu \in \mathcal{S}} e^{-b(\nu)}$ can be treated by Theorem 2.

Given a vector $\boldsymbol{\rho} = (\rho_i)_{i=1 \leq i \leq N}$ with $\rho_i > 1 \forall i$, we define $\boldsymbol{\lambda} = (\lambda_i)_{1 \leq i \leq N}$ such that

$$\lambda_i = \log \rho_i > 0 \quad \forall 1 \leq i \leq N.$$

In Section 6.1 and 6.2, we study the error analysis of quasi-optimal methods based on Taylor and Legendre expansions respectively. A computational comparison of our proposed estimate with existing results is showing in Section 6.3.

6.1. Error analysis of quasi-optimal Taylor approximations. We start with the quasi-optimal methods corresponding to a basic coefficient bound of the form $\boldsymbol{\rho}^{-\boldsymbol{\nu}}$ (see Proposition 1). These are reasonable schemes for Taylor approximations of elliptic problems with the random fields composed of non-overlapping basis functions. The convergence result is stated in the following proposition.

Proposition 3. Consider the Taylor series $\sum_{\boldsymbol{\nu} \in \mathcal{S}} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}}$ of u . Assume that

$$\|t_{\boldsymbol{\nu}}\|_{V(D)} \leq \frac{\|f\|_{V^*(D)}}{\delta} \boldsymbol{\rho}^{-\boldsymbol{\nu}} \quad (6.1)$$

holds for all $\boldsymbol{\nu} \in \mathcal{S}$, as in Proposition 1. Denote by Λ_M the set of indices corresponding to M largest bounds in (6.1). For any $\varepsilon > 0$, there exists $M_\varepsilon > 0$ depending on ε such that

$$\sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_M} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}} \right\|_{V(D)} \leq \frac{\|f\|_{V^*(D)}}{\delta} C_u(\varepsilon) M \exp \left(- \left(\frac{MN! \prod_{i=1}^N \lambda_i}{(1+\varepsilon)} \right)^{\frac{1}{N}} \right) \quad (6.2)$$

for all $M > M_\varepsilon$.

Proof. We have by triangle inequality

$$\sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_M} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}} \right\|_{V(D)} \leq \sum_{\boldsymbol{\nu} \notin \Lambda_M} \|t_{\boldsymbol{\nu}}\|_{V(D)} \leq \frac{\|f\|_{V^*(D)}}{\delta} \sum_{\boldsymbol{\nu} \notin \Lambda_M} \boldsymbol{\rho}^{-\boldsymbol{\nu}}. \quad (6.3)$$

For $\boldsymbol{\nu} \in [0, \infty)^N$, define $b(\boldsymbol{\nu}) = \sum_{i=1}^N \lambda_i \nu_i$, so that $\boldsymbol{\rho}^{-\boldsymbol{\nu}} = e^{-b(\boldsymbol{\nu})} \forall \boldsymbol{\nu} \in \mathcal{S}$. We notice that the quasi-optimal index sets in this case are the Total Degree spaces:

$$\mathcal{P}_j \cap \mathbb{Z}^N = \left\{ \boldsymbol{\nu} \in \mathcal{S} : \boldsymbol{\rho}^{-\boldsymbol{\nu}} \geq e^{-j} \right\} = \left\{ \boldsymbol{\nu} \in \mathcal{S} : \sum_{i=1}^N \lambda_i \nu_i \leq j \right\}, \quad \forall j \in \mathbb{N}.$$

Since $\lambda_i > 0 \forall i$, it is easy to check that the map b satisfies Assumption 3 with $H_{\boldsymbol{\nu}}$ being constant $\forall \boldsymbol{\nu}$. Observing that $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau) = \left\{ \boldsymbol{\nu} \in [0, \infty)^N : \sum_{i=1}^N \lambda_i \nu_i \leq 1 \right\}$, we can specify $|\mathcal{P}| = \frac{1}{N!(\lambda_1 \dots \lambda_N)}$.

We are now ready to apply Theorem 2 to obtain

$$\begin{aligned} \sum_{\boldsymbol{\nu} \notin \Lambda_M} \boldsymbol{\rho}^{-\boldsymbol{\nu}} &\leq C_u(\varepsilon)M \exp\left(-\left(\frac{M}{|\mathcal{P}|(1+\varepsilon)}\right)^{1/N}\right) \\ &\leq C_u(\varepsilon)M \exp\left(-\left(\frac{MN! \prod_{i=1}^N \lambda_i}{(1+\varepsilon)}\right)^{1/N}\right), \end{aligned}$$

which proves (6.2). \square

We proceed to analyze the quasi-optimal Taylor approximations based on best analytical bound provided by Proposition 1. Although this method is not easily implementable, an asymptotic error estimate can be obtained as a simple corollary of Theorem 2. It is reasonable to assume that the set \mathbf{Ad} of all admissible $(\delta, \boldsymbol{\rho})$ is bounded: as seen through several examples in Figure 1, the domains of uniform ellipticity do not expand infinitely in complex plane.

Proposition 4. *Consider the Taylor series $\sum_{\boldsymbol{\nu} \in \mathcal{S}} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}}$ of u . Assume*

$$\|t_{\boldsymbol{\nu}}\|_{V(D)} \leq \inf_{(\delta, \boldsymbol{\rho}) \in \mathbf{Ad}} \frac{\|f\|_{V^*(D)}}{\delta} \boldsymbol{\rho}^{-\boldsymbol{\nu}} \quad (6.4)$$

holds for all $\boldsymbol{\nu} \in \mathcal{S}$, as in (3.13), with \mathbf{Ad} being bounded. Denote by Λ_M the set of indices corresponding to M largest bounds in (6.4). For any $\varepsilon > 0$, there exists $M_\varepsilon > 0$ depending on ε such that

$$\sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_M} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}} \right\|_{V(D)} \leq \|f\|_{V^*(D)} C_u(\varepsilon) M \exp\left(-\left(\frac{M}{|\mathcal{P}|(1+\varepsilon)}\right)^{1/N}\right) \quad (6.5)$$

for all $M > M_\varepsilon$. Here, $\mathcal{P} = \left\{ \boldsymbol{\nu} \in [0, \infty)^N : \sum_{i=1}^N (\log \rho_i) \nu_i \leq 1 \quad \forall (\delta, \boldsymbol{\rho}) \in \mathbf{Ad} \right\}$.

Proof. First, we have

$$\sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_M} t_{\boldsymbol{\nu}} \mathbf{y}^{\boldsymbol{\nu}} \right\|_{V(D)} \leq \sum_{\boldsymbol{\nu} \notin \Lambda_M} \|t_{\boldsymbol{\nu}}\|_{V(D)} \leq \|f\|_{V^*(D)} \sum_{\boldsymbol{\nu} \notin \Lambda_M} \inf_{(\delta, \boldsymbol{\rho}) \in \mathbf{Ad}} \frac{\boldsymbol{\rho}^{-\boldsymbol{\nu}}}{\delta}. \quad (6.6)$$

Recall that $\lambda_i = \log \rho_i \quad \forall i$. With abuse of notation, we say $(\delta, \boldsymbol{\lambda}) \in \mathbf{Ad}$ iff $(\delta, \boldsymbol{\rho}) \in \mathbf{Ad}$. For $\boldsymbol{\nu} \in [0, \infty)^N$, define $b(\boldsymbol{\nu}) = \sup_{(\delta, \boldsymbol{\lambda}) \in \mathbf{Ad}} \left(\log \delta + \sum_{i=1}^N \lambda_i \nu_i \right)$, so that $\inf_{(\delta, \boldsymbol{\rho}) \in \mathbf{Ad}} \frac{\boldsymbol{\rho}^{-\boldsymbol{\nu}}}{\delta} = e^{-b(\boldsymbol{\nu})} \quad \forall \boldsymbol{\nu} \in \mathcal{S}$. The quasi-optimal index sets in this case are:

$$\mathcal{P}_j \cap \mathbb{Z}^N = \left\{ \boldsymbol{\nu} \in \mathcal{S} : \sup_{(\delta, \boldsymbol{\lambda}) \in \mathbf{Ad}} \left(\log \delta + \sum_{i=1}^N \lambda_i \nu_i \right) \leq j \right\}, \quad \forall j \in \mathbb{N}.$$

We will show that b fulfills Assumption 3. It is easy to check that b is convex. As a

consequence, for any $\tau > 0$, \mathcal{P}_τ and $\bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau)$ are convex (and Jordan measurable). Since \mathbf{Ad} is bounded, there exist $0 < c < C$ such that $c|\boldsymbol{\nu}| < b(\boldsymbol{\nu}) < C|\boldsymbol{\nu}|$ as $\boldsymbol{\nu} \rightarrow \infty$. Now, let $\tau \geq \tau' > 0$, it gives

$$\frac{1}{\tau'} \left(\log \delta + \sum_{i=1}^N \lambda_i \tau' \nu_i \right) \leq \frac{1}{\tau} \left(\log \delta + \sum_{i=1}^N \lambda_i \tau \nu_i \right), \quad \forall (\delta, \boldsymbol{\lambda}) \in \mathbf{Ad}, \boldsymbol{\nu} \in [0, \infty)^N,$$

since $\delta < 1$. Hence, $H_{\boldsymbol{\nu}}(\tau') \leq H_{\boldsymbol{\nu}}(\tau)$, $\forall \boldsymbol{\nu} \in [0, \infty)^N$.

We can apply Theorem 2 to get the asymptotic estimate

$$\sum_{\boldsymbol{\nu} \notin \Lambda_M} \inf_{(\delta, \boldsymbol{\rho}) \in \mathbf{Ad}} \frac{\boldsymbol{\rho}^{-\boldsymbol{\nu}}}{\delta} \leq C_u(\varepsilon) M \exp \left(- \left(\frac{M}{|\mathcal{P}|(1+\varepsilon)} \right)^{1/N} \right), \quad (6.7)$$

where $\mathcal{P} := \bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_\tau) = \left\{ \boldsymbol{\nu} \in [0, \infty)^N : \sum_{i=1}^N \lambda_i \nu_i \leq 1 \quad \forall (\delta, \boldsymbol{\lambda}) \in \mathbf{Ad} \right\}$. Combining (6.6) and (6.7) gives (6.5), concluding the proof. \square

6.2. Error analysis of quasi-optimal Legendre approximations. For the first example, we consider quasi-optimal methods for Legendre approximations of elliptic PDEs with the random field consisting of basis functions with disjoint supports. In [6], these problems were computationally treated with bounds of type (6.1) and Total Degree index sets with some success. However, those bounds are not analytically optimal, as the true exponential decay of coefficients is penalized by a large multiplier. In the following, we establish a convergence analysis for the sharper upper bound $\boldsymbol{\rho}^{-\boldsymbol{\nu}} \prod_{i=1}^N \sqrt{2\nu_i + 1}$ of Legendre coefficients (see Section 3.3). Whether the quasi-optimal method corresponding to this estimate outperforms Total Degree approximations in computation is an interesting subject to study next.

Proposition 5. *Consider the Legendre series $\sum_{\boldsymbol{\nu} \in \mathcal{S}} v_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}}$ of u . Assume that*

$$\|v_{\boldsymbol{\nu}}\|_{V(D)} \leq C_{\boldsymbol{\rho}, \delta} \boldsymbol{\rho}^{-\boldsymbol{\nu}} \prod_{i=1}^N \sqrt{2\nu_i + 1} \quad (6.8)$$

holds for all $\boldsymbol{\nu} \in \mathcal{S}$, as in Proposition 2. Denote by Λ_M the set of indices corresponding to M largest bounds in (6.8). For any $\varepsilon > 0$, there exists a constant $M_\varepsilon > 0$ depending on ε such that

$$\left\| u - \sum_{\boldsymbol{\nu} \in \Lambda_M} v_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}} \right\|_{V(D) \otimes L^2_\rho(\Gamma)}^2 \leq C_{\boldsymbol{\rho}, \delta}^2 C_u(\varepsilon) M \exp \left(-2 \left(\frac{MN! \prod_{i=1}^N \lambda_i}{(1+\varepsilon)} \right)^{1/N} \right) \quad (6.9)$$

for all $M > M_\varepsilon$.

Proof. First, we have

$$\left\| u - \sum_{\boldsymbol{\nu} \in \Lambda_M} v_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}} \right\|_{V(D) \otimes L^2_\rho(\Gamma)}^2 = \sum_{\boldsymbol{\nu} \notin \Lambda_M} \|v_{\boldsymbol{\nu}}\|_{V(D)}^2 \leq C_{\boldsymbol{\rho}, \delta}^2 \sum_{\boldsymbol{\nu} \notin \Lambda_M} \boldsymbol{\rho}^{-2\boldsymbol{\nu}} \prod_{i=1}^N (2\nu_i + 1).$$

For $\boldsymbol{\nu} \in [0, \infty)^N$, define $b(\boldsymbol{\nu}) = \sum_{i=1}^N (2\lambda_i \nu_i - \log(2\nu_i + 1))$, so that $\boldsymbol{\rho}^{-2\boldsymbol{\nu}} \prod_{i=1}^N (2\nu_i + 1) = e^{-b(\boldsymbol{\nu})} \forall \boldsymbol{\nu} \in \mathcal{S}$. We notice that the quasi-optimal index sets in this case given by:

$$\mathcal{P}_j \cap \mathbb{Z}^N = \left\{ \boldsymbol{\nu} \in \mathcal{S} : \sum_{i=1}^N (2\lambda_i \nu_i - \log(2\nu_i + 1)) \leq j \right\}, \forall j \in \mathbb{N}.$$

We proceed to prove b satisfies Assumption 3. It is easy to check that $b(\boldsymbol{\nu})$ is continuous. As $\boldsymbol{\nu} \rightarrow \infty$, $\lambda_{\min} |\boldsymbol{\nu}| < b(\boldsymbol{\nu}) < 2\lambda_{\max} |\boldsymbol{\nu}|$, where $\lambda_{\min} = \min_{1 \leq i \leq N} \lambda_i$ and $\lambda_{\max} = \max_{1 \leq i \leq N} \lambda_i$. Also, observing $\log(at + 1) \geq t \log(a + 1)$ for every $a \geq 0$, $0 \leq t \leq 1$, we have

$$H_{\boldsymbol{\nu}}(\tau') = \sum_{i=1}^N \left(2\lambda_i \nu_i - \frac{1}{\tau'} \log(2\tau' \nu_i + 1) \right) \leq \sum_{i=1}^N \left(2\lambda_i \nu_i - \frac{1}{\tau} \log(2\tau \nu_i + 1) \right) = H_{\boldsymbol{\nu}}(\tau)$$

for all $\tau, \tau' \in (0, \infty)$, $\tau \geq \tau'$.

We prove $\mathcal{P} = \bigcap_{\tau \in \mathbb{R}^+} (\frac{1}{\tau} \mathcal{P}_{\tau})$ is Jordan measurable and $|\mathcal{P}| = \frac{1}{2^N N! \prod_{i=1}^N \lambda_i}$ by showing that

$$\mathcal{P} = \left\{ \boldsymbol{\nu} \in [0, \infty)^N : \sum_{i=1}^N 2\lambda_i \nu_i \leq 1 \right\}.$$

Indeed, let $\boldsymbol{\nu}$ be contained in \mathcal{P} , then $\tau \boldsymbol{\nu} \in \mathcal{P}_{\tau}$, $\forall \tau > 0$, i.e.,

$$\sum_{i=1}^N (2\tau \lambda_i \nu_i - \log(2\tau \nu_i + 1)) \leq \tau, \quad \forall \tau > 0.$$

Dividing both sides by τ gives

$$\sum_{i=1}^N \left(2\lambda_i \nu_i - \frac{1}{\tau} \log(2\tau \nu_i + 1) \right) \leq 1, \quad \forall \tau > 0. \quad (6.10)$$

Taking the limit of (6.10) as $\tau \rightarrow \infty$, we have $\sum_{i=1}^N 2\lambda_i \nu_i \leq 1$.

Conversely, assuming $\boldsymbol{\nu} \in [0, \infty)^N$ such that $\sum_{i=1}^N 2\lambda_i \nu_i \leq 1$, there holds

$$\sum_{i=1}^N (2\tau \lambda_i \nu_i - \log(2\tau \nu_i + 1)) \leq \sum_{i=1}^N 2\tau \lambda_i \nu_i \leq \tau, \quad \forall \tau > 0,$$

which proves $\boldsymbol{\nu} \in \frac{1}{\tau} \mathcal{P}_{\tau}$, $\forall \tau > 0$.

Finally, applying Theorem 2, we obtain

$$\sum_{\nu \notin \Lambda_M} \rho^{-2\nu} \prod_{i=1}^N (2\nu_i + 1) \leq C_u(\varepsilon) M \exp \left(-2 \left(\frac{MN! \prod_{i=1}^N \lambda_i}{1 + \varepsilon} \right)^{1/N} \right)$$

for all $M > M_\varepsilon$. This concludes our proof. \square

We remark that while the bound (6.8) is weaker than (6.1), its corresponding index sets are descending towards Total Degree sets. As a result, we are able to obtain the same convergence rate as Taylor approximations.

Now, we apply our framework to prove a convergence estimate for quasi-optimal Legendre approximations based on the coefficient exponential decay $\|v_\nu\|_{V(D)} \leq \|f\|_{V^*(D)} \frac{|\nu|!}{\nu!} \alpha^\nu$. Unlike other upper bounds discussed so far, this decay is established by real analysis argument [13]. In the case of affine linear random fields, i.e. $a(x, \mathbf{y}) = a_0(x) + \sum_{i=1}^N y_i \psi_i(x)$, $\alpha = (\alpha_i)_{1 \leq i \leq N}$ is specified by $\alpha_i = \frac{\|\psi_i\|_{L^\infty(D)}}{a_{\min} \sqrt{3}}$. A development and implementation of quasi-optimal method can be found in [7]; however, no error estimate has been provided. In the following result, similar to the aforementioned works, we assume $\sum_{i=1}^N \alpha_i < 1$, which is necessary for the summability of sequence $\left(\frac{|\nu|!}{\nu!} \alpha^\nu \right)_{\nu \in \mathcal{S}}$.

Proposition 6. *Consider the Legendre series $\sum_{\nu \in \mathcal{S}} v_\nu L_\nu$ of u . Assume there exists a vector $\alpha = (\alpha_i)_{1 \leq i \leq N}$ with $\alpha_i > 0 \forall i$ and $\sum_{i=1}^N \alpha_i < 1$ such that*

$$\|v_\nu\|_{V(D)} \leq \|f\|_{V^*(D)} \frac{|\nu|!}{\nu!} \alpha^\nu \quad (6.11)$$

for all $\nu \in \mathcal{S}$. Denote by Λ_M the set of indices corresponding to M largest bounds in (6.11). For any $\varepsilon > 0$, there exists a constant $M_\varepsilon > 0$ depending on ε such that

$$\left\| u - \sum_{\nu \in \Lambda_M} v_\nu L_\nu \right\|_{V(D) \otimes L^2_\rho(\Gamma)}^2 \leq \|f\|_{V^*(D)}^2 C_u(\varepsilon) M \exp \left(- \left(\frac{M}{|\mathcal{P}|(1 + \varepsilon)} \right)^{1/N} \right) \quad (6.12)$$

for all $M > M_\varepsilon$. Here, $\mathcal{P} = \left\{ \nu \in (0, \infty)^N : \sum_{i=1}^N \lambda_i \nu_i - \log \frac{|\nu|^{|\nu|}}{\prod_{i=1}^N \nu_i^{\nu_i}} < \frac{1}{2} \right\}$.

Proof. From (6.11), we have

$$\left\| u - \sum_{\nu \in \Lambda_M} v_\nu L_\nu \right\|_{V(D) \otimes L^2_\rho(\Gamma)}^2 = \sum_{\nu \notin \Lambda_M} \|v_\nu\|_{V(D)}^2 \leq \|f\|_{V^*(D)}^2 \sum_{\nu \notin \Lambda_M} \alpha^{2\nu} \left(\frac{|\nu|!}{\nu!} \right)^2.$$

Let $\lambda_i = -\log \alpha_i > 0 \forall 1 \leq i \leq N$ and Γ denote the gamma function. Also, let ψ_0, ψ_1 and ψ_2 be the di-, tri- and tetra-gamma functions respectively: $\psi_0 = (\log \Gamma)'$, $\psi_1 = \psi_0' = (\log \Gamma)''$, $\psi_2 = \psi_1' = (\log \Gamma)'''$. For $\nu \in [0, \infty)^N$, define $b(\nu) = 2 \sum_{i=1}^N \lambda_i \nu_i - 2 \log \frac{\Gamma(|\nu|+1)}{\prod_{i=1}^N \Gamma(\nu_i+1)}$, so that $\alpha^{2\nu} \left(\frac{|\nu|!}{\nu!} \right)^2 = e^{-b(\nu)} \forall \nu \in \mathcal{S}$. The quasi-optimal index sets in this case are given

by:

$$\mathcal{P}_j \cap \mathbb{Z}^N = \left\{ \boldsymbol{\nu} \in \mathcal{S} : \sum_{i=1}^N \lambda_i \nu_i - \log \frac{\Gamma(|\boldsymbol{\nu}| + 1)}{\prod_{i=1}^N \Gamma(\nu_i + 1)} \leq \frac{j}{2} \right\}.$$

We proceed to prove b satisfies Assumption 3. First, since $\sum_{i=1}^N \alpha_i < 1$, one can find $p \in (0, 1)$ such that $\sum_{i=1}^N \alpha_i^p < 1$ and, by Theorem 7.2 in [13], have $\left(\frac{|\boldsymbol{\nu}|!}{\boldsymbol{\nu}!} \boldsymbol{\alpha}^{p\boldsymbol{\nu}} \right)_{\boldsymbol{\nu} \in \mathcal{S}}$ ℓ^1 -summable. This gives $\left(\frac{|\boldsymbol{\nu}|!}{\boldsymbol{\nu}!} \right)^2 \boldsymbol{\alpha}^{2p\boldsymbol{\nu}} < 1$ as $\boldsymbol{\nu} \rightarrow \infty$ and there follows

$$\left(\frac{|\boldsymbol{\nu}|!}{\boldsymbol{\nu}!} \right)^2 \boldsymbol{\alpha}^{2\boldsymbol{\nu}} < \boldsymbol{\alpha}^{(2-2p)\boldsymbol{\nu}}, \text{ i.e., } b(\boldsymbol{\nu}) > (2-2p) \sum_{i=1}^N \lambda_i \nu_i \text{ as } \boldsymbol{\nu} \rightarrow \infty.$$

Next, define $g(\tau) = \frac{1}{\tau} \log \left(\frac{\Gamma(|\tau\boldsymbol{\nu}|+1)}{\prod_{i=1}^N \Gamma(\tau\nu_i+1)} \right)$ to be a mapping from $(0, \infty)$ to \mathbb{R} . We will prove $H_{\boldsymbol{\nu}}$ is decreasing by showing g is an increasing function. Observing that

$$g(\tau) = \sum_{q=2}^N \frac{1}{\tau} \log \left(\frac{\Gamma \left(\tau \sum_{i=1}^q \nu_i + 1 \right)}{\Gamma \left(\tau \sum_{i=1}^{q-1} \nu_i + 1 \right) \Gamma(\tau\nu_q + 1)} \right),$$

without loss of generality, we can assume $N = 2$. Consider the first derivative of g :

$$\begin{aligned} g'(\tau) &= -\frac{1}{\tau^2} \log \left(\frac{\Gamma(\tau\nu_1 + \tau\nu_2 + 1)}{\Gamma(\tau\nu_1 + 1)\Gamma(\tau\nu_2 + 1)} \right) + \frac{1}{\tau^2} \frac{\Gamma'(\tau\nu_1 + \tau\nu_2 + 1)}{\Gamma(\tau\nu_1 + \tau\nu_2 + 1)} (\tau\nu_1 + \tau\nu_2) \\ &\quad - \frac{1}{\tau^2} \frac{\Gamma'(\tau\nu_1 + 1)}{\Gamma(\tau\nu_1 + 1)} \tau\nu_1 - \frac{1}{\tau^2} \frac{\Gamma'(\tau\nu_2 + 1)}{\Gamma(\tau\nu_2 + 1)} \tau\nu_2. \end{aligned}$$

Then $g'(\tau) \geq 0 \ \forall \tau > 0$ iff $h(\nu_1 + \nu_2) \geq h(\nu_1) + h(\nu_2)$, $\forall \nu_1, \nu_2 \geq 0$, where $h(s) := s\psi_0(s+1) - \log(\Gamma(s+1))$.

We have $h''(s) = s\psi_2(s+1) + \psi_1(s+1) > 0$ for any $s \geq 0$, see Theorem 1, [17], so h is convex. Combining with the fact that $h(0) = 0$, this implies the superadditivity of h in $[0, \infty)$, as desired. Note that for $\boldsymbol{\nu} \in (0, \infty)^N$, g is strictly increasing in $(0, \infty)$.

Since $H_{\boldsymbol{\nu}}$ is decreasing, define the limiting set $\mathcal{P} = \bigcup_{\tau \in \mathbb{R}^+} \left(\frac{1}{\tau} \mathcal{P}_{\tau} \right)$. We will characterize \mathcal{P} and show it is Jordan measurable. Without loss of generality, we can ignore the set of points of \mathcal{P} in the coordinate hyperplanes, since it is of measure zero. Using the strictly increasing property of g for $\boldsymbol{\nu} \in (0, \infty)^N$, it gives

$$\bigcup_{\tau \in \mathbb{R}^+} \left(\frac{1}{\tau} \mathcal{P}_{\tau} \right) = \left\{ \boldsymbol{\nu} \in (0, \infty)^N : \sum_{i=1}^N \lambda_i \nu_i - \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \log \frac{\Gamma(|\tau\boldsymbol{\nu}| + 1)}{\prod_{i=1}^N \Gamma(\tau\nu_i + 1)} < \frac{1}{2} \right\}$$

Applying Stirling's formula, see, e.g., [1], yields

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \log \frac{\Gamma(|\tau \boldsymbol{\nu}| + 1)}{\prod_{i=1}^N \Gamma(\tau \nu_i + 1)} &= \log \lim_{\tau \rightarrow \infty} \left(\frac{|\tau \boldsymbol{\nu}|^{|\tau \boldsymbol{\nu}| + \frac{1}{2}} e^{-|\tau \boldsymbol{\nu}|} (2\pi)^{\frac{1}{2}}}{\prod_{i=1}^N (\tau \nu_i)^{\tau \nu_i + \frac{1}{2}} e^{-\tau \nu_i} (2\pi)^{\frac{1}{2}}} \right)^{\frac{1}{\tau}} \\ &= \log \lim_{\tau \rightarrow \infty} \frac{\tau^{|\boldsymbol{\nu}| + \frac{1}{2\tau}} |\boldsymbol{\nu}|^{|\boldsymbol{\nu}| + \frac{1}{2\tau}} (2\pi)^{\frac{1-N}{2\tau}}}{\tau^{|\boldsymbol{\nu}| + \frac{N}{2\tau}} \prod_{i=1}^N \nu_i^{\nu_i + \frac{1}{2\tau}}} = \log \frac{|\boldsymbol{\nu}|^{|\boldsymbol{\nu}|}}{\prod_{i=1}^N \nu_i^{\nu_i}}, \end{aligned}$$

and we obtain

$$\mathcal{P} = \left\{ \boldsymbol{\nu} \in (0, \infty)^N : \sum_{i=1}^N \lambda_i \nu_i - \log \frac{|\boldsymbol{\nu}|^{|\boldsymbol{\nu}|}}{\prod_{i=1}^N \nu_i^{\nu_i}} < \frac{1}{2} \right\}.$$

For the Jordan measurability of \mathcal{P} , we prove \mathcal{P} is convex. It is enough to show the function $G(\boldsymbol{\nu}) := \log \frac{|\boldsymbol{\nu}|^{|\boldsymbol{\nu}|}}{\prod_{i=1}^N \nu_i^{\nu_i}}$ is concave in $(0, \infty)^N$. Denote by $\nabla^2 G$ the Hessian matrix of G and again assume $N = 2$, we have

$$\nabla^2 G = \begin{pmatrix} 1/(\nu_1 + \nu_2) - 1/\nu_1 & 1/(\nu_1 + \nu_2) \\ 1/(\nu_1 + \nu_2) & 1/(\nu_1 + \nu_2) - 1/\nu_2 \end{pmatrix}.$$

Let $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$, it gives $\boldsymbol{x}^\top (\nabla^2 G) \boldsymbol{x} = \frac{(x_1 + x_2)^2}{\nu_1 + \nu_2} - \frac{x_1^2}{\nu_1} - \frac{x_2^2}{\nu_2} \leq 0$, by employing Cauchy-Schwarz inequality. Thus, $\nabla^2 G$ is negative semidefinite, which implies the concavity of G .

We can apply Theorem 2 to get the asymptotic estimate

$$\sum_{\boldsymbol{\nu} \notin \Lambda_M} \alpha^{2\nu} \left(\frac{|\boldsymbol{\nu}|!}{\boldsymbol{\nu}!} \right)^2 \leq C_u(\varepsilon) M \exp \left(- \left(\frac{M}{|\mathcal{P}|(1 + \varepsilon)} \right)^{1/N} \right).$$

The proof is now complete. \square

6.3. A computational comparison of our proposed estimate with previously established rates of convergence. Most of the established explicit error estimates for quasi-optimal approximations concerns the coefficient bounds of the form

$$\|t_{\boldsymbol{\nu}}\|_{V(D)} \leq \rho^{-\nu}. \quad (6.13)$$

We therefore compare our result with those from other approaches in estimating the truncation error of $\sum_{\boldsymbol{\nu} \in \mathcal{S}} \|t_{\boldsymbol{\nu}}\|_{V(D)}$, given (6.13). Recall, we proved in Proposition 3 that

$$\sum_{\boldsymbol{\nu} \notin \Lambda_M} \|t_{\boldsymbol{\nu}}\|_{V(D)} \leq C_u(\varepsilon) M \exp \left(- \left(\frac{MN! \prod_{i=1}^N \lambda_i}{(1 + \varepsilon)} \right)^{\frac{1}{N}} \right). \quad (6.14)$$

In [14], application of the Stechkin estimation gives $\sum_{\nu \notin \Lambda_M} \|t_\nu\|_{V(D)} \leq \|(\|t_\nu\|)\|_{\ell^p(\mathcal{S})} M^{1-\frac{1}{p}}$ for every $0 < p < 1$. Applying (6.13), there holds

$$\sum_{\nu \notin \Lambda_M} \|t_\nu\|_{V(D)} \leq \left(\prod_{i=1}^N \frac{1}{1 - e^{-p\lambda_i}} \right)^{1/p} M^{1-\frac{1}{p}}. \quad (\text{stech})$$

We note that Stechkin inequality holds for every M and can accommodate a wider class of approximation problems than those considered herein, including best- M term approximations. Later development due to [6] computes $p \in (0, 1)$ minimizing (stech) for each M and obtains

$$\sum_{\nu \notin \Lambda_M} \|t_\nu\|_{V(D)} \leq M \exp \left(-\frac{1}{e} \left(M \prod_{i=1}^N \lambda_i \right)^{1/N} N\xi \right), \quad (\text{optim})$$

where ξ is the rate adjusting parameter varying from 0 to $(e-1)/e$. Large ξ gives stronger convergence but also require more restrictive minimum cardinality. The best convergence is only guaranteed in the limit $M \rightarrow \infty$.

Figure 4 shows a comparison of our error estimate with (stech) and (optim) in computing the series $\sum_{\nu \in \mathcal{S}} e^{-(\nu_1 + \nu_2 + 2\nu_3 + 4\nu_4)}$ (corresponding to sequence (P.2) in Section 4). (optim) is plotted at its best possible rate with $\xi = (e-1)/e$. We also plot the exact value of $\sum_{\nu \notin \Lambda_M} e^{-b(\nu)}$, which can be calculated using Ehrhart polynomial in this case, for reference.

We observe that while (stech) holds for any rate $M^{1-\frac{1}{p}}$, the attached coefficient is very large with small p and strong rates are not effective except at high cardinality; (optim) is slightly above (stech), and both of them show considerable discrepancy with the exact truncation error, verifying Stechkin inequality is not sharp. Estimate (6.14), on the other hand, is close to the true value, even with ε large. Besides, the actual minimum cardinality for the estimate to hold is shown more optimistic than those proven in theory: $M_\varepsilon \simeq 1$ for $\varepsilon = 4.0$, $M_\varepsilon \simeq 10$ for $\varepsilon = 1.0$ and $M_\varepsilon \simeq 10^3$ for $\varepsilon = 0.3$. For comparison, from Figure 2, the theoretical values are 10^2 , 10^3 and 10^5 respectively. Also notice that $(N!)^{1/N} \simeq N/e$, (optim) and (6.14) are similar, except for the rate adjusting parameters. While $1/(1+\varepsilon)^{1/N}$ in (6.14) can be close to 1, ξ is bounded by $(e-1)/e \simeq 0.65$, resulting in the best convergence attainable by (optim) approximately $M \exp \left(-0.65 \left(\frac{M}{|\mathcal{P}|} \right)^{1/N} \right)$.

We consider next the problem of finding a tight upper bounds of

$$\text{error} := \sup_{\mathbf{y} \in \Gamma} \left\| u(\mathbf{y}) - \sum_{\nu \in \Lambda_M} t_\nu \mathbf{y}^\nu \right\|_{V(D)},$$

assuming $u(\mathbf{z})$ is a holomorphic function in an open neighborhood of the polydisc \mathcal{O}_ρ with $\rho_1 = \dots = \rho_N > 1 \ \forall N$. We note that (6.14) holds here, since the exponential decay (6.13) occurs (see Section 3.3), with $\lambda_i = \log \rho_i =: \lambda, \forall i$. An isotropic estimate introduced in [6],

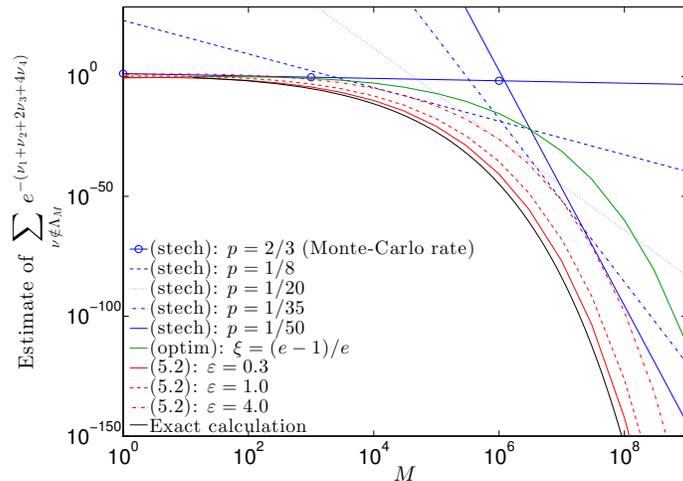


Figure 4: A comparison of our error estimate in computing the series $\sum_{\nu \in \mathcal{S}} e^{-(\nu_1 + \nu_2 + 2\nu_3 + 4\nu_4)}$ with those resulting from some previous approaches.

when applied to this error, gives

$$\text{error} \leq (1 - e^{-\lambda/2})^{-N} \exp\left(\frac{\lambda N}{2e} \log(1 - \epsilon) \sqrt[N]{M}\right), \quad (\text{optim-b})$$

where $\epsilon = \frac{e-1}{e} \left(1 - \frac{1.09}{\sqrt[N]{M}}\right)$. This bound is obtained based on an optimization of a Stechkin-type estimation, also presented in [6],

$$\text{error} \leq (1 - e^{-\lambda/2})^{-N} M^{-1/p} (1 - e^{-p\lambda/2})^{-N/p}, \quad (\text{steck-b})$$

for $p > 0$. Another nice result due to [4, 6], employing complex analysis technique, proves

$$\text{error} \leq \frac{1}{e^\lambda - 1} e^{-\lambda J},$$

for $M = \binom{N+J}{J}$, which implies

$$\text{error} \leq \frac{1}{e^\lambda - 1} \exp\left(-\lambda(MN!)^{1/N}\right) \quad (\text{complex})$$

in asymptotic regime.

Figure 5 plots estimate (6.14) and the upper bounds listed above in case $\lambda = 1$ and $N = 8$ (corresponding to sequence (P.3) in Section 4). The exact truncation error in computing the series $\sum_{\nu \in \mathcal{S}} \exp(-\sum_{i=1}^8 \nu_i)$ is also shown. It is interesting to see the (optim-b) curve is almost tangent to the (steck-b) lines, elucidating that (optim-b) is obtained by an optimization of (steck-b). Again, estimate (6.14) exhibits a much better approximation of the exact truncation error than (steck-b) and (optim-b). It should, however, be noted

that (optim-b) is proved to hold with relatively small cardinalities ($M > 1.09^N$), which are not covered by our analysis. The best convergence rate here is given by (complex). The advantage of (complex) lies in the fact that unlike other approaches, it seeks to approximate the remainder of Taylor series $\left\| u(\mathbf{y}) - \sum_{\nu \in \Lambda_M} t_\nu \mathbf{y}^\nu \right\|_V$ directly without using triangle inequality. Figure 5 shows a discrepancy between (complex) and exact calculation of $\sum_{\nu \in S} \exp(-\sum_{i=1}^8 \nu_i)$, revealing triangle inequality is not sharp in all cases. We are, unfortunately, not aware of an extension of (complex) outside the isotropic setting.

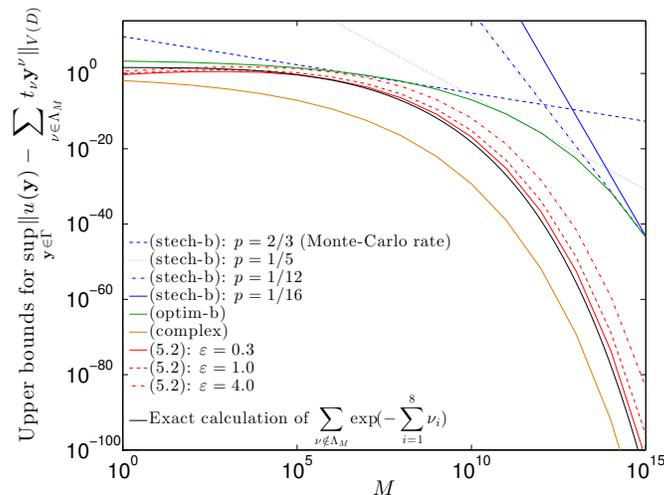


Figure 5: A comparison of our error estimate with those resulting from some previous approaches in an isotropic setting.

7. Concluding remarks. We present a new approach for analyzing the convergence of quasi-optimal Taylor and Legendre approximations for parameterized PDEs with deterministic and stochastic coefficients. The advantage of our analysis framework, which is demonstrated through several theoretical examples herein, includes its applicability to a general class of quasi-optimal polynomial approximations and the sharp estimates of their asymptotic errors. This work is restricted to linear elliptic equations with input coefficients depending affinely on the parameter. We expect similar results to hold in different settings with finite parametric dimension, particularly nonlinear elliptic PDEs, initial value problems and parabolic equations [12, 22, 23, 25], as our analysis only depends on the polynomial coefficient estimates.

Developing algorithms for identifying quasi-optimal subspaces is the next natural and essential step. Two potential types of procedures for building the subspaces corresponding to sharp estimates of the coefficients c_ν includes a priori and a posteriori approaches. In the first approach, the estimates for c_ν are derived a priori using knowledge on the input coefficient $a(x, \mathbf{y})$. Analytical studies reveal that if the complex continuation of $a(x, \mathbf{y})$ is an analytic function in \mathbb{C}^N then a theoretical decaying rate $\rho^{-\nu}$ of c_ν (with $\rho = (\rho_i)_{1 \leq i \leq N}$ representing the size of certain N -dimensional complex domains where real part of $a(x, \mathbf{y})$ is bounded away from 0) can be proved. The exploration of polynomial subspaces thus reduces

to the specification of such domains (or ρ in particular), which is expectedly significantly less computational demanding. Recent study [6] for a priori constructed Total Degree subspace found that while the theoretical estimates were not sharp, they could still provide good prediction on the anisotropy of the index sets. However, in practice, most analytical coefficient bounds lead to subspaces much more complicated than Total Degree and the determination of ρ in several cases is nontrivial, see [6, 12–14]. It is important to develop, implement and test of effectiveness of a priori algorithms in such settings.

Research on *a posteriori* procedures may be pursued in three directions. The first strategy finds the quasi-optimal index set using the theoretical coefficient estimates, but with ρ determined sharply in an a posteriori manner (by exact calculation of the decaying rate of c_ν in each direction i , $i = 1, \dots, N$), instead of a priori (by the definition of $a(x, \mathbf{y})$ as in above). The second strategy adaptively builds nested sequence $(\Lambda_M)_{M \geq 0}$ of quasi-optimal index sets Λ_M at a cost that scales linearly in $\#(\Lambda_M)$. Given Λ_M , we construct Λ_{M+1} by enriching Λ_M with the most effective indices ν in its neighborhood (denoted by $\mathcal{M}(\Lambda_M)$), which results in the best residual reduction. The third strategy first evaluates $u(\mathbf{y})$ on certain finite subset of Γ and then constructs the quasi-optimal subspace based on estimates of coefficients $c_\nu = \int_\Gamma u(\mathbf{y}) \Psi_\nu(\mathbf{y}) d\mathbf{y}$ using non-intrusive methods, e.g., Monte-Carlo, collocation. We expect the exploration cost for this approach, mostly coming from the evaluation of $u(\mathbf{y})$, to be a fraction of cost for computing the solution.

Finally, the development of quasi-optimal methods for another class of polynomial approximation: non-intrusive interpolation or collocation methods, is an important problem to study. These methods are practical and convenient in that they allow the use of legacy, black-box deterministic numerical solver and the simultaneous approximation of parameterized solutions can be considered as a modular post-processing step. With observation that the accuracy of the interpolation operator \mathcal{I}_{Λ_M} is dictated by the inequality

$$\begin{aligned} \|u - \mathcal{I}_{\Lambda_M}[u]\|_{L^\infty(\Gamma)} &\leq (1 + \mathbb{L}_{\Lambda_M}) \inf_{v \in \mathbb{P}_{\Lambda_M}} \|u - v\|_{L^\infty(\Gamma)} \\ &\leq (1 + \mathbb{L}_{\Lambda_M}) \|u - u_{\Lambda_M}\|_{L^\infty(\Gamma)}, \end{aligned}$$

where \mathbb{L}_{Λ_M} denotes the Lebesgue constant, we expect that the interpolation schemes in the quasi-optimal subspaces recover the convergence rates described in this work. However, to construct a non-intrusive hierarchical interpolant, two difficult challenges need to be addressed. First, the number of interpolation points needs to remain equal to the dimension of the polynomial space, thus, they must be nested and increase linearly. Second, to guarantee the accuracy of $\mathcal{I}_{\Lambda_M}[u]$, the Lebesgue constant must grow slowly with respect to the total number of collocation points, and we will need to explore the selections of abscissas which optimize this growth.

Acknowledgements. The authors wish to graciously thank Prof. Ronald DeVore for his interested in our work, his patience in discussing the analysis of "best M -term" approximations, and his tremendously helpful insights into the theoretical developments we pursued in this paper.

This material is based upon work supported in part by the U.S. Air Force of Scientific Research under grant number 1854-V521-12 and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics

program under contract and award numbers ERKJ259 and ERKJE45; and by the Laboratory Directed Research and Development program at the Oak Ridge National Laboratory, which is operated by UT-Battelle, LLC., for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

REFERENCES

- [1] M. ABRAMOWITZ AND I. STEGUN, *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1965.
- [2] I. BABUSKA, F. NOBILE, AND R. TEMPONE, *A stochastic collocation method for elliptic partial differential equation with random input data*, SIAM J. Numer. Anal., 45 (2007), pp. 1005–1034.
- [3] I. BABUSKA, R. TEMPONE, AND G. ZOURARIS, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.
- [4] T. BAGBY, L. BOS, AND N. LEVENBERG, *Multivariate simultaneous approximation*, Constr. Approx., 18 (2002), pp. 569–577.
- [5] V. BALDONI, N. BERLINE, J. DELOERA, B. DUTRA, M. KÖPPE, S. MOREINIS, G. PINTO, M. VERGNE, AND J. WU, *A user's guide for LattE integrale v1.7.1*. software package LattE is available at <http://www.math.ucdavis.edu/~latte/>, 2013.
- [6] J. BECK, F. NOBILE, L. TAMELLINI, AND R. TEMPONE, *Convergence of quasi-optimal stochastic galerkin methods for a class of pdes with random coefficients*, Computers and Mathematics with Applications, 67 (2014), pp. 732–751.
- [7] J. BECK, R. TEMPONE, F. NOBILE, AND L. TAMELLINI, *On the optimal polynomial approximation of stochastic pdes by galerkin and collocation methods*, Math. Models and Methods Appl. Sci., 22 (2012).
- [8] M. BECK AND S. ROBINS, *Computing the Continuous Discretely: Integer-Point Enumeration in Polyhedra*, Springer, 2007.
- [9] M. BIERI, R. ANDREEV, AND C. SCHWAB, *Sparse tensor discretization of elliptic spdes*, SIAM J. Sci. Comput., 31 (2009), pp. 4281–4304.
- [10] A. BUFFA, Y. MADAY, A. PATERA, C. PRUD'HOMME, AND G. TURINICI, *A priori convergence of the greedy algorithm for the parametrized reduced basis method*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 595–603.
- [11] A. CHKIFA, A. COHEN, R. DEVORE, AND C. SCHWAB, *Sparse adaptive taylor approximation algorithms for parametric and stochastic elliptic pdes*, Modél. Math. Anal. Numér., 47 (2013), pp. 253–280.
- [12] A. CHKIFA, A. COHEN, AND C. SCHWAB, *Breaking the curse of dimensionality in sparse polynomial approximation of parametric pdes*, J. Math. Pures Appl., (2014), p. accepted.
- [13] A. COHEN, R. DEVORE, AND C. SCHWAB, *Convergence rates of best n -term galerkin approximations for a class of elliptic spdes*, Found Comput Math, 10 (2010), pp. 615–646.
- [14] ———, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic pdes*, Analysis and Applications, 9 (2011), pp. 11–47.
- [15] J. DELOERA, R. HEMMECKE, J. TAUZER, AND R. YOSHIDA, *Effective lattice point counting in rational convex polytopes*, Journal of Symbolic Computation, 38 (2004), pp. 1273–1302.
- [16] R. DEVORE, *Nonlinear approximation*, Acta. Numer, 7 (1998), pp. 51–150.
- [17] A. ELBERT AND A. LAFORGIA, *On some properties of the gamma function*, Proc. Amer. Math. Soc., 128 (2000), pp. 2667–2673.

- [18] G. S. FISHMAN, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1996.
- [19] O. FRINK, *Jordan measure and riemann integration*, Ann. of Math., 34 (1933), pp. 518–526.
- [20] R. GHANEM AND P. SPANOS, *Stochastic finite elements: a spectral approach*, Springer-Verlag, New York, 1991.
- [21] P. GRUBER, *Convex and Discrete Geometry*, Springer Grundlehren der mathematischen Wissenschaften, 2007.
- [22] M. HANSEN AND C. SCHWAB, *Analytic regularity and nonlinear approximation of a class of parametric semilinear elliptic pdes*, Math. Nachr., 286 (2013), pp. 832–860.
- [23] ———, *Sparse adaptive approximation of high dimensional parametric initial value problems*, Vietnam Journal of Mathematics, 41 (2013), pp. 181–215.
- [24] M. HERVÉ, *Analyticity in infinite-dimensional spaces*, De Gruyter, Berlin, 1989.
- [25] V. H. HOANG AND C. SCHWAB, *Sparse tensor galerkin discretizations for parametric and random parabolic pdes - analytic regularity and generalized polynomial chaos approximation*, SIAM J. Mathematical Analysis, 45 (2013), pp. 3050–3083.
- [26] L. LEWIN, *Polylogarithms and Associated Functions*, New York: North-Holland, 1981.
- [27] ———, ed., *Structural Properties of Polylogarithms*, Providence, RI: Amer. Math. Soc., 1991.
- [28] M. LOÈVE, *Probability theory. I*, vol. 45 of Graduate Texts in Mathematics, Springer-Verlag, New York, 4th ed., 1977.
- [29] ———, *Probability theory. II*, vol. 46 of Graduate Texts in Mathematics, Springer-Verlag, New York, 4th ed., 1978.
- [30] R. MILANI, A. QUARTERONI, AND G. ROZZA, *Reduced basis methods in linear elasticity with many parameters*, Comp. Meth. Appl. Mech. Engg., 197 (2008), pp. 4812–4829.
- [31] E. NEUMAN, *Inequalities and bounds for the incomplete gamma function*, Results. Math., 63 (2013), pp. 1209–1214.
- [32] F. NOBILE, R. TEMPONE, AND C. WEBSTER, *A sparse grid stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal., 46 (2008), pp. 2309–2345.
- [33] R. STANLEY, *Enumerative Combinatorics*, vol. I, Cambridge, 1997.
- [34] T. TAO, *An introduction to measure theory*, vol. 126 of Graduate Studies in Mathematics, American Mathematical Society, 2011.
- [35] R. TODOR AND C. SCHWAB, *Convergence rates of sparse chaos approximations of elliptic problems with stochastic coefficients*, IMA J. Numer. Anal., 27 (2007), pp. 232–261.
- [36] N. WIENER, *The homogeneous chaos*, Amer. J. Math., 60 (1938), pp. 897–936.