Vietnam Journal of Mathematics (2023) 51:151–175
https://doi.org/10.1007/s10013-022-00584-1

ORIGINAL ARTICLE

Sparse Polynomial Approximations for Affine Parametric Saddle Point Problems

Peng Chen[1] · Omar Ghattas[2,3,4]

Received: 10 December 2021 / Accepted: 15 July 2022 / Published online: 19 September 2022

Abstract

In this work we study convergence properties of sparse polynomial approximations for a class of affine parametric saddle point problems, including the Stokes equations for viscous incompressible flow, mixed formulation of diffusion equations for groundwater flow, time-harmonic Maxwell equations for electromagnetics, etc. Due to the lack of knowledge or intrinsic randomness, the (viscosity, diffusivity, permeability, permittivity, etc.) coefficients of such problems are uncertain and can often be represented or approximated by high- or countably infinite-dimensional random parameters equipped with suitable probability distributions, and the coefficients affinely depend on a series of either globally or locally supported basis functions, e.g., Karhunen–Loève expansion, piecewise polynomials, or adaptive wavelet approximations. We consider sparse polynomial approximations of the parametric solutions, in particular sparse Taylor approximations, and study their convergence properties for these parametric problems. Under suitable sparsity assumptions on the parametrization of the random coefficients, we show the algebraic convergence rates $O(N^{-r})$ for the sparse polynomial approximations of the parametric solutions based on the results for affine parametric elliptic PDEs (Cohen, A. et al.: Anal. Appl. **9**, 11–47, 2011), (Bachmayr, M., et al.: ESAIM Math. Model. Numer. Anal. **51**, 321–339, 2017), (Cohen, A., DeVore, R.: Acta Numer. **24**, 1–159, 2015), (Chkifa, A., et al.: J. Math. Pures Appl. **103**, 400–428, 2015), (Chkifa, A., et al.: ESAIM Math. Model. Numer. Anal. **47**, 253–280, 2013), (Cohen, A., Migliorati, G.: Contemp. Comput. Math., 233–282, 2018), with the rate r depending only on a sparsity parameter in the parametrization, not on the number of active parameter dimensions or the number of polynomial terms N. We note that parametric saddle point problems were considered in (Cohen, A., DeVore, R.: Acta Numer. **24**, 1–159, 2015, Section 2.2) with the anticipation that the same results on the approximation of the solution map obtained for elliptic PDEs can be extended to more general saddle point problems. In this paper, we consider a general formulation of saddle point problems, different from that presented in (Cohen, A., DeVore, R.: Acta Numer. **24**, 1–159, 2015, Section 2.2), and obtain convergence rates for the two variables, e.g., velocity and pressure in Stokes equations, which are different for the case of locally supported basis functions.

This paper is dedicated to Prof. Alfio Quarteroni on the occation of his 70th birthday.

✉ Peng Chen
peng@cc.gatech.edu

Extended author information available on the last page of the article.

Keywords Sparse polynomial approximation · Saddle point problems · Parametric PDEs · Dimension-independent convergence

Mathematics Subject Classification (2010) 41A10 · 41A58 · 41A63 · 65N15

1 Introduction

Computational simulations based on mathematical models are increasing used for decision making (design, control, allocation of resources, determination of policy, etc.). For such cases, it is critical to account for uncertainties in the inputs, and thus output predictions of these models. One fundamental approach to characterize these uncertainties is by probabilistic modeling, where the uncertain input can be represented by a finite number of random variables or by random fields that can be represented by a large or even infinite number of random variables. We refer to these random variables as parameters and equip them with suitable probability measures. With these parameters as uncertain inputs, we often need to conduct statistical analysis of the model outputs, such as sensitivity analysis with respect to the parameters, computation of statistical moments via integration of outputs in the parameter space, and risk analysis that predicts the failure probability of the system under the uncertainty. To perform these statistical analyses, various numerical approximation methods have been developed largely in the last few decades, such as Monte Carlo and quasi Monte Carlo methods, generalized polynomial chaos, stochastic collocation and Galerkin methods, and model and parameter reduction methods.

The Monte Carlo method has been widely employed in practice because of several advantages, such as very simple and embarrassingly parallel implementation and dimension-independent convergence. However, it has a slow convergence rate of $O(N^{-1/2})$, where N is the number of samples, requiring a large number of simulations to achieve sufficient accuracy. New methods such as (high-order) quasi Monte Carlo [29, 36] and multi-level/multi-index Monte Carlo [22, 32] have been proposed to achieve faster convergence and reduced computational cost. Sparse polynomial approximations such as stochastic Galerkin and collocation methods based on (generalized) polynomial chaos and sparse grids have been developed that improve the convergence to a great extent for problems depending smoothly on the parameters; see, e.g., [1, 2, 31, 40, 49, 50]. Practical algorithms to construct such sparse polynomial approximations, such as adaptive [13, 30], least-squares [20, 39], and compressive sensing [27, 43] constructions, have also been actively developed. Another class of methods known as model reduction, including reduced basis methods, achieve quasi optimal convergence (in terms of Kolmogorov widths [7]) and considerable computational reduction for many-query simulations [5, 7, 9, 10, 15] by exploring the intrinsic structure of the solution manifolds. More recently, deep neural networks has been applied to solve high-dimensional parametric problems [6, 37, 38, 41, 46].

One critical challenge faced by polynomial based approximation methods for high-dimensional parametric problems is the so-called curse of dimensionality, i.e., convergence rates that severely deteriorate with the parameter dimension. In recent work [3, 4, 11, 17, 18, 23–25, 28, 46, 48, 52], it has been demonstrated that the curse of dimensionality can be effectively broken with dimension-independent convergence rates achieved under certain sparsity assumptions on the countably infinite-dimensional parametrization of the uncertain input. For instance, in [25] analytic regularity of the parametric solution with respect to the parameters was obtained for elliptic partial differential equations. This leads to upper bounds for the coefficients of Taylor expansion of the parametric solution. Under an ℓ^s-summability of the coefficients of the expansion that represent the random input, the Taylor coefficients were demonstrated to also satisfy the ℓ^s-summability. Then a dimension-independent convergence rate of a sparse Taylor approximation—truncation of a Taylor expansion of the parametric solution into a suitable sparse index set—were achieved by Stechkin's lemma. This analysis has been extended to sparse Legendre polynomial approximation [25], sparse polynomial interpolation [21], and sparse polynomial integration [44] for elliptic problems as well as for certain parabolic and nonlinear problems [18, 23].

In this work, we consider affine parametric saddle point problems that cover a wide range of applications, such as the Stokes equations for viscous incompressible flow, mixed formulation of the Poisson equation for groundwater flow, and time-harmonic Maxwell equations for electromagnetic wave propagation; see [8, 42] and the references therein. These applications require better understanding of the approximability and convergence of parametric saddle point problems in a high- or infinite-dimensional parametric setting, which is the aim and main contribution of this work based on the results for affine parametric elliptic PDEs [4, 18, 19, 23, 25, 26]. In particular, our contributions are presented in several sections structured as follows: In Section 2, we formulate an abstract saddle point problem with affine parametrization, and demonstrate the well-posedness of the parametric saddle point problem through several specific examples. Moreover, we consider both globally and locally supported basis functions for the affine parametrization with suitable sparsity assumptions for each of them. In Section 3, we consider a Taylor expansion of the solution of the parametric saddle point problem with respect to the parameters and its sparse Taylor approximation. In the case of globally supported basis functions, we prove the analytic regularity of the parametric solution with respect to the parameters, and prove the ℓ^s-summability of the Taylor coefficients. In the case of locally supported basis functions, we prove a weighted ℓ^2-summability of the Taylor coefficients, based on which we obtain the ℓ^s-summability of the Taylor coefficients. Based on the ℓ^s-summability, we prove dimension-independent convergence rates of the sparse Taylor approximations, for both arbitrary sparse index set and a downward closed sparse index set. In particular, our formulation of the saddle point problems is different from that presented in [23, Section 2.2], and leads to convergence results for the two variables in our saddle point formulation of the three examples, e.g., velocity and pressure in Stokes equations, which are different for the case of locally supported basis functions. This is not considered in [23, Section 2.2]. The last section provides conclusions and several ongoing and future research directions.

2 Affine Parametric Saddle Point Problems

2.1 An Abstract Saddle Point Formulation

Let $\mathcal{V}$ and $\mathcal{Q}$ denote two Hilbert spaces equipped with inner products $(\cdot,\cdot)_{\mathcal{V}}$, $(\cdot,\cdot)_{\mathcal{Q}}$ and induced norms $\|v\|_{\mathcal{V}} = (v,v)_{\mathcal{V}}^{1/2}$ $\forall v \in \mathcal{V}$, and $\|\cdot\|_{\mathcal{Q}}^2 = (q,q)_{\mathcal{Q}}^{1/2}$ $\forall q \in \mathcal{Q}$. Let $\mathcal{V}'$ and $\mathcal{Q}'$ denote the duals of $\mathcal{V}$ and $\mathcal{Q}$, respectively. Let $\mathcal{K}$ denote a separable Banach space. We present an abstract formulation of the parametric saddle point problem as: given parameter $\kappa \in \mathcal{K}$, and data $f \in \mathcal{V}'$ and $g \in \mathcal{Q}'$, find $(u,p) \in \mathcal{V} \times \mathcal{Q}$ such that

$$\begin{cases} a(u,v;\kappa) + b(v,p) = f(v) & \forall v \in \mathcal{V}, \\ \qquad\qquad\quad b(u,q) = g(q) & \forall q \in \mathcal{Q}, \end{cases} \tag{1}$$

where the linear forms $f(v)$ and $g(q)$ represent the duality pairing $\langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}}$ and $\langle g, q\rangle_{\mathcal{Q}'\times\mathcal{Q}}$ for simplicity, $a(\cdot,\cdot;\kappa) : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ is a parametric bilinear form, and $b(\cdot,\cdot) : \mathcal{V}\times\mathcal{Q} \to \mathbb{R}$ is a bilinear form. Moreover, we make the following assumptions on the bilinear forms. First, let $\mathcal{V}^0$ denote the kernel of the bilinear form b in $\mathcal{V}$, i.e.,

$$\mathcal{V}^0 := \{v \in \mathcal{V} : b(v,q) = 0\ \forall q \in \mathcal{Q}\}.$$

Assumption 2.1 Suppose the bilinear forms $a(\cdot,\cdot;\kappa)$ and $b(\cdot,\cdot)$ are uniformly continuous, i.e., there exist constants $\gamma > 0$ independent of κ and $\delta > 0$ such that

$$\begin{aligned} a(w,v;\kappa) &\le \gamma\|w\|_{\mathcal{V}}\|v\|_{\mathcal{V}} \quad \forall w,v \in \mathcal{V},\\ b(v,q) &\le \delta\|v\|_{\mathcal{V}}\|q\|_{\mathcal{Q}} \quad \forall v \in \mathcal{V}, q \in \mathcal{Q}. \end{aligned} \tag{2}$$

Moreover, we assume that $a(\cdot,\cdot;\kappa)$ is uniformly coercive in $\mathcal{V}^0$, i.e., there exists a constant $\alpha > 0$ independent of κ such that

$$a(v,v;\kappa) \ge \alpha\|v\|_{\mathcal{V}}^2 \quad \forall v \in \mathcal{V}^0. \tag{3}$$

Furthermore, we assume that $b(\cdot,\cdot)$ satisfies the inf-sup (compatibility) condition, i.e., there exists a constant $\beta > 0$ such that

$$\inf_{q\in\mathcal{Q}}\sup_{v\in\mathcal{V}} \frac{b(v,q)}{\|v\|_{\mathcal{V}}\|q\|_{\mathcal{Q}}} \ge \beta.$$

The classical results of existence, uniqueness, and a-priori estimates for the parametric saddle point problem (1) are stated in the following theorem.

Theorem 2.1 [42, Theorem 16.4] *Under Assumption 2.1, for every $\kappa \in \mathcal{K}$, there exists a unique solution $(u,p) \in \mathcal{V}\times\mathcal{Q}$ of the parametric saddle point problem* (1)*, such that the following a-priori estimates hold*

$$\|u\|_{\mathcal{V}} \le C_u < \infty \quad \textit{and} \quad \|p\|_{\mathcal{Q}} \le C_p < \infty, \tag{4}$$

where for notational convenience, the constants C_u and C_p are short for

$$C_u = \frac{1}{\alpha}\|f\|_{\mathcal{V}'} + \frac{\alpha+\gamma}{\alpha\beta}\|g\|_{\mathcal{Q}'} \quad \textit{and} \quad C_p = \frac{\alpha+\gamma}{\alpha\beta}\|f\|_{\mathcal{V}'} + \frac{\gamma(\alpha+\gamma)}{\alpha\beta^2}\|g\|_{\mathcal{Q}'}. \tag{5}$$

Remark 2.1 We remark that the saddle point problem considered in [23, Section 2.2] has the form: given parameter $\kappa \in \mathcal{K}$, find $u \in \mathcal{V}$ such that

$$B(u,v;\kappa) = L(v;\kappa) \quad \forall v \in \mathcal{V},$$

where the bilinear form B satisfies inf-sup condition, which is different from what we consider in (1) to cover the examples in the next section. Moreover, we can obtain different convergence rates for sparse polynomial approximations of u and p in (1), as shown in Section 3.

2.2 Examples

Let $D \subset \mathbb{R}^d$ ($d = 2, 3$) be an open and bounded physical domain with Lipschitz continuous boundary $\partial D = \Gamma$, which can be aligned to Dirichlet boundary Γ_0 and Neumann boundary Γ_1 such that $\Gamma = \Gamma_0 \cup \Gamma_1$ and $\Gamma_0 \cap \Gamma_1 = \emptyset$. Let $L^\infty(D)$ denote a space of essentially bounded measurable functions, i.e.,

$$L^{\infty}(D) = \left\{ v : \operatorname*{ess\,sup}_{x \in D} |v(x)| = \|v\|_{L^{\infty}(D)} < \infty \right\}.$$

Let $L^2(D)$ denote a space of square integrable functions on D, i.e.,

$$L^2(D) = \left\{ v : \int_D |v|^2 dx = \|v\|^2_{L^2(D)} < \infty \right\}.$$

Let ∇, $\nabla \cdot$, $\nabla \times$ denote the gradient, divergence, and curl operators. We use the definition of the following Hilbert spaces by convention [8]

$$\begin{aligned} H^1(D) &:= \left\{ v \in L^2(D) : |\nabla v| \in L^2(D) \right\}, \\ H(\text{div}; D) &:= \left\{ \boldsymbol{v} \in (L^2(D))^d : \nabla \cdot \boldsymbol{v} \in L^2(D) \right\}, \\ H(\text{curl}; D) &:= \left\{ \boldsymbol{v} \in (L^2(D))^d : \nabla \times \boldsymbol{v} \in (L^2(D))^d \right\}, \end{aligned}$$

with corresponding norms

$$\begin{aligned} \|v\|^2_{H^1(D)} &:= \|v\|^2_{L^2(D)} + \|\nabla v\|^2_{L^2(D)}, \\ \|\boldsymbol{v}\|^2_{H(\text{div};D)} &:= \|\boldsymbol{v}\|^2_{(L^2(D))^d} + \|\nabla \cdot \boldsymbol{v}\|^2_{L^2(D)}, \\ \|\boldsymbol{v}\|^2_{H(\text{curl};D)} &:= \|\boldsymbol{v}\|^2_{(L^2(D))^d} + \|\nabla \times \boldsymbol{v}\|^2_{L^2(D)}. \end{aligned}$$

Moreover, for functions with vanishing values on Γ_0, we define

$$\begin{aligned} H^1_0(D) &:= \left\{ v \in H^1(D) : v = 0 \text{ on } \Gamma_0 \right\}, \\ H_0(\text{div}; D) &:= \{ v \in H(\text{div}; D) : \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma_0 \}, \\ H_0(\text{curl}; D) &:= \{ \boldsymbol{v} \in H(\text{curl}; D) : \boldsymbol{v} \times \boldsymbol{n} = 0 \text{ on } \Gamma_0 \}, \end{aligned}$$

where $\boldsymbol{n}$ is the unit normal vector along the boundary. In what follows, we present several classical problems in (mixed) variational formulations. These formulations are preferred due to several reasons [8]: the presence of a physical constraint, physical importance of the variables appearing in the formulations, better accommodation of finite dimensional approximation and/or available data. For the simplicity of presentation, we assume homogeneous Dirichlet and/or Neumann boundary conditions for all the examples.

2.2.1 Stokes Flow

We consider a flow of a viscous incompressible fluid with low velocity in a domain D, which can be described by Stokes equations in the variational form as: given parameter $\kappa \in L^{\infty}(D)$, data $\boldsymbol{f} \in (L^2(D))^d$, find $(\boldsymbol{u}, p) \in (H^1_0(D))^d \times L^2(D)$ such that

$$\left\{ \begin{aligned} \int_D 2\kappa \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) dx - \int_D (\nabla \cdot \boldsymbol{v})\, p dx &= \int_D \boldsymbol{f} \cdot \boldsymbol{v} dx \quad \forall \boldsymbol{v} \in (H^1_0(D))^d, \\ \int_D (\nabla \cdot \boldsymbol{u})\, q dx &= 0 \quad \forall q \in L^2(D), \end{aligned} \right. \tag{6}$$

where $\boldsymbol{u}$ is the velocity, p is the pressure, $\kappa > 0$ is the shear viscosity, $\boldsymbol{f} \in \mathbb{R}^d$ is the body force, and $\boldsymbol{\varepsilon}(\boldsymbol{u}) \in \mathbb{R}^{d \times d}$ is the strain rate tensor defined as

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{1}{2} \left(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T \right).$$

Note that for the weaker condition $\boldsymbol{f} \in (H^{-1}(D))^d$, the formal expression $\int_D \boldsymbol{f} \cdot \boldsymbol{v}$ represents the duality pairing $\langle \boldsymbol{f}, \boldsymbol{v}\rangle_{\mathcal{V}'\times\mathcal{V}}$ with $\mathcal{V} = (H_0^1(D))^d$.

Problem (6) can be identified in the abstract saddle point formulation (1) in the spaces $\mathcal{K} = L^\infty(D)$, $\mathcal{V} = (H_0^1(D))^d$ and $\mathcal{Q} = L^2(D)$ with the bilinear forms

$$\begin{aligned} a(\boldsymbol{w}, \boldsymbol{v}; \kappa) &:= \int_D 2\kappa \boldsymbol{\varepsilon}(\boldsymbol{w}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) dx \quad \forall \boldsymbol{w}, \boldsymbol{v} \in \mathcal{V}, \\ b(\boldsymbol{v}, q) &:= -\int_D (\nabla \cdot \boldsymbol{v})\, q dx \quad \forall \boldsymbol{v} \in \mathcal{V}, \forall q \in \mathcal{Q}. \end{aligned}$$

Then Assumption 2.1 is satisfied with the constants

$$\gamma = 2\gamma_2 \operatorname*{ess\,sup}_{x\in D} \kappa(x), \quad \delta = 1, \quad \alpha = 2\gamma_1 \operatorname*{ess\,inf}_{x\in D} \kappa(x), \quad \text{and } \beta = \frac{1}{\sqrt{1 + C_p}}, \tag{7}$$

where the constants γ_1, γ_2 are determined by the Korn's inequality [34], i.e.,

$$\gamma_1 \|\boldsymbol{v}\|_{\mathcal{V}}^2 \le \int_D \boldsymbol{\varepsilon}(\boldsymbol{v}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) dx \le \gamma_2 \|\boldsymbol{v}\|_{\mathcal{V}}^2 \quad \forall \boldsymbol{v} \in \mathcal{V},$$

and C_p is determined by the Poincaré's inequality [42], i.e.,

$$\int_D |\boldsymbol{v}|^2 dx \le C_p \int_D |\nabla \cdot \boldsymbol{v}|^2 dx, \quad \forall \boldsymbol{v} \in \mathcal{V}.$$

Thus the inf-sup constant β is obtained as: for any $q \in \mathcal{Q}$, by taking $\nabla \cdot \boldsymbol{v} = q$,

$$\sup_{\boldsymbol{v}\in\mathcal{V}} \frac{b(\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{\mathcal{V}} \|q\|_{\mathcal{Q}}} \ge \frac{\|q\|_{\mathcal{Q}}^2}{\|\boldsymbol{v}\|_{\mathcal{V}} \|q\|_{\mathcal{Q}}} = \frac{\|\nabla \cdot \boldsymbol{v}\|_{L^2(D)}}{\|\boldsymbol{v}\|_{H^1(D)}} \ge \frac{1}{\sqrt{1 + C_p}} =: \beta.$$

Therefore, Theorem 2.1 holds for the Stokes problem (6) with these constants.

2.2.2 Diffusion

Diffusion equations are widely used in modelling various physical phenomena. In many applications it is the flux rather than the state that is of interesting. For instance in thermo-diffusion problems heat flux may be more important than the temperature field. For such consideration, we present the diffusion problem in the variational formulation: given parameter $\kappa \in L^\infty(D)$ and data $f \in L^2(D)$, find $(\boldsymbol{u}, p) \in H_0(\text{div}; D) \times L^2(D)$ such that

$$\begin{cases} \int_D \kappa \boldsymbol{u} \cdot \boldsymbol{v} dx + \int_D (\nabla \cdot \boldsymbol{v}) p dx = 0 \quad \forall \boldsymbol{v} \in H_0(\text{div}; D), \\ \int_D (\nabla \cdot \boldsymbol{u}) q dx = -\int_D f q dx \quad \forall q \in L^2(D), \end{cases} \tag{8}$$

where p is the state, e.g., temperature field, the auxiliary variable $\boldsymbol{u} = \kappa^{-1}\nabla p$ represents the flux, $\kappa > 0$ is the (inverse) diffusion coefficient, f is a source term.

By defining the bilinear forms

$$\begin{aligned} a(\boldsymbol{w}, \boldsymbol{v}; \kappa) &:= \int_D \kappa \boldsymbol{u} \cdot \boldsymbol{v} dx \quad \forall \boldsymbol{w}, \boldsymbol{v} \in \mathcal{V}, \\ b(\boldsymbol{v}, q) &:= \int_D (\nabla \cdot \boldsymbol{v}) q dx \quad \forall \boldsymbol{v} \in \mathcal{V}, \forall q \in \mathcal{Q}, \end{aligned}$$

in the Hilbert spaces $\mathcal{V} = H_0(\text{div}; D)$ and $\mathcal{Q} = L^2(D)$, we can identify the diffusion problem (8) in the abstract saddle point formulation (1) with $\mathcal{K} = L^\infty(D)$. Assumption 2.1

is satisfied with the following constants

$$\gamma = \operatorname*{ess\,sup}_{x\in D} \kappa(x), \quad \delta = 1, \quad \alpha = \operatorname*{ess\,inf}_{x\in D} \kappa(x), \quad \text{and } \beta = \frac{1}{\sqrt{1+C_p}},$$

where β is obtained the same as in the Stokes problem. Note that the bilinear form $a(\cdot,\cdot;\kappa)$ is coercive in $\mathcal{V}^0$, in which $\nabla\cdot\boldsymbol{v}$ vanishes, even it is not coercive in $\mathcal{V}$.

2.2.3 Time Harmonic Maxwell System

The foundation of classical electromagnetism, optics, and electric circuits can be described by Maxwell equations. The time harmonic Maxwell system is considered when the propagation of electromagnetic waves at a given frequency is studied or when the Fourier transform in time is used. In the mixed variational formulation, the Maxwell system can be stated as: given parameter $\kappa \in L^\infty(D)$, and data $\boldsymbol{f} \in (L^2(D))^d$, find $(\boldsymbol{u}, p) \in H_0(\text{curl}; D) \times H_0^1(D)$ such that

$$\begin{cases} \displaystyle\int_D \kappa(\nabla\times\boldsymbol{u})\cdot(\nabla\times\boldsymbol{v})dx - \omega^2\int_D \varepsilon\boldsymbol{u}\cdot\boldsymbol{v}dx + \int_D \nabla p\cdot\boldsymbol{v}dx \\ \qquad\qquad = \displaystyle\int_D \boldsymbol{f}\cdot\boldsymbol{v}dx \quad \forall\boldsymbol{v}\in H_0(\text{curl}; D), \\ \displaystyle\int_D \nabla q\cdot\boldsymbol{u}dx = 0 \quad \forall q\in H_0^1(D), \end{cases}$$

where $\boldsymbol{u}$ is the electric field vector, p is the auxiliary variable, ω is a frequency, $\boldsymbol{f} = i\omega\boldsymbol{j}$ with current source field vector $\boldsymbol{j}$, $\kappa > 0$ denotes the (inverse) magnetic permeability, $\varepsilon > 0$ denotes the electric permittivity. Here we only consider κ as a varying parameter and fix ε for simplicity.

By defining the bilinear forms

$$\begin{aligned} a(\boldsymbol{w},\boldsymbol{v};\kappa) &:= \int_D \kappa(\nabla\times\boldsymbol{u})\cdot(\nabla\times\boldsymbol{v})dx - \omega^2\int_D \varepsilon\boldsymbol{u}\cdot\boldsymbol{v}dx \quad \forall\boldsymbol{w},\boldsymbol{v}\in\mathcal{V}, \\ b(\boldsymbol{v},q) &:= \int_D \nabla p\cdot\boldsymbol{v}dx \quad \forall\boldsymbol{v}\in\mathcal{V}, \forall q\in\mathcal{Q}, \end{aligned}$$

in the Hilbert spaces $\mathcal{V} = H_0(\text{curl}; D)$ and $\mathcal{Q} = H_0^1(D)$, we can express the time harmonic Maxwell system as in the abstract saddle point formulation (1) with $\mathcal{K} = L^\infty(D)$. Moreover, we can verify Assumption 2.1 with the following constants

$$\gamma = \operatorname*{ess\,sup}_{x\in D} \kappa(x), \quad \delta = 1, \quad \text{and } \beta = \frac{1}{\sqrt{1+C_p}},$$

and

$$\alpha = \frac{1}{1+C_f}\left(\operatorname*{ess\,inf}_{x\in D}\kappa(x) - \omega^2 C_f \operatorname*{ess\,sup}_{x\in D}\varepsilon(x)\right).$$

We consider the case that $\alpha > 0$ in this work. It is straightforward to verify γ and δ. To verify β, for any $q\in\mathcal{Q}$, by taking $\boldsymbol{v} = \nabla q$, we have

$$\sup_{\boldsymbol{v}\in\mathcal{V}} \frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{\mathcal{V}}\|q\|_{\mathcal{Q}}} \geq \frac{\|\nabla q\|^2_{(L^2(D))^d}}{\|\nabla q\|_{(L^2(D))^d}\|q\|_{H_0^1(D)}} = \frac{\|\nabla q\|_{(L^2(D))^d}}{\|q\|_{H_0^1(D)}} \geq \frac{1}{\sqrt{1+C_p}} =: \beta,$$

noting that $\nabla \times \nabla q = 0, \forall q \in \mathcal{Q}$, in the first inequality. To verify α, by Friedrichs's inequality [8], there exists a constant C_f such that

$$\int_D |\boldsymbol{v}|^2 dx \leq C_f \int_D |\nabla \times \boldsymbol{v}|^2 dx \quad \forall \boldsymbol{v} \in \mathcal{V}.$$

Therefore, we have

$$\begin{aligned} a(\boldsymbol{v}, \boldsymbol{v}; \kappa) &\geq \operatorname*{ess\,inf}_{x \in D} \kappa(x) \int_D |\nabla \times \boldsymbol{v}|^2 dx - \omega^2 \operatorname*{ess\,sup}_{x \in D} \varepsilon(x) \int_D |\boldsymbol{v}|^2 dx \\ &\geq \left(\operatorname*{ess\,inf}_{x \in D} \kappa(x) - \omega^2 C_f \operatorname*{ess\,sup}_{x \in D} \varepsilon(x) \right) \int_D |\nabla \times \boldsymbol{v}|^2 dx \\ &\geq \frac{1}{1 + C_f} \left(\operatorname*{ess\,inf}_{x \in D} \kappa(x) - \omega^2 C_f \operatorname*{ess\,sup}_{x \in D} \varepsilon(x) \right) \|\boldsymbol{v}\|_{\mathcal{V}}^2 \quad \forall \boldsymbol{v} \in \mathcal{V}. \end{aligned}$$

2.3 Affine Parametrization

In this section, we present an affine parametrization for the parameter κ. We first present a common structure of the bilinear form $a(\cdot, \cdot; \kappa)$ in (1) appearing in many saddle point problems such as the Stokes equations, mixed formulation of the Poisson equation, and time-harmonic Maxwell's equations, that is affine with respect to the parameter $\kappa \in \mathcal{K}$, i.e., it can be written as

$$a(w, v; \kappa) = a_0(w, v) + a_1(w, v; \kappa) \quad \forall w, v \in \mathcal{V}, \tag{9}$$

where $a_1(w, v; \kappa)$ depends linearly on κ such that for any $\kappa \in \mathcal{K}$ there hold

$$\begin{aligned} a_1(v, v; \kappa) &\geq c_1 \operatorname*{ess\,inf}_{x \in D} |\kappa(x)| \, \|v\|_{\mathcal{V}}^2 \quad \forall v \in \mathcal{V}^0, \\ a_1(w, v; \kappa) &\leq C_1 \operatorname*{ess\,sup}_{x \in D} |\kappa(x)| \, \|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \quad \forall w, v \in \mathcal{V}, \\ a_1(w, v; \kappa) &\leq \frac{1}{2} \left(a_1(w, w; |\kappa|) + a_1(v, v; |\kappa|) \right) \quad \forall w, v \in \mathcal{V}, \end{aligned} \tag{10}$$

for constants $c_1, C_1 > 0$ independent of κ, e.g., related to the Poincaré's or Friedrichs's constant in Stokes equations or time-harmonic Maxwell's equations. We shall consider this affine structure (9) with the properties (10) in what follows.

To parametrize κ, we consider a countably infinite-dimensional parameter space

$$U = [-1, 1]^{\mathbb{N}}.$$

We denote the element of the parameter space as $\boldsymbol{y} = (y_j)_{j \geq 1} \in U$ and equip the parameter space with the probability measure

$$d\mu(\boldsymbol{y}) = \bigotimes_{j \geq 1} \frac{d\lambda(y_j)}{2},$$

where $d\lambda$ is the Lebesgue measure in $[-1, 1]$. To this end, we consider an affine parametrization for the representation and approximation of the parameter κ that is widely used in the literature [1, 2, 19, 22–25, 32, 47].

Assumption 2.2 The variation of the parameter κ in $\mathcal{K}$ can be represented by the parameter $\boldsymbol{y} \in U$ through the affine expansion

$$\kappa(x, \boldsymbol{y}) = \kappa_0(x) + \sum_{j \geq 1} y_j \kappa_j(x) \quad \forall (x, \boldsymbol{y}) \in D \times U \text{ and } \kappa_j \in \mathcal{K}, \forall j \geq 0. \tag{11}$$

Moreover, we assume there exist constants $0 < \theta < \Theta < \infty$ such that

$$\theta < \kappa_{\min} := \inf_{(x,\boldsymbol{y})\in D\times U} \kappa(x, \boldsymbol{y}) \leq \sup_{(x,\boldsymbol{y})\in D\times U} \kappa(x, \boldsymbol{y}) =: \kappa_{\max} < \frac{\Theta}{2},$$

and such that the coercivity and continuity conditions (3) and (2) are satisfied for the bilinear form $a(\cdot, \cdot; \kappa)$ at any $\kappa \in [\theta, \Theta]$.

The sequence $(\kappa_j)_{j\geq 0}$ could either be directly prescribed knowledge of the physical system or given by an affine representation or approximation of the random field κ. We present two specific examples, where we distinguish the parametrization in two classes representing globally and locally supported basis $(\kappa_j)_{j\geq 1}$, respectively.

1. *Globally supported basis*. One classical example comes from Karhunen–Loève expansion of a random field with finite second order moment, given by [45]

$$\kappa(x, \boldsymbol{y}) = \kappa_0(x) + \sum_{j=1}^{\infty} y_j \sqrt{\lambda_j} \psi_j(x), \tag{12}$$

where κ_0 is the mean of the random field, $(\lambda_j, \psi_j)_{j\geq 1}$ are the eigenpairs of the covariance of the random field. Here, we can identify $\kappa_j = \sqrt{\lambda_j}\psi_j$, $j \geq 1$, in the affine assumption (11).
2. *Locally supported basis*. Piecewise polynomials or wavelets can be employed to model or approximate the parameter field κ. A particular case is the weighted piecewise constant basis representation

$$\kappa(x, \boldsymbol{y}) = \kappa_0 + \sum_{j=1}^{J} y_j w_j \chi_j(x), \tag{13}$$

where w_j is the weight and χ_j is the characteristic function in the subdomain/element D_j, $j = 1, \ldots, J$, where $D = \cup_{j=1}^{J} D_j$ and $D_i \cap D_j = \emptyset$ for $i \neq j$. In this example, we can identify $\kappa_j = w_j \chi_j$, $j = 1, \ldots, J$.

Assumption 2.2 guarantees the well-posedness of the parametric saddle point problem (1). To study the convergence property of certain approximation of its solution or related quantity of interest, we make the following assumptions to cover the globally and locally supported basis representations as considered in [25] and [4], respectively.

Assumption 2.3 For the parametrization (11) under Assumption 2.2, assume for some $s \in (0, 1)$ there holds $(\|\kappa_j\|_{\mathcal{K}})_{j\geq 1} \in \ell^s(\mathbb{N})$, i.e.,

$$\sum_{j\geq 1} \|\kappa_j\|_{\mathcal{K}}^s < \infty. \tag{14}$$

Remark 2.2 As discussed in [4], for the Karhunen–Loève expansion (12), the ℓ^s-summability condition (14) is satisfied when $\sup_{j\geq 1} \|\psi_j\|_{\mathcal{K}} \leq C$ for some $C < \infty$, and $(\sqrt{\lambda_j})_{j\geq 1} \in \ell^s(\mathbb{N})$. However, it is not satisfied for any $s \in (0, 1)$ in the case of the locally supported representation (13) when $|w_j| \lesssim\!\!\!\!\!/ \; j^{-1}$, i.e., $(|w_j|)_{j\geq 1} \notin \ell^1(\mathbb{N})$, as $J \to \infty$. To accommodate such a case, we make the following assumption.

Assumption 2.4 For the parametrization (11) under Assumption 2.2, assume there exists a sequence $\boldsymbol{\rho} = (\rho_j)_{j\geq 1}$ with $\rho_j > 1$, such that

$$\sum_{j\geq 1} \rho_j |\kappa_j(x)| \leq \kappa_0(x) - \epsilon \quad \forall x \in D, \tag{15}$$

for some $\theta < \epsilon < \kappa_{\min}$, and such that $(\rho_j^{-1})_{j\geq 1} \in \ell^t(\mathbb{N})$ for some $t \in (0, \infty)$.

Remark 2.3 We can see that Assumption 2.4 is satisfied for the locally supported representation (13) for $J \to \infty$, as in [4]. For instance, we can take $\rho_j^{-1} \sim |w_j|$ and $\rho_j |w_j| \leq \kappa_{\min} - \epsilon$ as $|w_j| \to 0$, such that $\rho_j > 1$ and (15) holds, then $(\rho_j^{-1})_{j\geq 1} \in \ell^t(\mathbb{N})$ whenever $(|w_j|)_{j\geq 1} \in \ell^t(\mathbb{N})$ for any $t \in (0, \infty)$.

3 Sparse Polynomial Approximations

Let $\mathcal{F}$ denote a multi-index set with finitely supported multi-index $\boldsymbol{\nu} = (\nu_j)_{j\geq 1}$, i.e., $\boldsymbol{\nu} \in \mathcal{F}$ if and only if $|\boldsymbol{\nu}| = \sum_{j\geq 1} \nu_j < \infty$. For any $\boldsymbol{\nu} \in \mathcal{F}$, we define the multi-factorial $\boldsymbol{\nu}!$, multi-monomial $\boldsymbol{y}^{\boldsymbol{\nu}}$ for $\boldsymbol{y} \in U$, and partial derivative $\partial^{\boldsymbol{\nu}} \psi(\boldsymbol{y})$ for a differentiable parametric map $\psi(\boldsymbol{y})$ as

$$\boldsymbol{\nu}! := \prod_{j\geq 1} \nu_j!, \quad \boldsymbol{y}^{\boldsymbol{\nu}} := \prod_{j\geq 1} y_j^{\nu_j}, \quad \partial^{\boldsymbol{\nu}} \psi(\boldsymbol{y}) := \frac{\partial^{|\boldsymbol{\nu}|} \psi(\boldsymbol{y})}{\partial^{\nu_1} y_1 \partial^{\nu_2} y_2 \cdots},$$

where we use the convention $0! := 1$, $0^0 := 1$, and $\partial^0 \psi(\boldsymbol{y}) / \partial^0 y_j = \psi(\boldsymbol{y})$. For such a differentiable map ψ, we consider the Taylor power series

$$T_{\mathcal{F}} \psi(\boldsymbol{y}) := \sum_{\boldsymbol{\nu} \in \mathcal{F}} t_{\boldsymbol{\nu}}^{\psi} \boldsymbol{y}^{\boldsymbol{\nu}}, \tag{16}$$

with Taylor coefficients $t_{\boldsymbol{\nu}}^{\psi}$ defined as

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

Let $(\Lambda_N)_{N\geq 1} \subset \mathcal{F}$ denote a sequence of index sets with N indices that exhaust $\mathcal{F}$, i.e., any finite set $\Lambda \subset \mathcal{F}$ is contained in all Λ_N for $N \geq N_0$ with N_0 sufficiently large. We define the truncation of the power series (16) in Λ_N as

$$T_{\Lambda_N} \psi(\boldsymbol{y}) := \sum_{\boldsymbol{\nu} \in \Lambda_N} t_{\boldsymbol{\nu}}^{\psi}(\boldsymbol{y}) \boldsymbol{y}^{\boldsymbol{\nu}},$$

which we call *sparse Taylor approximation*. We are interested in two questions: (1) if the sparse Taylor approximation for the solution of the parametric saddle point problem (1) is convergent; (2) if so, how fast it converges with respect to N. To answer these questions, we carry out two types of analyses corresponding to Assumptions 2.3 and 2.4, respectively. The first type is to obtain the analytic regularity property of the parametric solution in a complex domain covering the parameter space. This analyticity leads to upper bounds for the Taylor coefficients $(t_{\boldsymbol{\nu}}^u, t_{\boldsymbol{\nu}}^p)$ at each $\boldsymbol{\nu} \in \mathcal{F}$ by Cauchy's integral formula, which implies a $\ell^s(\mathcal{F})$-summability of the coefficients. The second type is to derive a weighted $\ell^2(\mathcal{F})$-summability of the Taylor coefficient based on the affine structure of the parametrization; then the $\ell^s(\mathcal{F})$-summability of the Taylor coefficients is obtained by using Hölder's inequality. Due to the $\ell^s(\mathcal{F})$-summability, a best N-term dimension-independent convergence rate of a suitable Taylor approximation is achieved using Stechkin's lemma. These analyses are based on the results in [25] and [4] for studying parametric elliptic PDEs, which we extend to dealing with the parametric saddle point problem (1) under Assumptions 2.3 and 2.4, respectively.

3.1 ℓ^s-Summability by Analytic Regularity

Let $z = (z_j)_{j\geq 1}$ denote a sequence of complex numbers with $z_j \in \mathbb{C}$, $j \geq 1$, i.e., $z \in \mathbb{C}^{\mathbb{N}}$. Let $\mathcal{U}$ denote a polydisc defined as

$$\mathcal{U} := \left\{ z \in \mathbb{C}^{\mathbb{N}} : |z_j| \leq 1 \text{ for every } j \geq 1 \right\}.$$

Then we can extend the parametrization of κ in (11) from $U = [-1, 1]^{\mathbb{N}}$ to $\mathcal{U}$, i.e.,

$$\kappa(x, z) = \kappa_0(x) + \sum_{j\geq 1} z_j \kappa_j(x) \quad \forall (x, z) \in D \times \mathcal{U},$$

for which, under Assumption 2.2, we have

$$\kappa_{\min} \leq \Re(\kappa(x, z)) \leq |\kappa(x, z)| \leq 2\kappa_{\max}.$$

For two constants r and R such that

$$0 < \theta < r < \kappa_{\min} < 2\kappa_{\max} < R < \Theta < \infty,$$

where θ and Θ are given in Assumption 2.2, we define the complex set

$$\mathcal{A}_r^R = \left\{ z \in \mathbb{C}^{\mathbb{N}} : r \leq |\kappa(x, z)| \leq R \text{ for every } x \in D \right\}.$$

By the equivalence of Babuška Theorem and Brezzi Theorem for saddle point problems [51], and the extension of the Babuška theorem to complex function space [23, Theorem 2.2], Theorem 2.1 holds for $z \in \mathcal{A}_r^R$ under Assumptions 2.1 and 2.2 in complex function spaces $\mathcal{V}$ and $\mathcal{Q}$, i.e., there exists a unique solution $(u(z), p(z)) \in \mathcal{V} \times \mathcal{Q}$ $\forall z \in \mathcal{A}_r^R$, which satisfies the a-priori estimates in (4). In fact, Theorem 2.1 holds for $z \in \mathcal{A}_{\tilde{r}}^R$ for any $\tilde{r} \geq \theta$ due to Assumption 2.2 on the coercivity condition of the sesquilinear form $a(\cdot, \cdot; \kappa)$. Moreover, we observe that $\mathcal{U} \in \mathcal{A}_r^R$ by definition so that Theorem 2.1 also holds for $z \in \mathcal{U}$.

Lemma 3.1 *Let (u, p) and $(\tilde{u}, \tilde{p})$ denote the solutions of the parametric saddle point problem (1) at $\kappa \in \mathcal{A}_r^R$ and $\tilde{\kappa} \in \mathcal{A}_r^R$, respectively, then we have*

$$\|u - \tilde{u}\|_{\mathcal{V}} \leq \frac{1}{\alpha} C_1 C_u \|\kappa - \tilde{\kappa}\|_{\mathcal{K}} \quad \text{and} \quad \|p - \tilde{p}\|_{\mathcal{Q}} \leq \frac{\alpha + \gamma}{\alpha + \beta} C_1 C_u \|\kappa - \tilde{\kappa}\|_{\mathcal{K}}, \tag{17}$$

where the constants α, β and γ are given in Theorem 2.1, C_1 and C_u are given in (10) and (5).

Proof By subtracting (1) at κ from it at $\tilde{\kappa}$, we have

$$\begin{cases} a(u - \tilde{u}, v; \kappa) + b(v, p - \tilde{p}) = -a(\tilde{u}, v; \kappa - \tilde{\kappa}) \quad \forall v \in \mathcal{V}, \\ b(u - \tilde{u}, q) = 0 \quad \forall q \in \mathcal{Q}. \end{cases}$$

By Theorem 2.1, the following a-priori estimates hold

$$\|u - \tilde{u}\|_{\mathcal{V}} \leq \frac{1}{\alpha} \|\mathrm{a}\|_{\mathcal{V}'} \quad \text{and} \quad \|p - \tilde{p}\|_{\mathcal{Q}} \leq \frac{\alpha + \gamma}{\alpha + \beta} \|\mathrm{a}\|_{\mathcal{V}'}, \tag{18}$$

where we denote $\mathrm{a}(v) = -a(\tilde{u}; v; \kappa - \tilde{\kappa})$ $\forall v \in \mathcal{V}$. By the affine dependence of $a(\cdot, \cdot; \kappa)$ on κ as in (9) and the bound (10) and (4), we have

$$\|\mathrm{a}\|_{\mathcal{V}'} \leq C_1 \|\tilde{u}\|_{\mathcal{V}} \|\kappa - \tilde{\kappa}\|_{\mathcal{K}} \leq C_1 C_u \|\kappa - \tilde{\kappa}\|_{\mathcal{K}}.$$

Thus, we conclude by inserting this bound in (18). □

Lemma 3.2 *For every $z \in \mathcal{A}_r^R$, the complex derivative $(\partial_{z_j} u(z), \partial_{z_j} p(z))$ with respect to z_j for each $j \geq 1$ is well-defined for the solution $(u(z), p(z))$ of the parametric saddle point problem (1), which is given by: find $(\partial_{z_j} u(z), \partial_{z_j} p(z)) \in \mathcal{V} \times \mathcal{Q}$ such that*

$$\begin{cases} a(\partial_{z_j} u, v; \kappa) + b(v, \partial_{z_j} p) = -a(u, v; \kappa_j) \quad \forall v \in \mathcal{V}, \\ \qquad\qquad b(\partial_{z_j} u, q) = 0 \quad \forall q \in \mathcal{Q}. \end{cases}$$

Note that we use $a(u, v; \kappa_j) = \int_D \kappa_j (\nabla \times u) \cdot (\nabla \times v) dx$ by slight abuse of notation for the time harmonic Maxwell system, which is bounded.

Proof For any $z \in \mathcal{A}_r^R$ and $j \geq 1$, for $h \in \mathbb{C} \setminus \{0\}$ sufficiently small such that $|h| \|\kappa_j\|_{\mathcal{K}} \leq \epsilon < r$, we have

$$r - \epsilon \leq \Re(\kappa(x, z + h\boldsymbol{e}_j)) \leq |\kappa(x, z + h\boldsymbol{e}_j)| \leq R + \epsilon \quad \forall x \in D,$$

where $\boldsymbol{e}_j$ is the Kronecker sequence with 1 at index j and 0 at other indices, so that $(u(z + h\boldsymbol{e}_j), p(z + h\boldsymbol{e}_j)) \in \mathcal{V} \times \mathcal{Q}$ is a well-defined solution of (1) at $\kappa(z + h\boldsymbol{e}_j)$. Therefore, we have that the following difference quotients satisfy

$$u_h(z) := \frac{u(z + h\boldsymbol{e}_j) - u(z)}{h} \in \mathcal{V} \quad \text{and} \quad p_h(z) := \frac{p(z + h\boldsymbol{e}_j) - p(z)}{h} \in \mathcal{Q}.$$

Subtracting problem (1) at $\kappa(z + h\boldsymbol{e}_j)$ from its evaluation at $\kappa(z)$ and dividing by h, we obtain that $(u_h(z), p_h(z))$ is a unique solution of the following problem:

$$\begin{cases} a(u_h(z), v; \kappa(z)) + b(v, p_h(z)) = -a(u(z + h\boldsymbol{e}_j), v; \kappa_j) \quad \forall v \in \mathcal{V}, \\ \qquad\qquad b(u_h(z), q) = 0 \quad \forall q \in \mathcal{Q}. \end{cases} \tag{19}$$

Let $\mathrm{a}_h(v) = -a(u(z + h\boldsymbol{e}_j), v; \kappa_j)$. By Assumption 2.1, we have

$$|\mathrm{a}_h(v) - \mathrm{a}_0(v)| \leq \gamma \|u(z + h\boldsymbol{e}_j) - u(z)\|_{\mathcal{V}} \|v\|_{\mathcal{V}}.$$

By the stability estimates (17) in Lemma 3.1, we have

$$\|u(z + h\boldsymbol{e}_j) - u(z)\|_{\mathcal{V}} \leq \frac{1}{\alpha} C_1 C_u \|\kappa_j\|_{\mathcal{K}} |h|,$$

which converges to zero as $|h| \to 0$, so that $\mathrm{a}_h \to \mathrm{a}_0$ in $\mathcal{V}'$ as $|h| \to 0$. Consequently, (u_h, p_h) converges to (u_0, p_0) in $\mathcal{V} \times \mathcal{Q}$ by Theorem 2.1, which is the unique solution of (19) for $h = 0$. Therefore, $(\partial_{z_j} u, \partial_{z_j} p) = (u_0, p_0)$ by the uniqueness. □

To study the convergence rate of the Taylor approximation, we need to bound the Taylor coefficients under Assumption 2.3, for which we employ the Cauchy integral formula in a suitable complex domain. We call a sequence $\boldsymbol{\rho} = (\rho_j)_{j \geq 1}$ is r-admissible

$$\text{if } \sum_{j \geq 1} \rho_j |\kappa_j(x)| \leq \kappa_0(x) - r \text{ and } \rho_j > 1 \text{ for every } j \geq 1. \tag{20}$$

By this definition, if $\boldsymbol{\rho}$ is r-admissible, Theorem 2.1 holds in a larger polydisc

$$\mathcal{U}_{\boldsymbol{\rho}} := \left\{ z \in \mathbb{C}^{\mathbb{N}} : |z_j| \leq \rho_j \text{ for every } j \geq 1 \right\}.$$

This is because $\mathcal{U}_{\boldsymbol{\rho}} \subset \mathcal{A}_r^R$, as it can be readily shown that

$$|\kappa(x, z)| \geq \kappa_0(x) - \sum_{j \geq 1} \rho_j |\kappa_j(x)| \geq r$$

and

$$|\kappa(x, \boldsymbol{z})| \leq \kappa_0(x) + \sum_{j\geq 1} \rho_j |\kappa_j(x)| \leq 2\kappa_0(x) - r < R.$$

Lemma 3.3 *Under Assumptions 2.1 and 2.2, for a sequence $\boldsymbol{\rho}$ satisfying (20), for the Taylor coefficients $t^u_{\boldsymbol{\nu}}$ and $t^p_{\boldsymbol{\nu}}$ defined in (16) we have the following bounds*

$$\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}} \leq C_u \boldsymbol{\rho}^{-\boldsymbol{\nu}} \quad \text{and} \quad \|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}} \leq C_p \boldsymbol{\rho}^{-\boldsymbol{\nu}} \qquad \forall \boldsymbol{\nu} \in \mathcal{F}, \tag{21}$$

where C_u and C_p are given in (5), $\rho^{-0} = 1$ by convention for any $\rho > 0$.

We follow the proof in [25, Lemma 2.4] for elliptic problems and adjust it for the saddle point problem (1) here.

Proof For any $\boldsymbol{\nu} \in \mathcal{F}$, let $J = \max\{j \in \mathbb{N} : \nu_j \neq 0\}$. For such J, let $\boldsymbol{z}^0_J$ denote a truncated complex sequence for any $\boldsymbol{z} \in \mathcal{U}$ defined as

$$(\boldsymbol{z}^0_J)_j = z_j \ \text{ for } 1 \leq j \leq J \quad \text{and} \quad (\boldsymbol{z}^0_J)_j = 0 \ \text{ for } j > J. \tag{22}$$

Then for the solution (u, p) of (1) at $\boldsymbol{z}^0_J$, we have the a-priori estimates (4) by Theorem 2.1 under Assumptions 2.1 and 2.2. Given the sequence $\boldsymbol{\rho}$, we define a new sequence $\tilde{\boldsymbol{\rho}}$ as

$$\tilde{\rho}_j = \rho_j + \varepsilon \ \text{ if } j \geq J \quad \text{and} \quad \tilde{\rho}_j = \rho_j \ \text{ if } j > J, \varepsilon := \frac{r - \theta}{2\|\sum_{1\leq j\leq J} |\kappa_j|\|_{\mathcal{K}}},$$

which implies $\mathcal{U}_{\tilde{\boldsymbol{\rho}}} \subset \mathcal{A}^R_{\tilde{r}}$ with $\tilde{r} = (r+\theta)/2 > \theta$. As the coercivity condition (3) is satisfied for any $\boldsymbol{z} \in \mathcal{A}^R_{\tilde{r}}$ under Assumption 2.2, Theorem 2.1 and Lemma 3.2 hold. Therefore, $u(\boldsymbol{z}^0_J)$ is analytic with respect to each z_j, $1 \leq j \leq J$ on the polydisc $\mathcal{U}_{\tilde{\boldsymbol{\rho}},J}$, which is an open neighborhood of $\mathcal{U}_{\boldsymbol{\rho},J}$ defined as

$$\mathcal{U}_{\boldsymbol{\rho},J} = \left\{(z_1, \ldots, z_J) \in \mathbb{C}^J : |z_j| \leq \rho_j \text{ for every } 1 \leq j \leq J\right\}.$$

Therefore, by the Cauchy integral formula [33, Theorem 2.1.2], we have for u

$$u(\tilde{\boldsymbol{z}}^0_J) = (2\pi i)^{-J} \int_{|z_1|=\rho_1} \cdots \int_{|z_J|=\rho_J} \frac{u(\boldsymbol{z}^0_J)}{(\tilde{z}_1 - z_1)\cdots(\tilde{z}_J - z_J)} dz_1 \cdots dz_J.$$

By taking the derivative $\partial^{\boldsymbol{\nu}}$ on both sides and evaluating it at $\mathbf{0}$, we have

$$\partial^{\boldsymbol{\nu}} u(\mathbf{0}) = \boldsymbol{\nu}!(2\pi i)^{-J} \int_{|z_1|=\rho_1} \cdots \int_{|z_J|=\rho_J} \frac{u(\boldsymbol{z}^0_J)}{z_1^{\nu_1}\cdots z_J^{\nu_J}} dz_1 \cdots dz_J,$$

so that

$$\frac{1}{\boldsymbol{\nu}!}\|\partial^{\boldsymbol{\nu}} u(\mathbf{0})\|_{\mathcal{V}} \leq \sup_{\boldsymbol{z}^0_J \in \mathcal{U}_{\boldsymbol{\rho}}} \|u(\boldsymbol{z}^0_J)\|_{\mathcal{V}} \prod_{1\leq j\leq J} \rho_j^{-\nu_j} \leq C_u \boldsymbol{\rho}^{-\boldsymbol{\nu}},$$

which is (21) for u. The same argument is applied to derive the bound for p. □

Lemma 3.4 *Under Assumption 2.3, there exists a $\frac{r+\theta}{2}$-admissible sequence $\boldsymbol{\rho}$, i.e, it satisfies (20) with r replaced by $\frac{r+\theta}{2}$, such that*

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t^u_{\boldsymbol{\nu}}\|^s_{\mathcal{V}} < \infty \quad \text{and} \quad \sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t^p_{\boldsymbol{\nu}}\|^s_{\mathcal{Q}} < \infty.$$

This result for the saddle point problems here can be proved following that in [25, Sec. 3] for elliptic problems.

Proof By Lemma 3.3, we only need to prove there exists a $\frac{r+\theta}{2}$-admissible sequence $\boldsymbol{\rho}$ such that

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} \boldsymbol{\rho}^{-s\boldsymbol{\nu}} < \infty. \tag{23}$$

This is done in a constructive way by specification of $\boldsymbol{\rho}$. By Assumption 2.3, we have $(\|\kappa_j\|_{\mathcal{K}})_{j\geq 1} \in \ell^s(\mathbb{N}) \subset \ell^1(\mathbb{N})$, so that there exists a sufficiently large J such that

$$\sum_{j>J} \|\kappa_j\|_{\mathcal{K}} \leq \frac{r-\theta}{12}.$$

Then we choose $\tau > 1$ such that

$$(\tau - 1)\sum_{j\leq J} \|\kappa_j\|_{\mathcal{K}} \leq \frac{r-\theta}{4}.$$

For any $\boldsymbol{\nu} \in \mathcal{F}$, we specify the sequence $\boldsymbol{\rho}$ as

$$\rho_j := \tau,\ j \leq J; \quad \rho_j := \max\left\{1, \frac{(r-\theta)\nu_j}{4\|\kappa_j\|_{\mathcal{K}} \sum_{i>J} \nu_i}\right\},\ j > J, \tag{24}$$

with the convention that $\nu_j/(\sum_{i>J} \nu_i) = 0$ if $\sum_{i>J} \nu_i = 0$. Then we have

$$\sum_{j\geq 1} \rho_j |\kappa_j(x)| \leq \sum_{j\geq 1} |\kappa_j(x)| + \frac{r-\theta}{2} \leq \kappa_0(x) - \frac{r+\theta}{2},$$

where in the second inequality we have used Assumption 2.2, i.e., for any $x \in D$,

$$r < \kappa_0(x) + \inf_{\boldsymbol{y}\in U} \sum_{j\geq 1} y_j \kappa_j(x) = \kappa_0(x) - \sum_{j\geq 1} |\kappa_j(x)|.$$

Therefore, $\boldsymbol{\rho}$ is $\frac{r+\theta}{2}$-admissible. By results in [25, Sec. 3], (23) holds for the choice (24). □

3.2 ℓ^s-Summability by Weighted ℓ^2-Summability

The ℓ^s-summability of the Taylor coefficients is guaranteed by the ℓ^s-summability of $(\|\kappa_j\|_{\mathcal{K}})_{j\geq 1}$ in Assumption 2.3 as shown in the last section. However, as indicated in Remark 2.3, $(\|\kappa_j\|_{\mathcal{K}})_{j\geq 1}$ may not be ℓ^s-summable for any $s \in (0, 1)$, as considered in [4] for coercive elliptic PDEs. In this case, Assumption 2.4 may still hold, in particular for locally supported $(\kappa_j)_{j\geq 1}$, for which we prove the ℓ^s-summability of the Taylor coefficients $(\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and the ℓ^t-summability of the Taylor coefficients $(\|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}$, where $s = \frac{2t}{2+t}$ for $t \in (0, \infty)$ given in Assumption 2.4.

Lemma 3.5 *Under Assumption 2.4, we have*

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t^u_{\boldsymbol{\nu}}\|^s_{\mathcal{V}} < \infty \quad \textit{and} \quad \sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t^p_{\boldsymbol{\nu}}\|^t_{\mathcal{Q}} < \infty,$$

where $s = \frac{2t}{2+t} \in (0, 2)$ *for* $t \in (0, \infty)$ *given in Assumption 2.4.*

The different summability results for u and p can be proved by following that in [4] for elliptic problem with necessary adjustment to the saddle point problems here.

Proof For a sequence $\boldsymbol{\rho}$ satisfying (15) in Assumption 2.4, we define the scaling function $R_{\boldsymbol{\rho}}(\boldsymbol{y}) := (\rho_j y_j)_{j\geq 1}$. By Assumption (15) we have for any $x \in D$

$$\inf_{\boldsymbol{y}\in U} \kappa(x, R_{\boldsymbol{\rho}}(\boldsymbol{y})) = \kappa_0(x) + \inf_{\boldsymbol{y}\in U} \sum_{j\geq 1} \rho_j y_j \kappa_j(x) \geq \kappa_0(x) - \sum_{j\geq 1} \rho_j |\kappa_j(x)| \geq \epsilon > \theta,$$

so that $a(\cdot, \cdot; \kappa)$ is coercive by Assumption 2.2. Under Assumption 2.1, there exists a unique $(u(R_{\boldsymbol{\rho}}(\boldsymbol{y})), p(R_{\boldsymbol{\rho}}(\boldsymbol{y}))) \in \mathcal{V} \times \mathcal{Q}$ for every $\boldsymbol{y} \in U$ such that

$$\begin{cases} a(u(R_{\boldsymbol{\rho}}(\boldsymbol{y})), v; \kappa(R_{\boldsymbol{\rho}}(\boldsymbol{y}))) + b(v, p(R_{\boldsymbol{\rho}}(\boldsymbol{y}))) = f(v) & \forall v \in \mathcal{V}, \\ b(u(R_{\boldsymbol{\rho}}(\boldsymbol{y})), q) = g(q) & \forall q \in \mathcal{Q}. \end{cases} \tag{25}$$

By the definition of the Taylor coefficients in (16), we have at $\boldsymbol{\nu} = \boldsymbol{0}$ that $(t^u_{\boldsymbol{0}}, t^p_{\boldsymbol{0}}) = (u(\boldsymbol{0}), p(\boldsymbol{0}))$, which satisfy the a-priori estimates (4) by Theorem 2.1, i.e.,

$$\|t^u_{\boldsymbol{0}}\|_{\mathcal{V}} \leq C_u \quad \text{and} \quad \|t^p_{\boldsymbol{0}}\|_{\mathcal{Q}} \leq C_p.$$

For any other $\boldsymbol{\nu} \in \mathcal{F}$, by taking the partial derivative $\partial^{\boldsymbol{\nu}}$ for (25), we obtain

$$\begin{cases} a(\boldsymbol{\rho}^{\boldsymbol{\nu}} \partial^{\boldsymbol{\nu}} u(R_{\boldsymbol{\rho}}(\boldsymbol{y})), v; \kappa(R_{\boldsymbol{\rho}}(\boldsymbol{y}))) + b(v, \boldsymbol{\rho}^{\boldsymbol{\nu}} \partial^{\boldsymbol{\nu}} p(R_{\boldsymbol{\rho}}(\boldsymbol{y}))) \\ = - \displaystyle\sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\nu_j \boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} \partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} u(R_{\boldsymbol{\rho}}(\boldsymbol{y})), v; \rho_j \kappa_j) \quad \forall v \in \mathcal{V}, \\ b(\boldsymbol{\rho}^{\boldsymbol{\nu}} \partial^{\boldsymbol{\nu}} u(R_{\boldsymbol{\rho}}(\boldsymbol{y})), q) = 0 \quad \forall q \in \mathcal{Q}, \end{cases}$$

where $\text{supp}\,\boldsymbol{\nu} = \{j \in \mathbb{N} : \nu_j \neq 0\}$. Taking division by $\boldsymbol{\nu}!$ on both sides, setting $\boldsymbol{y} = \boldsymbol{0}$, we have the saddle point problem for the Taylor coefficients $(t^u_{\boldsymbol{\nu}}, t^p_{\boldsymbol{\nu}}) \in \mathcal{V} \times \mathcal{Q}$

$$\begin{cases} a(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, v; \kappa(R_{\boldsymbol{\rho}}(\boldsymbol{y}))) + b(v, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^p_{\boldsymbol{\nu}}) \\ = - \sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, v; \rho_j \kappa_j) \quad \forall v \in \mathcal{V}, \\ b(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, q) = 0 \quad \forall q \in \mathcal{Q}. \end{cases} \tag{26}$$

Therefore, $t^u_{\boldsymbol{\nu}} \in \mathcal{V}^0$ by the second equation. We shall show that $(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^p_{\boldsymbol{\nu}}) \in \mathcal{V} \times \mathcal{Q}$ is a bounded solution of (26) for any $\boldsymbol{\nu} \in \mathcal{F}$. First it is so for $\boldsymbol{\nu} = \boldsymbol{0}$. Then by induction we assume that $(\boldsymbol{\rho}^{\boldsymbol{\mu}} t^u_{\boldsymbol{\mu}}, \boldsymbol{\rho}^{\boldsymbol{\mu}} t^p_{\boldsymbol{\mu}}) \in \mathcal{V} \times \mathcal{Q}$ are bounded solutions of (26) (being $\boldsymbol{\nu}$ replaced by $\boldsymbol{\mu}$) for any $\boldsymbol{\mu} \preceq \boldsymbol{\nu}$, i.e., $\mu_j \leq \nu_j \ \forall j \geq 1$, and $\boldsymbol{\mu} \neq \boldsymbol{\nu}$, then by Theorem 2.1 we have $(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^p_{\boldsymbol{\nu}}) \in \mathcal{V} \times \mathcal{Q}$ is the unique solution of (26), such that

$$\begin{aligned} \boldsymbol{\rho}^{\boldsymbol{\nu}} \|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}} &\leq \frac{1}{\alpha} \sup_{\|v\|_{\mathcal{V}}=1} \sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, v; \rho_j \kappa_j), \\ \boldsymbol{\rho}^{\boldsymbol{\nu}} \|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}} &\leq \frac{\alpha + \gamma}{\alpha\beta} \sup_{\|v\|_{\mathcal{V}}=1} \sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, v; \rho_j \kappa_j), \end{aligned} \tag{27}$$

where by (10) and $|\boldsymbol{\nu}|_0 = \#\{j \in \mathbb{N} : \nu_j > 0\} < \infty$ for any $\boldsymbol{\nu} \in \mathcal{F}$ we have

$$\begin{aligned} &\sup_{\|v\|_{\mathcal{V}}=1} \sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, v; \rho_j \kappa_j) \\ &\leq C_1 |\boldsymbol{\nu}|_0 (\|\kappa_0\|_{\mathcal{K}} - \epsilon) \max_{j\geq 1}(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} \|t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}\|_{\mathcal{V}}) < \infty. \end{aligned} \tag{28}$$

Therefore, by taking the test functions as $(v, q) = (\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^p_{\boldsymbol{\nu}})$, we obtain

$$\begin{aligned} a(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \kappa_0) &= - \sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \rho_j \kappa_j) \\ &\leq \frac{1}{2} \sum_{j\in \text{supp}\,\boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, \boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}; \rho_j |\kappa_j|) + a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \rho_j |\kappa_j|), \end{aligned} \tag{29}$$

where for the inequality we used the assumption (10). Therefore, by (15), we have

$$\sum_{j\in \text{supp}\, \boldsymbol{\nu}} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \rho_j|\kappa_j|) \leq a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \kappa_0 - \epsilon),$$

which, together with (29) leads to

$$a\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2}\right) \leq \frac{1}{2} \sum_{j\in \text{supp}\, \boldsymbol{\nu}} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, \boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}; \rho_j|\kappa_j|\right). \tag{30}$$

By Assumption 2.2, we have

$$a(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \theta) \geq \alpha \|\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}\|^2_{\mathcal{V}} \geq 0,$$

so that by the affine structure (9) there holds

$$\begin{aligned} a\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2}\right) &= a(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \theta) + a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right) \\ &\geq a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right). \end{aligned} \tag{31}$$

Hence, from (30) and (31) we obtain

$$a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right) \leq \frac{1}{2} \sum_{j\in \text{supp}\, \boldsymbol{\nu}} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, \boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}; \rho_j|\kappa_j|\right).$$

Summing over $|\boldsymbol{\nu}| = k$ for any $k \geq 1$ for both sides, we have

$$\begin{aligned} &\sum_{|\boldsymbol{\nu}|=k} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right) \\ &= \frac{1}{2} \sum_{|\boldsymbol{\nu}|=k} \sum_{j\in \text{supp}\, \boldsymbol{\nu}} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}, \boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{e}_j} t^u_{\boldsymbol{\nu}-\boldsymbol{e}_j}; \rho_j|\kappa_j|\right) \\ &= \frac{1}{2} \sum_{|\boldsymbol{\nu}|=k-1} \sum_{j\geq 1} a_1(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \rho_j|\kappa_j|) \\ &\leq \sum_{|\boldsymbol{\nu}|=k-1} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0-\epsilon}{2}\right) \\ &\leq \sup_{x\in D} \frac{\kappa_0(x)-\epsilon}{\kappa_0(x)+\epsilon-2\theta} \sum_{|\boldsymbol{\nu}|=k-1} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right), \end{aligned}$$

where we used Assumption 2.4 in the first inequality. By denoting

$$\sigma = \sup_{x\in D} \frac{\kappa_0(x)-\epsilon}{\kappa_0(x)+\epsilon-2\theta} < 1, \qquad \text{since } \theta < \epsilon,$$

we obtain

$$\sum_{|\boldsymbol{\nu}|=k} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right) \leq \sigma^k a_1\left(t^u_{\boldsymbol{0}}, t^u_{\boldsymbol{0}}; \frac{\kappa_0+\epsilon}{2} - \theta\right).$$

Summing over $k \geq 1$, we have

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0+\epsilon}{2} - \theta\right) \leq \frac{1}{1-\sigma} a_1\left(t^u_{\boldsymbol{0}}, t^u_{\boldsymbol{0}}; \frac{\kappa_0+\epsilon}{2} - \theta\right) < \infty.$$

By the coercivity condition (10) in $\mathcal{V}^0$, for any $\boldsymbol{\nu} \neq \mathbf{0}$, as $t^u_{\boldsymbol{\nu}} \in \mathcal{V}^0$ we have

$$a_1\left(\boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}, \boldsymbol{\rho}^{\boldsymbol{\nu}} t^u_{\boldsymbol{\nu}}; \frac{\kappa_0 + \epsilon}{2} - \theta\right) \geq c_1 \inf_{x \in D}\left(\frac{\kappa_0(x) + \epsilon}{2} - \theta\right)(\boldsymbol{\rho}^{\boldsymbol{\nu}} \|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})^2, \tag{32}$$

where $\inf_{x \in D} \kappa_0(x) > \epsilon > \theta$ by Assumption 2.4. Therefore, we obtain

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} (\boldsymbol{\rho}^{\boldsymbol{\nu}} \|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})^2 < \infty. \tag{33}$$

By Hölder's inequality, we have

$$\begin{aligned}\sum_{\boldsymbol{\nu} \in \mathcal{F}} \|t^u_{\boldsymbol{\nu}}\|^s_{\mathcal{V}} &= \sum_{\boldsymbol{\nu} \in \mathcal{F}} (\boldsymbol{\rho}^{\boldsymbol{\nu}} \|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})^s \boldsymbol{\rho}^{-s\boldsymbol{\nu}} \\ &\leq \left(\sum_{\boldsymbol{\nu} \in \mathcal{F}} (\boldsymbol{\rho}^{\boldsymbol{\nu}} \|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})^2\right)^{s/2} \left(\sum_{\boldsymbol{\nu} \in \mathcal{F}} \boldsymbol{\rho}^{-\frac{2s}{2-s}\boldsymbol{\nu}}\right)^{(2-s)/2},\end{aligned}$$

where the first term is finite by (33). For the second term, with $t = \frac{2s}{2-s}$, i.e., $s = \frac{2t}{2+t}$, we have

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} \boldsymbol{\rho}^{-\frac{2s}{2-s}\boldsymbol{\nu}} = \prod_{j \geq 1}\left(\sum_{k=0}^{\infty} \rho_j^{-tk}\right) = \prod_{j \geq 1}(1 - \rho_j^{-t})^{-1}.$$

As $(\rho_j^{-1})_{j \geq 1} \in \ell^t(\mathbb{N})$, there exists $J \in \mathbb{N}$ such that $\rho_j^{-t} < \frac{1}{2}$ for all $j > J$. Note that $g(x) := -\log(1-x) - 2x < 0$ as $g(0) = 0$ and $g'(x) = \frac{1}{1-x} - 2 < 0$ for $0 < x < \frac{1}{2}$, which implies $(1 - \rho_j^{-t})^{-1} < \exp(2\rho_j^{-t})$ for $j > J$, so that

$$\prod_{j \geq 1}(1 - \rho_j^{-t})^{-1} < \exp\left(2\sum_{j > J} \rho_j^{-t}\right) \prod_{j \leq J}(1 - \rho_j^{-t})^{-1},$$

which is finite as $(\rho_j^{-1})_{j \geq 1} \in \ell^t(\mathbb{N})$. Therefore, $(\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})_{\boldsymbol{\nu} \in \mathcal{F}} \in \ell^s(\mathcal{F})$.

By (33), there exists a constant $C_2 > 0$ such that

$$\sup_{\boldsymbol{\nu} \in \mathcal{F}} \|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}} \leq C_2 \boldsymbol{\rho}^{-\boldsymbol{\nu}}. \tag{34}$$

Therefore, by (27) and (28), we have

$$\|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}} \leq C_3 \boldsymbol{\rho}^{-\boldsymbol{\nu}} |\boldsymbol{\nu}|_0 \leq C_3 \boldsymbol{\rho}^{-\boldsymbol{\nu}} \prod_{j \geq 1}(1 + \nu_j), \tag{35}$$

where

$$C_3 = C_1 C_2 \frac{\alpha + \gamma}{\alpha\beta}(\|\kappa_0\|_{\mathcal{K}} - \epsilon) < \infty,$$

and we used the fact $|\boldsymbol{\nu}|_0 \leq \prod_{j \geq 1}(1 + \nu_j)$ for any $\boldsymbol{\nu} \in \mathcal{F}$ in the second inequality. Hence, we have

$$\sum_{\boldsymbol{\nu} \in \mathcal{F}} \|t^p_{\boldsymbol{\nu}}\|^t_{\mathcal{Q}} \leq (C_3)^t \sum_{\boldsymbol{\nu} \in \mathcal{F}} \prod_{j \geq 1} \rho_j^{-t\nu_j}(1 + \nu_j)^t = (C_3)^t \prod_{j \geq 1} \sum_{k=0}^{\infty} \rho_j^{-tk}(1 + k)^t, \tag{36}$$

where for each $j \geq 0$ we have

$$\sum_{k=0}^{\infty} \rho_j^{-tk}(1 + k)^t = 1 + \rho_j^{-t} \sum_{k=0} \rho_j^{-tk}(2 + k)^t.$$

As $(\rho_j^{-1})_{j\geq 1} \in \ell^t(\mathbb{N})$, there exists $J > 0$ such that $\rho_j^{-1} < \frac{1}{4}$ for any $j > J$. Moreover, for any $t > 0$, there exist $c_1 > 0$ and $1 < c_2 < 2$ such that $(2+k)^t \leq c_1 c_2^k$ for $k \geq 0$, so that

$$\sum_{k=0}^{\infty} \rho_j^{-tk}(2+k)^t \leq c_1 \sum_{k=0}^{\infty} (\rho_j^{-1} c_2)^k = c_1(1-\rho_j^{-1}c_2)^{-1} \leq 2c_1.$$

As $\rho_j > 1$, there exists $C_j < \infty$ for each $j \geq 1$ such that

$$\sum_{k=0}^{\infty} \rho_j^{-tk}(1+k)^t \leq C_j.$$

Therefore, we have

$$\prod_{j\geq 1}\sum_{k=0}^{\infty} \rho_j^{-tk}(1+k)^t \leq \prod_{j\leq J} C_j \prod_{j>J}(1+2c_1\rho_j^{-t}) \leq \exp\left(2c_1 \sum_{j>J} \rho_j^{-t}\right) \prod_{j\leq J} C_j, \quad (37)$$

which is finite when $(\rho_j^{-1})_{j\geq 1} \in \ell^t(\mathbb{N})$. Note that in the second inequality, we used $1+x \leq e^x$ for $x \geq 0$. Hence $(\|t_{\boldsymbol{\nu}}^p\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^t(\mathcal{F})$ from (36). □

Remark 3.1 We remark that the weighted ℓ^2-summability for $(\|t_{\boldsymbol{\nu}}^u\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}$ in Lemma 3.5 is a result of the coercivity property (32) (where the ℓ^2-norm shows up) of $a_1(\cdot,\cdot;\kappa) : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$. However, the weighted ℓ^2-summability cannot be shown for $(\|t_{\boldsymbol{\nu}}^p\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}$, where $t_{\boldsymbol{\nu}}^p$ only appears in $b(\cdot,\cdot) : \mathcal{V}\times\mathcal{Q} \to \mathbb{R}$ that holds the inf-sup condition. Instead, by this condition, we can bound the Taylor coefficient $t_{\boldsymbol{\nu}}^p$ as in (35) by (28).

3.3 Dimension-independent Convergence

As a consequence of the summability obtained in the Sections 3.1 and 3.2, we obtain the following convergence results.

Theorem 3.1 *Under Assumptions 2.1 and 2.2, there exist two sequences of index sets $(\Lambda_N^u)_{N\geq 1}$ and $(\Lambda_N^p)_{N\geq 1}$ with indices $\boldsymbol{\nu} \in \mathcal{F}$ corresponding to the N largest Taylor coefficients $\|t_{\boldsymbol{\nu}}^u\|_{\mathcal{V}}$ and $\|t_{\boldsymbol{\nu}}^p\|_{\mathcal{Q}}$, respectively, such that*

$$\begin{aligned} \sup_{\boldsymbol{y}\in U} \|u(\boldsymbol{y}) - T_{\Lambda_N^u} u(\boldsymbol{y})\|_{\mathcal{V}} &\leq \|(\|t_{\boldsymbol{\nu}}^u\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s(\mathcal{F})} N^{-r(s)}, \\ \sup_{\boldsymbol{y}\in U} \|p(\boldsymbol{y}) - T_{\Lambda_N^p} p(\boldsymbol{y})\|_{\mathcal{Q}} &\leq \|(\|t_{\boldsymbol{\nu}}^p\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s(\mathcal{F})} N^{-r(s)}, \end{aligned} \quad (38)$$

under Assumption 2.3, and

$$\begin{aligned} \sup_{\boldsymbol{y}\in U} \|u(\boldsymbol{y}) - T_{\Lambda_N^u} u(\boldsymbol{y})\|_{\mathcal{V}} &\leq \|(\|t_{\boldsymbol{\nu}}^u\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s(\mathcal{F})} N^{-r(s)}, \\ \sup_{\boldsymbol{y}\in U} \|p(\boldsymbol{y}) - T_{\Lambda_N^p} p(\boldsymbol{y})\|_{\mathcal{Q}} &\leq \|(\|t_{\boldsymbol{\nu}}^p\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^t(\mathcal{F})} N^{-r(t)}, \end{aligned} \quad (39)$$

under Assumption 2.4, where the dimension-independent convergence rate r is given by

$$r(s) = \frac{1}{s} - 1, \quad s < 1. \quad (40)$$

The convergence results are due to the application of Stechkin's Lemma [24, Lemma 5.5], as also used in [25] for elliptic problems, which we briefly present below for the saddle point problems.

Proof At first, by Lemmas 3.4 and 3.5 for Assumption 2.3 and Assumption 2.4, respectively, for any $s < 1$, we have

$$\sup_{\boldsymbol{y}\in U}\left\|\sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{y}^{\boldsymbol{\nu}} t_{\boldsymbol{\nu}}^{u}\right\|_{\mathcal{V}} \leq \sup_{\boldsymbol{y}\in U}\sum_{\boldsymbol{\nu}\in\mathcal{F}} |\boldsymbol{y}^{\boldsymbol{\nu}}|\, \|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}} \leq \sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}} \leq \sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}}^{s} < \infty,$$

which implies that the Taylor power series $T_{\mathcal{F}}u$ defined in (16) is uniformly convergent. Secondly, for any $\boldsymbol{y} \in U$ and $\varepsilon > 0$, by Lemma 3.1, there exists $J_1 > 0$ such that for any $J \geq J_1$

$$B_1 := \|u(\boldsymbol{y}) - u(\boldsymbol{y}_J^0)\|_{\mathcal{V}} \leq \frac{1}{\alpha} C_1 C_u \|\kappa(\boldsymbol{y}) - \kappa(\boldsymbol{y}_J^0)\|_{\mathcal{K}} < \frac{\varepsilon}{2},$$

under Assumptions 2.3 or 2.4, where $\boldsymbol{y}_J^0$ is defined in the same way as in (22). Moreover, for any $J \geq J_1$, by the analytic regularity of $u(\boldsymbol{y}_J^0)$ in the complex domain $\mathcal{U}_{\boldsymbol{\rho}}$ as indicated in Lemma 3.2, there exists $K > 0$ such that for any $\Lambda = \{\boldsymbol{\nu} \in \mathcal{F} : \nu_j > K \text{ for } j \leq J \text{ and } \nu_j = 0 \text{ for } j > J\}$ there holds

$$B_2 := \|u(\boldsymbol{y}_J^0) - T_\Lambda u(\boldsymbol{y}_J^0)\|_{\mathcal{V}} < \frac{\varepsilon}{2}.$$

By the definition of Λ we have $T_\Lambda u(\boldsymbol{y}_J^0) = T_\Lambda u(\boldsymbol{y})$. Hence, we have

$$\|u(\boldsymbol{y}) - T_\Lambda u(\boldsymbol{y})\|_{\mathcal{V}} \leq B_1 + B_2 < \varepsilon,$$

which implies that the Taylor power series $T_{\mathcal{F}}u(\boldsymbol{y})$ converges to $u(\boldsymbol{y})$ for every $\boldsymbol{y} \in U$. Consequently,

$$\sup_{\boldsymbol{y}\in U}\|u(\boldsymbol{y}) - T_{\Lambda_N^u} u(\boldsymbol{y})\|_{\mathcal{V}} = \sup_{\boldsymbol{y}\in U}\left\|\sum_{\boldsymbol{\nu}\notin\Lambda_N^u} \boldsymbol{y}^{\boldsymbol{\nu}} t_{\boldsymbol{\nu}}^{u}\right\|_{\mathcal{V}} \leq \sum_{\boldsymbol{\nu}\notin\Lambda_N^u} \|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}},$$

which concludes for the error of the Taylor approximation of u by using Stechkin's Lemma [24, Lemma 5.5], i.e., for a non-increasing arrangement of $(\|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}$, there holds

$$\sum_{\boldsymbol{\nu}\notin\Lambda_N^u} \|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}} \leq \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} \|t_{\boldsymbol{\nu}}^{u}\|_{\mathcal{V}}^{s}\right)^{1/s} N^{-r(s)},$$

with $r(s)$ defined in (40). The same result holds for the error of the Taylor approximation of p by using the same argument. □

Remark 3.2 We remark that the convergence results (38) and (39) are obtained under different assumptions, and cannot be implied by one another. In fact, it is clear that (39) cannot be implied by (38) as explained in Remark 2.2. On the other hand, (38) cannot be implied by (39) as shown in the following simple example: let $\kappa_0 = 1$ and $\kappa_j = j^{-2}$ for $j \geq 1$, then by (38) we have the convergence rate N^{-r} for any $r < 1$ arbitrarily close to 1. However, by (39), for which there exists $(\rho_j^{-1})_{j\geq 1} \in \ell^t(\mathbb{N})$ with $t > 1$ satisfying (15), we can only obtain a convergence rate of N^{-r} for $r = \frac{1}{s} - 1 = \frac{1}{t} - \frac{1}{2} < \frac{1}{2}$ for $\sup_{\boldsymbol{y}\in U}\|u(\boldsymbol{y}) - T_{\Lambda_N^u} u(\boldsymbol{y})\|_{\mathcal{V}}$, and $r = \frac{1}{t} - 1 < 0$, i.e., non-convergent, for $\sup_{\boldsymbol{y}\in U}\|p(\boldsymbol{y}) - T_{\Lambda_N^p} p(\boldsymbol{y})\|_{\mathcal{Q}}$.

Theorem 3.1 states the existence of such index sets $\Lambda_N^u \subset \mathcal{F}$ and $\Lambda_N^p \subset \mathcal{F}$ that lead to the dimension-independent convergence rates. However, there is no particular structure of these index sets. To guide more practical algorithm development, we consider a particular structure of these index sets, namely, downward closed set $\Lambda \subset \mathcal{F}$, also known as admissible set or monotone set [19, 21, 26, 30], which satisfies

$$\text{if } \boldsymbol{\nu} \in \Lambda \text{ then } \boldsymbol{\mu} \in \Lambda \quad \forall \boldsymbol{\mu} \preceq \boldsymbol{\nu},$$

where we recall that $\boldsymbol{\mu} \preceq \boldsymbol{\nu}$ means $\mu_j \leq \nu_j$ for all $j \geq 1$.

We say that a sequence $(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ is monotonically decreasing

$$\text{if } \boldsymbol{\mu} \preceq \boldsymbol{\nu} \text{ then } \theta_{\boldsymbol{\nu}} \leq \theta_{\boldsymbol{\mu}}.$$

Lemma 3.6 *Let $(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ be a monotonically decreasing sequence of positive real numbers in $\ell^s(\mathcal{F})$ with $s < 1$, then there exists a sequence of downward closed and nested index sets $(\Lambda_N)_{N\geq 1} \subset \mathcal{F}$ such that*

$$\sum_{\boldsymbol{\nu}\notin\Lambda_N} \theta_{\boldsymbol{\nu}} \leq \|(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s(\mathcal{F})} N^{-r(s)}, \quad r(s) = \frac{1}{s} - 1. \tag{41}$$

Proof By Stechkin's Lemma as in the proof of Theorem 3.1, there exists a sequence of index sets $(\Lambda_N)_{N\geq 1} \subset \mathcal{F}$ such that (41) holds. It is left to show that $(\Lambda_N)_{N\geq 1}$ can be taken as downward closed and nested. This is achieved by an induction argument. First, for $N = 1$, we take $\Lambda_1 = \{\boldsymbol{\nu}(1)\}$ with $\boldsymbol{\nu}(1) = \mathbf{0}$, then (41) holds. Suppose (41) holds for some $N > 1$ with downward closed and nested index set Λ_N, then we look for the next index $\boldsymbol{\nu}(N+1) \in \mathcal{F}$ such that $\Lambda_{N+1} := \Lambda_N \cup \{\boldsymbol{\nu}(N+1)\}$ is downward closed and (41) holds in Λ_{N+1}. Let $\mathcal{N}(\Lambda_N)$ denote the admissible forward neighbor set defined as

$$\mathcal{N}(\Lambda_N) = \{\boldsymbol{\nu} \in \mathcal{F} \setminus \Lambda_N : \boldsymbol{\nu} - \boldsymbol{e}_j \in \Lambda_N \text{ for every } j \in \mathbb{N} \text{ such that } \nu_j \neq 0\},$$

where we recall the Kronecker sequence $\boldsymbol{e}_j = (\delta_{ij})_{i\geq 1}$. Then we take

$$\boldsymbol{\nu}(N+1) = \underset{\boldsymbol{\mu}\in\mathcal{N}(\Lambda_N)}{\operatorname{argmax}}\ \theta_{\boldsymbol{\mu}}.$$

By the definition of the admissible forward neighbor set $\mathcal{N}(\Lambda_N)$, we have $\Lambda_{N+1} := \Lambda_N \cup \{\boldsymbol{\nu}\}$ is downward closed for any $\boldsymbol{\nu} \in \mathcal{N}(\Lambda_N)$. Moreover, the sequence $(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ is monotonically decreasing, which implies $\theta_{\boldsymbol{\nu}(N+1)} \leq \theta_{\boldsymbol{\nu}(N)}$ since $\boldsymbol{\nu}(N) \preceq \boldsymbol{\nu}(N+1)$ for every $N \geq 1$, which satisfies the Stechkin's Lemma for decreasing sequence to hold (41) in Λ_{N+1}. This concludes. □

Let $(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ be a real sequence. Then the sequence $(\theta^*_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ with

$$\theta^*_{\boldsymbol{\nu}} := \max_{\boldsymbol{\nu}\preceq\boldsymbol{\mu}} \theta_{\boldsymbol{\mu}} \quad \forall \boldsymbol{\nu} \in \mathcal{F}, \tag{42}$$

is monotonically decreasing. If the sequence $(\theta^*_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ is $\ell^s(\mathcal{F})$-summable, then we denote a $\ell^s_m(\mathcal{F})$-norm for $(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ as

$$\|(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s_m(\mathcal{F})} = \|(\theta^*_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s(\mathcal{F})}.$$

We provide the dimension-independent convergence rates for the case of downward closed and nested index sets for saddle point problems, following that in [26] for elliptic problems.

Theorem 3.2 *Under Assumptions 2.1 and 2.2, there exist two sequences of downward closed and nested index sets $(\Lambda^u_N)_{N\geq 1}$ and $(\Lambda^p_N)_{N\geq 1}$ with indices $\boldsymbol{\nu} \in \mathcal{F}$ corresponding to* *the N largest Taylor coefficients $\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}}$ and $\|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}}$, respectively, such that*

$$\sup_{\boldsymbol{y}\in U}\|u(\boldsymbol{y})-T_{\Lambda^u_N}u(\boldsymbol{y})\|_{\mathcal{V}} \le \|(\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s_m(\mathcal{F})}N^{-r(s)},$$

$$\sup_{\boldsymbol{y}\in U}\|p(\boldsymbol{y})-T_{\Lambda^p_N}p(\boldsymbol{y})\|_{\mathcal{Q}} \le \|(\|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^s_m(\mathcal{F})}N^{-r(s)},$$

under Assumption 2.3, and

$$\sup_{\boldsymbol{y}\in U}\|u(\boldsymbol{y})-T_{\Lambda^u_N}u(\boldsymbol{y})\|_{\mathcal{V}} \le \|(\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^t_m(\mathcal{F})}N^{-r(t)},$$

$$\sup_{\boldsymbol{y}\in U}\|p(\boldsymbol{y})-T_{\Lambda^p_N}p(\boldsymbol{y})\|_{\mathcal{Q}} \le \|(\|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^t_m(\mathcal{F})}N^{-r(t)},$$

under Assumption 2.4, where the dimension-independent convergence rate r is given by

$$r(s)=\frac{1}{s}-1,\quad s<1.$$

Proof By Theorem 3.1 and Lemma 3.6, we only need to show that $(\|t^u_{\boldsymbol{\nu}}\|^*_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and $(\|t^p_{\boldsymbol{\nu}}\|^*_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}$, the associated monotone envelopes defined in (42) for $(\|t^u_{\boldsymbol{\nu}}\|_{\mathcal{V}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and $(\|t^p_{\boldsymbol{\nu}}\|_{\mathcal{Q}})_{\boldsymbol{\nu}\in\mathcal{F}}$, respectively, are $\ell^s(\mathcal{F})$-summable under Assumption 2.3, and $\ell^t(\mathcal{F})$-summable under Assumption 2.4. Under Assumption 2.3, by Lemma 3.3 we have

$$|t^u_{\boldsymbol{\nu}}\|^*_{\mathcal{V}} \le C_u\boldsymbol{\rho}^{-\boldsymbol{\nu}} \quad\text{and}\quad \|t^p_{\boldsymbol{\nu}}\|^*_{\mathcal{Q}} \le C_p\boldsymbol{\rho}^{-\boldsymbol{\nu}} \quad \forall\boldsymbol{\nu}\in\mathcal{F},$$

since $(\boldsymbol{\rho}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ is monotonically decreasing by (20). Moreover, as shown in Lemma 3.4, $(\boldsymbol{\rho}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ is $\ell^s(\mathcal{F})$-summable, which concludes. Under Assumption 2.4, we have by (34) and (35) that

$$\|t^u_{\boldsymbol{\nu}}\|^*_{\mathcal{V}} \le C_2\boldsymbol{\rho}^{-\boldsymbol{\nu}} \quad\text{and}\quad \|t^p_{\boldsymbol{\nu}}\|^*_{\mathcal{Q}} \le C_3\theta^*_{\boldsymbol{\nu}} \quad \forall\boldsymbol{\nu}\in\mathcal{F},$$

since both $(\boldsymbol{\rho}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and $(\theta^*_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ are monotonically decreasing, where we denote

$$\theta_{\boldsymbol{\nu}}=\boldsymbol{\rho}^{-\boldsymbol{\nu}}\prod_{j\ge1}(1+\nu_j)\quad \boldsymbol{\nu}\in\mathcal{F}.$$

The $\ell^t(\mathcal{F})$-summability of $(\boldsymbol{\rho}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ can be shown as in (36). For the $\ell^t(\mathcal{F})$-summability of $(\theta^*_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$, we proceed as follows. As $(\rho_j^{-1})_{j\ge1}\in\ell^t(\mathbb{N})$, there exists a J such that $\rho_j^{-1}<1/4$ for all $j>J$, which implies

$$\frac{\theta_{\boldsymbol{\nu}+\boldsymbol{e}_j}}{\theta_{\boldsymbol{\nu}}}=\frac{(1+\nu_j+1)}{(1+\nu_j)\rho_j}<1\quad \forall j>J. \tag{43}$$

Moreover, as $\rho_j>1$ there exists $K\in\mathbb{N}$ such that $(1+k+1)/(1+k)<\rho_j$ for all $j\le J$ when $k>K$, so that

$$\frac{\theta_{\boldsymbol{\nu}+\boldsymbol{e}_j}}{\theta_{\boldsymbol{\nu}}}=\frac{(1+\nu_j+1)}{(1+\nu_j)\rho_j}<1\quad \forall j\le J \text{ and } \nu_j>K. \tag{44}$$

By defining a sequence of functions $(\theta_j^{(J,K)})_{j\ge1}$ as

$$\theta_j^{(J,K)}(k)=\begin{cases}\max_{k\le K}\rho_j^{-k}(1+k) & j\le J \text{ and } k\le K,\\ \rho_j^{-k}(1+k) & j>J \text{ or } k>K,\end{cases}$$

and defining a new sequence $(\Theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ as

$$\Theta_{\boldsymbol{\nu}}:=\prod_{j\ge1}\theta_j^{(J,K)}(\nu_j)\quad \forall\boldsymbol{\nu}\in\mathcal{F},$$

we have that $(\Theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ is monotonically decreasing by (43) and (44). Moreover, the monotone envelope of $(\theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ satisfies $\theta^*_{\boldsymbol{\nu}} \leq \Theta_{\boldsymbol{\nu}}$ for all $\boldsymbol{\nu} \in \mathcal{F}$. Therefore, we only need to show $(\Theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^t(\mathcal{F})$. By definition we have

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} \Theta^t_{\boldsymbol{\nu}} = \sum_{\boldsymbol{\nu}\in\mathcal{F}} \prod_{j\geq 1} (\theta_j^{(J,K)}(\nu_j))^t = \prod_{1\leq j\leq J} \sum_{k=0}^{\infty} (\theta_j^{(J,K)}(k))^t \prod_{j>J} \sum_{k=0}^{\infty} (\theta_j^{(J,K)}(k))^t. \tag{45}$$

Since $\rho_j > 1$, there exist a constant $C_j^{(K,J)} < \infty$ for each $j \geq 1$ such that

$$\sum_{k=0}^{\infty} \left(\theta_j^{(J,K)}(k)\right)^t = K \max_{k\leq K} \rho_j^{-tk}(1+k)^t + \sum_{k=K+1}^{\infty} \rho_j^{-tk}(1+k)^t < C_j^{(K,J)}.$$

Therefore, the first term of (45) can be bounded as

$$\prod_{1\leq j\leq J} \sum_{k=0}^{\infty} (\theta_j^{(J,K)}(k))^t \leq \prod_{1\leq j\leq J} C_j^{(K,J)} < \infty.$$

The second term of (45) can be bounded as in (37), i.e.,

$$\prod_{j>J} \sum_{k=0}^{\infty} \left(\theta_j^{(J,K)}(k)\right)^t = \prod_{j>J} \sum_{k=0}^{\infty} \rho_j^{-tk}(1+k)^t \leq \exp\left(2c_1 \sum_{j>J} \rho_j^{-1}\right),$$

which is finite when $(\rho_j^{-1})_{j\geq 1} \in \ell^t(\mathbb{N})$. Hence, $(\Theta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^t(\mathcal{F})$, which concludes. □

Remark 3.3 Note that the same convergence rate is obtained in Theorem 3.2 for downward closed and nested index sets as in Theorem 3.1 for more general index sets under Assumption 2.3. While under Assumption 2.4, the convergence rates for the Taylor approximation of u becomes different. Specifically, the convergence rate from $N^{-r(s)}$ is deteriorated to $N^{-r(t)}$ with $r(s) > r(t)$, as $s = \frac{2t}{2+t} < t$, for downward closed and nested index sets. This deterioration is due to the bound (34), which may be crude and the convergence rate may not be optimal.

4 Conclusions

We studied sparse polynomial approximations for parametric saddle point problems, which covered such problems as Stokes, mixed formulation of the Poisson, and time-harmonic Maxwell problems. We considered the setting of a random input parameter parametrized by a countably infinite number of independent parameters as the coefficients of an affine expansion on a series of basis functions. Both globally and locally supported basis functions were considered, which led to different assumptions on the sparsity of the parametrization. Based on the two different sparsity assumptions and the results in [4, 25] for affine parametric elliptic PDEs, we proved the ℓ^s-summability of the coefficients of the Taylor expansion of the parametric solutions by different approaches—analytic regularity and weighted ℓ^2-summability, respectively, for the saddle point problems. By the ℓ^s-summability we obtained the dimension-independent algebraic convergence rates of the sparse polynomial approximations, thus breaking the curse of dimensionality for high or infinite dimensional parametric saddle point problems. Moreover, we considered sparse polynomial approximations of the parametric solutions on downward closed and nested multi-index sets, which also have the dimension-independent convergence rates.

The analysis in this work can serve as a guideline for error estimates of model reduction techniques such as reduced basis methods constructed by greedy algorithms [16]. Note that we only considered uniformly distributed parameters in this work. We are interested in studying more general distributions such as Gaussian or log-normal random fields for saddle point problems, motivated by their recent analysis for elliptic PDEs [3, 11, 28]. Finally, we mention a particular type of parametric saddle point problem—optimality systems arising from stochastic PDE-constrained optimal control [12, 14, 35]. Application of the analysis to such problems are interesting.

Acknowledgements This work is partially supported by NSF under grant DMS-2012453, DOE under grants DE-SC0019303 and DE-SC0021239, Simons Foundation under grant 56065.

References

1. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **45**, 1005–1034 (2007)
2. Babuška, I., Tempone, R., Zouraris, G.E.: Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. **42**, 800–825 (2004)
3. Bachmayr, M., Cohen, A., DeVore, R., Migliorati, G.: Sparse polynomial approximation of parametric elliptic PDEs. Part II: lognormal coefficients. ESAIM Math. Model. Numer. Anal. **51**, 341–363 (2017)
4. Bachmayr, M., Cohen, A., Migliorati, G.: Sparse polynomial approximation of parametric elliptic PDEs. Part I: affine coefficients. ESAIM Math. Model. Numer. Anal. **51**, 321–339 (2017)
5. Benner, P., Ohlberger, M., Cohen, A., Willcox, K.: Model Reduction and Approximation: Theory and Algorithms. SIAM, Philadelphia (2017)
6. Berner, J., Grohs, P., Jentzen, A.: Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. SIAM J. Math. Data Sci. **2**, 631–657 (2020)
7. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., Wojtaszczyk, P.: Convergence rates for greedy algorithms in reduced basis methods. SIAM J. Math. Anal. **43**, 1457–1472 (2011)
8. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics, vol. 44. Springer, Berlin (2013)
9. Boyaval, S., Le Bris, C., Lelièvre, T., Maday, Y., Nguyen, N.C., Patera, A.T.: Reduced basis techniques for stochastic problems. Arch. Comput. Methods Eng. **17**, 435–454 (2010)
10. Buffa, A., Maday, Y., Patera, A.T., Prud'homme, C., Turinici, G.: A priori convergence of the greedy algorithm for the parametrized reduced basis method. ESAIM Math. Model. Numer. Anal. **46**, 595–603 (2012)
11. Chen, P.: Sparse quadrature for high-dimensional integration with Gaussian measure. ESAIM Math. Model. Numer. Anal. **52**, 631–657 (2018)
12. Chen, P., Quarteroni, A.: Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraints. SIAM/ASA J. Uncertain. Quantif. **2**, 364–396 (2014)
13. Chen, P., Quarteroni, A.: A new algorithm for high-dimensional uncertainty quantification based on dimension-adaptive sparse grid approximation and reduced basis methods. J. Comput. Phys. **298**, 176–193 (2015)
14. Chen, P., Quarteroni, A., Rozza, G.: Multilevel and weighted reduced basis method for stochastic optimal control problems constrained by Stokes equations. Numer. Math. **133**, 67–102 (2016)
15. Chen, P., Quarteroni, A., Rozza, G.: Reduced basis methods for uncertainty quantification. SIAM/ASA J. Uncertain. Quantif. **5**, 813–869 (2017)
16. Chen, P., Schwab, Ch.: Sparse-grid, reduced-basis Bayesian inversion. Comput. Methods Appl. Mech. Eng. **297**, 84–115 (2015)
17. Chen, P., Villa, U., Ghattas, O.: Hessian-based adaptive sparse quadrature for infinite-dimensional Bayesian inverse problems. Comput. Methods Appl. Mech. Eng. **327**, 147–172 (2017)

18. Chkifa, A., Cohen, A., Schwab, Ch.: Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. J. Math. Pures Appl. **103**, 400–428 (2015)
19. Chkifa, A., Cohen, A., DeVore, R., Schwab, Ch.: Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. ESAIM Math. Model. Numer. Anal. **47**, 253–280 (2013)
20. Chkifa, A., Cohen, A., Migliorati, G., Nobile, F., Tempone, R.: Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs. ESAIM Math. Model. Numer. Anal. **49**, 815–837 (2015)
21. Chkifa, A., Cohen, A., Schwab, Ch.: High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. Found. Comput. Math. **14**, 601–633 (2014)
22. Cliffe, K.A., Giles, M.B., Scheichl, R., Teckentrup, A.L.: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Visual. Sci. **14**, 3 (2011)
23. Cohen, A., DeVore, R.: Approximation of high-dimensional parametric PDEs. Acta Numer. **24**, 1–159 (2015)
24. Cohen, A., DeVore, R., Schwab, Ch.: Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**, 615–646 (2010)
25. Cohen, A., Devore, R., Schwab, Ch.: Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. Anal. Appl. **9**, 11–47 (2011)
26. Cohen, A., Migliorati, G.: Multivariate approximation in downward closed polynomial spaces. In: Dick, J., Kuo, F.Y., Woźniakowski, H. (eds.) Contemporary Computational Mathematics - a Celebration of the 80th Birthday of Ian Sloan, pp. 233–282. Springer, Cham (2018)
27. Doostan, A., Owhadi, H.: A non-adapted sparse approximation of PDEs with stochastic inputs. J. Comput. Phys. **230**, 3015–3034 (2011)
28. Ernst, O.G., Sprungk, B., Tamellini, L.: Convergence of sparse collocation for functions of countably many Gaussian random variables (with application to elliptic PDEs). SIAM J. Numer. Anal. **56**, 877–905 (2018)
29. Gantner, R.N., Schwab, Ch.: Computational higher order quasi-Monte Carlo integration. In: Cools, R., Nuyens, D. (eds.) Monte Carlo and Quasi-Monte Carlo Methods, pp. 271–288. Springer, Cham (2016)
30. Gerstner, T., Griebel, M.: Dimension–adaptive tensor–product quadrature. Computing **71**, 65–87 (2003)
31. Ghanem, R.G., Spanos, P.D.: Stochastic finite elements: a spectral approach. Courier corporation (2003)
32. Haji-Ali, A.-L., Nobile, F., Tempone, R.: Multi-index Monte Carlo: when sparsity meets sampling. Numer. Math. **132**, 767–806 (2016)
33. Hervé, M.: Analyticity in Infinite Dimensional Spaces. De Gruyter Studies in Mathematics, vol. 10. Walter de Gruyter, Berlin (1989)
34. Horgan, C.O.: Korn's inequalities and their applications in continuum mechanics. SIAM Rev. **37**, 491–511 (1995)
35. Kunoth, A., Schwab, Ch.: Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs. SIAM J. Control Optim. **51**, 2442–2471 (2013)
36. Kuo, F.Y., Schwab, Ch., Sloan, I.H.: Quasi-monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. SIAM J. Numer. Anal. **50**, 3351–3374 (2012)
37. Kutyniok, G., Petersen, P., Raslan, M., Schneider, R.: A theoretical analysis of deep neural networks and parametric PDEs. Constr. Approx. **55**, 73–125 (2022)
38. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. arXiv:2010.08895 (2020)
39. Narayan, A., Jakeman, J.D., Zhou, T.: A Christoffel function weighted least squares algorithm for collocation approximations. Math. Comput. **86**, 1913–1947 (2017)
40. Nobile, F., Tempone, R., Webster, C.G.: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**, 2309–2345 (2008)
41. O'Leary-Roseberry, T., Villa, U., Chen, P., Ghattas, O.: Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs. Comput. Methods Appl. Mech. Eng. **388**, 114199 (2022)
42. Quarteroni, A.: Numerical Models for Differential Problems, 2nd edn. Springer, Milano (2013)
43. Rauhut, H., Schwab, Ch.: Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations. Math. Comput. **86**, 661–700 (2017)
44. Schillings, C., Schwab, Ch.: Sparse, adaptive Smolyak quadratures for Bayesian inverse problems. Inverse Probl. **29**, 065011 (2013)
45. Schwab, Ch., Todor, R.A.: Karhunen–loève approximation of random fields by generalized fast multipole methods. J. Comput. Phys. **217**, 100–122 (2006)
46. Schwab, Ch., Zech, J.: Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ. Anal. Appl. **17**, 19–55 (2019)

47. Soize, C.: Random vectors and random fields in high dimension: Parametric model-based representation, identification from data, and inverse problems. In: Ghanem, R., Higdon, D., Owhadi, H. (eds.) Handbook of Uncertainty Quantification, pp. 883–935. Springer, Cham (2017)
48. Tran, H., Webster, C.G., Zhang, G.: Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients. Numer. Math. **137**, 451–493 (2017)
49. Xiu, D., Hesthaven, J.S.: High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. **27**, 1118–1139 (2005)
50. Xiu, D., Karniadakis, G.E.: The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**, 619–644 (2002)
51. Xu, J., Zikatanov, L.: Some observations on babuška and Brezzi theories. Numer. Math. **94**, 195–202 (2003)
52. Zech, J., Schwab, Ch.: Convergence rates of high dimensional Smolyak quadrature. ESAIM Math. Model. Numer. Anal. **54**, 1259–1307 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Peng Chen[1] · Omar Ghattas[2,3,4]

Omar Ghattas
omar@oden.utexas.edu

[1] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30308, USA

[2] Oden Institute for Computational Engineering & Sciences, The University of Texas at Austin, Austin, TX 78712, USA

[3] Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78712, USA

[4] Department of Geological Sciences, The University of Texas at Austin, Austin, TX 78712, USA