Acta Numerica (2015), pp. 1–159
doi:10.1017/S0962492915000033
© Cambridge University Press, 2015
Printed in the United Kingdom

Approximation of high-dimensional parametric PDEs*

Albert Cohen
Laboratoire Jacques-Louis Lions,
Université Pierre et Marie Curie,
75005 Paris, France
E-mail: cohen@ann.jussieu.fr

Ronald DeVore
Department of Mathematics,
Texas A&M University,
College Station, TX 77840, USA
E-mail: ronald.a.devore@gmail.com

Parametrized families of PDEs arise in various contexts such as inverse problems, control and optimization, risk assessment, and uncertainty quantification. In most of these applications, the number of parameters is large or perhaps even infinite. Thus, the development of numerical methods for these parametric problems is faced with the possible curse of dimensionality. This article is directed at (i) identifying and understanding which properties of parametric equations allow one to avoid this curse and (ii) developing and analysing effective numerical methods which fully exploit these properties and, in turn, are immune to the growth in dimensionality.

Part I of this article studies the smoothness and approximability of the solution map, that is, the map $a \mapsto u(a)$, where a is the parameter value and $u(a)$ is the corresponding solution to the PDE. It is shown that for many relevant parametric PDEs, the parametric smoothness of this map is typically holomorphic and also highly anisotropic, in that the relevant parameters are of widely varying importance in describing the solution. These two properties are then exploited to establish convergence rates of n-term approximations to the solution map, for which each term is separable in the parametric and physical variables. These results reveal that, at least on a theoretical level, the solution map can be well approximated by discretizations of moderate complexity, thereby showing how the curse of dimensionality is broken. This theoretical analysis is carried out through concepts of approximation theory such as best n-term approximation, sparsity, and n-widths. These notions

* This research was supported by ONR contracts N00014-11-1-0712 and N00014-12-1-0561, NSF grant DMS 1222715, the Institut Universitaire de France and ERC advanced grant BREAD.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

determine *a priori* the best possible performance of numerical methods and thus serve as a benchmark for concrete algorithms.

Part II of this article turns to the development of numerical algorithms based on the theoretically established sparse separable approximations. The numerical methods studied fall into two general categories. The first uses polynomial expansions in terms of the parameters to approximate the solution map. The second one searches for suitable low-dimensional spaces for simultaneously approximating all members of the parametric family. The numerical implementation of these approaches is carried out through adaptive and greedy algorithms. An *a priori* analysis of the performance of these algorithms establishes how well they meet the theoretical benchmarks.

CONTENTS

1 Overview 2

Part I: Smoothness and approximation results
2 Holomorphic extensions 20
3 Best n-term polynomial approximations 40
4 Estimating the n-widths of solution manifolds 74

Part II: Algorithms for parametric PDEs
5 Towards concrete algorithms 87
6 Sparse polynomial interpolation 94
7 Taylor approximation 113
8 Reduced basis methods 132

References 154

1. Overview

1.1. Parametric and stochastic PDEs

Partial differential equations (PDEs) are commonly used to model complex systems in a variety of physical contexts. When solving a given PDE, one typically fixes certain *parameters*: the shape of the physical domain, the diffusion or velocity field, the source term, the flux or reaction law, *etc.* We use the terminology *parametric PDEs* when some of these parameters are allowed to vary over a certain range of interest. When treating such parametric PDEs, one is interested in finding the solution for all parameters in the range of interest.

To describe such problems in their full generality, we adopt the formulation

$$\mathcal{P}(u, a) = 0, \tag{1.1}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

where a denotes the parameters, u is the unknown of the problem, and

$$\mathcal{P} : V \times X \to W \tag{1.2}$$

is a linear or nonlinear partial differential operator, with (X, V, W) a triplet of Banach spaces. We assume that the parameter a ranges over a compact set $\mathcal{A} \subset X$, and that for any a in this range there exists a unique solution $u = u(a) \in V$ to (1.1). This allows us to define the *solution map*

$$u : a \mapsto u(a), \tag{1.3}$$

which acts from X onto V and is well defined over $\mathcal{A}$. We also define the *solution manifold* as the family

$$\mathcal{M} = u(\mathcal{A}) = \{u(a) : a \in \mathcal{A}\}, \tag{1.4}$$

which gathers together all solutions as the parameter varies within its range.

One simple guiding example, which will be often used throughout this article, is the linear elliptic equation

$$\begin{aligned} -\operatorname{div}(a\nabla u) &= f \quad \text{on } D, \\ u &= 0 \quad \text{on } \partial D, \end{aligned} \tag{1.5}$$

set on a Lipschitz domain $D \subset \mathbb{R}^m$. Here, we fix the right-hand side f as a real-valued function and consider real-valued diffusion coefficients a as the parameter. The corresponding operator $\mathcal{P}$ is therefore given by

$$\mathcal{P}(u, a) = f + \operatorname{div}(a\nabla u). \tag{1.6}$$

A possible choice for the triplet of spaces is then

$$(X, V, W) = (L^\infty(D), H_0^1(D), H^{-1}(D)). \tag{1.7}$$

Indeed, if $u \in V$, $a \in X$ and $f \in W$, one then defines $\mathcal{P}(u, a)$ as an element of W by requiring that

$$\langle \mathcal{P}(u, a), v\rangle = \langle f, v\rangle + \int_D a\nabla u \cdot \nabla v, \quad v \in V, \tag{1.8}$$

where $\langle \cdot, \cdot \rangle$ is the duality bracket between $V' = W$ and V. Lax–Milgram theory ensures the existence and uniqueness of a solution $u(a)$ to (1.1) from V if, for some $r > 0$, the diffusion a satisfies the ellipticity condition

$$a(x) \geq r, \quad x \in D. \tag{1.9}$$

Therefore, a typical parameter range is a set $\mathcal{A} \subset \{a \in L^\infty(D) : a \geq r\}$, which in addition is assumed to be compact in L^∞.

Although elementary, the above example gathers important features that are present in other relevant examples of parametric PDEs. In particular, the solution map $a \mapsto u(a)$ acts from an infinite-dimensional space into another infinite-dimensional space. Also note that, while the operator $\mathcal{P}$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

of (1.6) is linear both in a and u (up to the constant additive term f), the solution map is *nonlinear*. Because of the high-dimensionality of the parameter space X, such problems represent a significant challenge when trying to capture this map numerically. One objective of this article is to understand which properties of this map allow for a successful numerical treatment. Concepts such as holomorphy, sparsity, and adaptivity are at the heart of our development.

The solution map may also be viewed as a function

$$(x, a) \mapsto u(x, a) \tag{1.10}$$

of both the physical variable $x \in D$ and the parametric variable $a \in \mathcal{A}$. The parametric variable has a particular status because its different instances are uncoupled: for any fixed instance $a = a_0$, we may solve the PDE exactly or approximately and therefore compute $u(x, a_0)$ for all values of x, while ignoring all other values $a \neq a_0$. This plays an important role for certain numerical methods which are based on solving the parametric PDE for different particular values of a, since this task can be parallelized.

Parametric PDEs occur in a variety of modelling contexts. We draw the following major distinctions in their setting.

- *Deterministic modelling.* The parameters are deterministic *design or control* variables, which may be tuned by the user so that the solution u, or some quantity of interest $Q(u)$, has prescribed properties. For instance, if the elliptic equation (1.5) is used to model the heat conduction in a component produced by an industrial process, one may want to design the material in order to minimize the heat flux on a certain part on the boundary $\Gamma \subset \partial D$, in which case the quantity of interest to be optimized is

$$Q(u) = \int_\Gamma a \nabla u \cdot \mathbf{n}, \tag{1.11}$$

  where $\mathbf{n}$ is the outer normal. This amounts to optimizing the function

$$a \mapsto F(a) := Q(u(a)) \tag{1.12}$$

  over $\mathcal{A}$.

- *Stochastic modelling.* The parameters are random variables with prescribed probability laws, which account for *uncertainties* in the model. Therefore a has a certain probability distribution μ supported on $\mathcal{A}$. One is then typically interested in the resulting probabilistic properties of the solution u, which is itself a random variable over $\mathcal{A}$ with values in V, or in the probabilistic properties of a quantity of interest $Q(u)$. For instance, if the elliptic equation (1.5) is used to model oil or ground water diffusion, a common way to deal with the uncertainty of the

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

underground porous media is to define a as a random field with some prescribed law. Then one might want to estimate the mean solution

$$\overline{u} := \mathbb{E}(u), \tag{1.13}$$

or its variance

$$V(u) := \mathbb{E}(\|u - \overline{u}\|_V^2), \tag{1.14}$$

or the average flux through a certain interface Γ, that is, $\mathbb{E}(Q(u))$ with $Q(u)$ as in (1.11), or the probability that this flux exceeds a certain quantity, *etc.*

In both deterministic and stochastic settings, the given application may require evaluation of $u(a^i)$ for a large number $n \gg 1$ of values $\{a^1, \ldots, a^n\}$ of the parameter a. This is the case, for instance, when using a descent method for optimizing a quantity of interest in the deterministic framework, or a Monte Carlo method for evaluating an expectation in the stochastic framework. Since each individual instance $u(a)$ is the solution of a PDE, its exact evaluation is typically out of reach. Instead, each query for $u(a^i)$ is approximately evaluated by a numerical solver, which may itself be computationally intensive for high accuracy.

In order to significantly reduce the number of computations needed for attaining a prescribed accuracy, alternative strategies, commonly referred to as *reduced modelling*, have been developed. Understanding which of these strategies are effective in the case where the parameter has large or infinite dimension, and why, is the subject of this article.

1.2. Affine representation of the parameters

So far our description of the set $\mathcal{A}$ of parameters allows it to be any compact subset of X. An important ingredient in both our theoretical and numerical developments is to identify any $a \in \mathcal{A}$ through a sequence of real numbers. We are especially interested in affine representations of $\mathcal{A}$. We say that a sequence $(\psi_j)_{j\geq 1}$ of functions $\psi_j \in X$ is an *affine representer* for $\mathcal{A}$, or *representer* for short, if we can write each a as

$$a = a(y) = \overline{a} + \sum_{j\geq 1} y_j \psi_j, \quad y := (y_j)_{j\geq 1}, \quad y_j = y_j(a), \tag{1.15}$$

where the y_j are real numbers, $\overline{a}$ is a fixed function from X, and the series converges in the norm of X for each $a \in \mathcal{A}$. We are making a slight abuse of notation here since we use a to represent a general element of $\mathcal{A}$ and also use a to represent the map

$$a : y \mapsto a(y), \tag{1.16}$$

from $\mathbb{R}^{\mathbb{N}}$ to X. However, the meaning will always be clear from the context.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

It is easy to see that for any compact set $\mathcal{A}$ in a Banach space X, affine representers exist. For example, if X has a Schauder basis then any such basis will be a representer. Even if X does not admit a Schauder basis, as is the case for our example $X = L^\infty(D)$, we can still find representers as follows. Choose any $\bar{a} \in \mathcal{A}$. Since $\mathcal{K} := \mathcal{A} - \bar{a}$ is compact, there exist finite-dimensional spaces $(X_n)_{n\geq 0}$, with $\dim(X_n) = n$, such that

$$\operatorname{dist}(\mathcal{K}, X_n)_X := \sup_{a\in\mathcal{K}} \min_{b\in X_n} \|a - b\|_X \to 0 \quad \text{as } n \to \infty. \tag{1.17}$$

We can also take the spaces X_n to be nested, that is

$$X_n \subset X_{n+1}, \quad n \geq 0. \tag{1.18}$$

Let $(\phi_{j,n})_{j=1,\dots,n}$ be any basis for X_n and define $N_n := n(n-1)/2$. The sequence

$$\psi_j := \phi_{j-N_n,n}, \quad j = N_n, \dots, N_{n+1}, \tag{1.19}$$

contains each of the bases $(\phi_{j,n})_{j=1,\dots,n}$ for all $n \geq 1$. Given any $a \in \mathcal{A}$, let a_n be a best approximation in X to $a - \bar{a}$ from X_n, with $a_0 := 0$. Then we can write

$$a = \bar{a} + \sum_{n=1}^{\infty} (a_n - a_{n-1}). \tag{1.20}$$

Each term $a_n - a_{n-1}$ is in X_n and hence can be written as a linear combination of the ψ_j. Therefore, $(\psi_j)_{j\geq 1}$ is a representer for $\mathcal{A}$.

Affine representations (1.15) often occur in the natural formulation of the parametric problem. For instance, if the diffusion coefficient a in (1.5) is piecewise constant over a fixed partition $\{D_j\}_{j=1,\dots,d}$ of the physical domain D, then it is natural to set

$$a(y) = \overline{a} + \sum_{j=1}^{d} y_j \chi_{D_j}, \tag{1.21}$$

where $\overline{a}$ is a constant and the χ_{D_j} are the characteristic functions of the subdomains D_j. Similarly, if the parameter a describes the shape of the boundary of the physical domain in a computer-aided design setting, a typical format is

$$a(y) = \overline{a} + \sum_{j=1}^{d} y_j B_j, \tag{1.22}$$

where $\overline{a}$ represents a nominal shape and the B_j are B-spline functions associated to control points. In these two examples, d is finite, yet possibly very large.

In the statistical context, if a is a second-order random field over a domain D, a frequently used choice in (1.15) is $\overline{a} := \mathbb{E}(a)$, the average field,

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and $(\psi_j)_{j\geq 0}$, the Karhunen–Loève basis, that is, the eigenfunctions of the covariance operator

$$v \mapsto R_a v := \int_D C_a(\cdot, x)v(x)\,\mathrm{d}x, \quad C_a(z,x) := \mathbb{E}((a(z)-\overline{a}(z))(a(x)-\overline{a}(x)). \tag{1.23}$$

Then the resulting scalar variables are centred and uncorrelated, that is, $\mathbb{E}(y_i) = 0$ and $\mathbb{E}(y_i y_j) = 0$ when $i \neq j$.

Even if an affine representation of the form (1.15) is not given in the formulation of the problem, one can be derived by taking any representation system $(\psi_j)_{j\geq 1}$ in the Banach space X. For example, if X admits a Schauder basis, then one can take any such basis $(\psi_j)_{j\geq 1}$ for X and arrive at such an expansion. In classical spaces X, such as L^p or Sobolev spaces, standard systems of approximation, such as Fourier series, splines, or wavelets can be used.

The advantage of the representation (1.15) is that a can now be identified through the sequence $(y_j)_{j\geq 1}$. When considering all $a \in \mathcal{A}$, we obtain a family of such sequences. Note that this family can be quite complicated. In order to simplify matters, we normalize the ψ_j, so that for any j,

$$\sup_{a\in\mathcal{A}} |y_j(a)| = 1. \tag{1.24}$$

Such a renormalization is usually possible because $\mathcal{A}$ is compact and $y_j(a)$ depends continuously on a. After this normalization, for each $a \in \mathcal{A}$, the sequence $(y_j(a))_{j\geq 1}$ belongs to the infinite-dimensional cube

$$U := [-1,1]^{\mathbb{N}}. \tag{1.25}$$

Note that taking a general sequence $(y_j)_{j\geq 1}$ from this cube, there may not be an $a \in \mathcal{A}$ with $y_j = y_j(a)$, $j \geq 1$. Also, if $\{\psi_j\}_{j\geq 1}$ is not a basis, the representation (1.15) may not be unique. We define

$$U_{\mathcal{A}} := \left\{(y_j)_{j\geq 1} \in U : \sum_{j\geq 1} y_j\psi_j \in \mathcal{A}\right\}. \tag{1.26}$$

We are mainly interested in representers $\bar{a}$, $(\psi_j)_{j\geq 1}$ for which

$$\bar{a} + \sum_{j\geq 1} y_j\psi_j \tag{1.27}$$

converges in X for each $(y_j) \in U$. We call such representers *complete*. In this case, we may define

$$a(U) := \left\{a = a(y) = \overline{a} + \sum_{j\geq 1} y_j\psi_j : (y_j)_{j\geq 1} \in U\right\}, \tag{1.28}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

so that

$$\mathcal{A} \subset a(U). \tag{1.29}$$

A typical case of a complete representer is when $(\|\psi_j\|_X)$ is a sequence in $\ell^1(\mathbb{N})$.

Once an affine representation has been chosen, the initial solution map $a \mapsto u(a)$ becomes equivalent to the map $y \mapsto u(a(y))$ which is defined on $U_{\mathcal{A}}$. With an abuse of notation, we write this new solution map as

$$y \mapsto u(y) := u(a(y)). \tag{1.30}$$

This is a Banach-space-valued function of an infinite number of variables. Note that in the case where the affine representation has a finite number d of terms, the range of y is $[-1,1]^d$. However, the infinite-dimensional case subsumes the finite-dimensional case, since the latter may be viewed as a particular case with $\psi_j = 0$ for $j > d$.

In the case of a complete representer, $a(y)$ is defined on all of U. However, we do not know whether the solution map u is defined on all of U. To guarantee this, the following assumption will be used often.

Assumption A. The parameter set $\mathcal{A}$ has a complete representer $(\psi_j)_{j\geq 1}$ and the solution map $a \mapsto u(a)$ is well defined on the whole set $a(U)$, or equivalently the solution map $y \mapsto u(y)$ is well defined on the whole set U.

This assumption naturally holds when the set $\mathcal{A}$ is exactly defined as $a(U)$.

1.3. Smoothness of the solution map

One objective of this article is to develop efficient numerical approximations to the solution maps of (1.3) or (1.30). One of the main difficulties is that these maps are high- or infinite-dimensional, in the sense that the dimension of the variable a or y is high or infinite. In order to understand what might be good strategies for constructing such approximations, we need first to understand the inherent properties of these maps that might allow us to circumvent this difficulty.

We initiate such a program in Section 2, where we first analyse the smoothness of the solution map $a \mapsto u(a)$. In the case of the elliptic equation (1.5), it is easily seen that this map is not only infinitely differentiable, but also admits a holomorphic extension to certain subdomains of the complex-valued $X = L^\infty(D)$. We propose two general approaches which allow us to establish similar holomorphy properties for other relevant instances of linear and nonlinear parametric PDEs. The first approach is based on the Ladyzhenskaya–Babuška–Brezzi (LBB) theory. It applies to a range of linear PDEs where the operator and the right-hand side have holomorphic dependence in a. These include parabolic and saddle-point problems, such

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

as the heat equations or the Stokes problem, with parameter a in the diffusion term, similar to (1.5). The second approach is based on the implicit function theorem in complex-valued Banach spaces. In contrast to the first approach, it can be applied to certain nonlinear PDEs.

Using the affine representation (1.15) of a, we then study the solution map $y \mapsto u(y)$ under Assumption A, which means that it is defined on the whole of U. In addition to holomorphy, an important property of u can be extracted from the affine representation (1.15). The functions ψ_j appearing in (1.15) have norms $\|\psi_j\|_X$ of varying size. Since the variable y_j is scaled to be in $[-1,1]$, when $\|\psi_j\|_X$ is small, this variable has a reduced effect on the variations of $u(y)$. Thus the variables $(y_j)_{j\geq 1}$ are not democratic, but rather they have varying importance. In other words, the map $y \mapsto u(y)$ is highly anisotropic. More specifically, we derive holomorphic extension results for this map on certain multivariate complex domains of tensor product type. In particular, we consider polydiscs of the general form

$$\mathcal{U}_\rho := \otimes\{|z_j| \leq \rho_j\} = \{z = (z_j)_{j\geq 1} \in \mathbb{C}^{\mathbb{N}} : |z_j| \leq \rho_j\}, \tag{1.31}$$

where $\rho = (\rho_j)_{j\geq 1}$ is a positive sequence which serves to describe the anisotropy of the solution map. We also consider polyellipses, which deviate less far from the real axis. These holomorphy domains play a key role in the derivation of approximation results.

Remark 1.1. While we are generally interested in real-valued solutions u to the parametric PDE (1.1), corresponding to real-valued parameters a or y, our analysis of holomorphic smoothness leads us naturally to complex-valued solutions, corresponding to complex-valued parameters. For this reason, the spaces X, V, W are always assumed to be complex-valued Banach spaces throughout this paper.

1.4. Approximation of the solution map

Reduced modelling methods seek to take advantage of the properties of the solution maps $a \mapsto u(a)$ or $y \mapsto u(y)$ such as holomorphy and anisotropy mentioned above. These properties suggest strategies for approximating these maps u by simple functions u_n, in which the physical variables x and the parametric variables a or y are separated and hence take the form

$$(x,a) \mapsto u_n(x,a) := \sum_{i=1}^{n} v_i(x)\phi_i(a), \tag{1.32}$$

or

$$(x,y) \mapsto u_n(x,y) := \sum_{i=1}^{n} v_i(x)\phi_i(y), \tag{1.33}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

where $\{v_1, \dots, v_n\}$ are functions of x living in the solution space V and $\{\phi_1, \dots, \phi_n\}$ are functions of a or y with values in $\mathbb{R}$ or $\mathbb{C}$.

We may view u_n as a rank n approximation to u, in analogy with low-rank approximation of matrices. We adopt the notations

$$a \mapsto u_n(a) = u_n(\cdot, a) \quad \text{and} \quad y \mapsto u_n(y) = u_n(\cdot, y), \tag{1.34}$$

for the above approximations.

Let us discuss the potential accuracy of separable approximations of the form (1.32). If our objective is to capture $u(a)$ for all $a \in \mathcal{A}$ with a prescribed accuracy $\varepsilon(n)$, this means that we search for an error bound in the uniform sense, that is, of the form

$$\|u - u_n\|_{L^\infty(\mathcal{A},V)} := \sup_{a\in\mathcal{A}} \|u(a) - u_n(a)\|_V \le \varepsilon(n). \tag{1.35}$$

For certain applications, in particular in the stochastic framework, we may instead decide to measure the error on average, for instance by searching for an error bound in the mean square sense,

$$\mathbb{E}(\|u - u_n\|_V^2) := \|u - u_n\|^2_{L^2(\mathcal{A},V,\mu)} = \int_{\mathcal{A}} \|u(a) - u_n(a)\|_V^2 \, \mathrm{d}\mu(a) \le \varepsilon(n)^2, \tag{1.36}$$

where μ is the probability measure for the distribution of a over $\mathcal{A}$. Since μ is a probability measure, we have for any v

$$\|v\|_{L^2(\mathcal{A},V,\mu)} \le \|v\|_{L^\infty(\mathcal{A},V)}. \tag{1.37}$$

Therefore the uniform bound is stronger than the average bound, in the sense that (1.35) implies (1.36) with the same value of $\varepsilon(n)$.

Likewise, for the approximation of the map $y \mapsto u(y)$, we may search for a uniform bound

$$\|u - u_n\|_{L^\infty(U_\mathcal{A},V)} := \sup_{y\in U_\mathcal{A}} \|u(y) - u_n(y)\|_V \le \varepsilon(n) \tag{1.38}$$

or a mean-square bound

$$\mathbb{E}(\|u - u_n\|_V^2) := \|u - u_n\|^2_{L^2(U_\mathcal{A},V,\mu)} = \int_{U_\mathcal{A}} \|u(y) - u_n(y)\|_V^2 \, \mathrm{d}\mu(y) \le \varepsilon(n)^2, \tag{1.39}$$

where μ is the probability measure for the distribution of y over $U_\mathcal{A}$.

Remark 1.2. We do not indicate the measure μ in our notation $L^\infty(\mathcal{A}, V)$ or $L^\infty(U_\mathcal{A}, V)$, since in all relevant examples considered in this article we always consider the exact supremum over a in $\mathcal{A}$ or over y in $U_\mathcal{A}$, rather than the essential supremum.

For any $a \in \mathcal{A}$, the approximation $u_n(a)$ belongs to

$$V_n := \operatorname{span}\{v_1, \dots, v_n\}, \tag{1.40}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

which is a fixed n-dimensional subspace of V. Ideal benchmarks for the performance of separable expansions of the form (1.32) may thus be defined by selecting optimal n-dimensional spaces for the approximation of $u(a)$ in either a uniform or an average sense.

For uniform error bounds, this benchmark is given by the concept of Kolmogorov's *n-width*, which is well known in approximation theory. If $\mathcal{K}$ is a compact set in a Banach space V, we define its Kolmogorov n-width as

$$d_n(\mathcal{K})_V := \inf_{\dim(V_n)\leq n} \max_{v\in\mathcal{K}} \min_{w\in V_n} \|v-w\|_V. \tag{1.41}$$

This quantity, first introduced in Kolmogorov (1936), describes the best achievable accuracy, in the norm of V, when approximating all possible elements of $\mathcal{K}$ by elements from a linear n-dimensional space V_n. Obviously, the optimal choice of V_n^* for approximation of $u(a)$ in a uniform sense corresponds to the space, if it exists, that reaches the above infimum when $\mathcal{K}$ is taken to be the solution manifold $\mathcal{M}$. The best achievable error in the uniform sense is thus given by

$$\varepsilon(n) := d_n(\mathcal{M})_V. \tag{1.42}$$

There exist many other notions of widths that are used to measure the size of compact sets. Here we only work with the above one and refer to Pinkus (1985) for a more general treatment.

For mean-square bounds, in the case where V is a Hilbert space, the corresponding benchmark is related to the concept of principal component analysis, which is of common use in statistics. By choosing an arbitrary orthonormal basis $(e_i)_{i\geq 0}$ of V, we may expand $u(a)$ according to

$$u(a) = \sum_{i\geq 0} z_i e_i, \quad z_i = z_i(a) := \langle u(a), e_i\rangle_V. \tag{1.43}$$

Approximation of $u(a)$ from an n-dimensional space of V is then equivalent to the approximation of $\mathbf{z}(a) := (z_i(a))_{i\geq 0}$ from an n-dimensional space of $\ell^2(\mathbb{N})$. The optimal space is then obtained via study of the correlation operator

$$R = (R_{i,j})_{i,j\geq 0}, \quad R_{i,j} := \mathbb{E}(z_i(a)z_j(a)) = \int_{a\in\mathcal{A}} z_i(a)z_j(a)\,\mathrm{d}\mu(a). \tag{1.44}$$

This operator is symmetric, positive, compact and of trace class. It therefore admits an orthonormal basis of eigenvectors $(\mathbf{g}_k)_{k\geq 1}$ associated to a positive, non-increasing and summable sequence $(\lambda_k)_{k\geq 1}$ of eigenvalues. The space

$$G_n := \operatorname{span}\{\mathbf{g}_1, \ldots, \mathbf{g}_n\} \tag{1.45}$$

minimizes the mean-square error between $\mathbf{z}$ and its orthogonal projection

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$P_G\mathbf{z}$ over all n-dimensional spaces G, with

$$\mathbb{E}(\|\mathbf{z} - P_{G_n}\mathbf{z}\|_{\ell^2}^2) = \sum_{k>n} \lambda_k. \tag{1.46}$$

In turn, the optimal space V_n^* is spanned by the functions

$$v_k := \sum_{i\geq 0} g_{k,i} e_i, \quad k = 1, \dots, n, \tag{1.47}$$

where $g_{k,i}$ is the ith component of $\mathbf{g}_k$, that is, $\mathbf{g}_k = (g_{k,i})_{i\geq 0}$. The best achievable error in the mean square sense is thus given by

$$\varepsilon(n)^2 := \sum_{k>n} \lambda_k. \tag{1.48}$$

Note that, in view of (1.37), we have the comparison

$$\sum_{k>n} \lambda_k \leq d_n(\mathcal{M})_V^2. \tag{1.49}$$

The optimal spaces described above are usually out of reach, both from an analytic and computational point of view, and it is therefore interesting to consider sub-optimal approximations. In addition, when considering the solution map $y \mapsto u(y)$, the tensor product structure of U allows us to consider approximations based on further separation between the parametric variables, that is, where each function ϕ_i in (1.33) is itself a product of univariate functions of the different y_j. The simplest examples of such approximations are multivariate polynomials, which have the general form

$$u_n(x, y) = \sum_{\nu\in\Lambda_n} v_\nu(x) y^\nu, \quad y^\nu := \prod_{j\geq 1} y_j^{\nu_j}, \tag{1.50}$$

where each index $\nu = (\nu_j)_{j\geq 1} \in \Lambda_n$ is a finitely supported sequence of positive integers, or equivalently such that $|\nu| = \sum_{j\geq 1} \nu_j < \infty$, and Λ_n is a set of n such sequences.

In Section 3 we obtain such polynomial approximations by taking finite portions of infinite polynomial expansions of u. Here we work under Assumption A, which means that $u(y)$ is defined on the whole of U. We consider two types of expansion:

- power series of the form

$$u(y) = \sum_{\nu\in\mathcal{F}} t_\nu y^\nu, \tag{1.51}$$

  where $\mathcal{F}$ is the set of all finitely supported sequences of positive integers;

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

- orthogonal series of the form

$$u(y) = \sum_{\nu \in \mathcal{F}} w_\nu P_\nu(y), \quad P_\nu(y) := \prod_{j \geq 1} P_{\nu_j}(y_j), \tag{1.52}$$

where P_k is the Legendre polynomial of degree k defined on $[-1, 1]$.

Using the holomorphy and anisotropy properties of the solution map $y \mapsto u(y)$ established in Section 2 for specific classes of parametric PDEs, we derive *a priori* bounds on the V-norms $\|t_\nu\|_V$ and $\|w_\nu\|_V$ of the coefficients which appear in these expansions. In this way, we are able to establish algebraic convergence rates n^{-s} for certain truncations of the above expansions, where n is the cardinality of the truncation set Λ_n.

One critical aspect of the truncation strategy is that we retain the n largest coefficients, which is a form of *nonlinear approximation*, also known as *sparse* or *best n-term* approximation. With such a choice for Λ_n, the exponent s in the convergence rate is related to the available ℓ^p-summability for $p < 1$ of the V-norms of the coefficients in the considered infinite expansion. In particular, for uniform approximation estimates, that is, in $L^\infty(U, V)$, we have

$$s = \frac{1}{p} - 1, \tag{1.53}$$

once the ℓ^p-summability of these V-norms has been proved. The main result from Section 3 shows that, under suitable assumptions, the ℓ^p-summability of the sequence $(\|\psi_j\|_X)_{j \geq 0}$ implies that the norms of the coefficients in the expansions (1.51) or (1.52) are also ℓ^p-summable.

The fact that we obtain the algebraic convergence rate n^{-s} despite the infinite-dimensional nature of the variable $(y_j)_{j \geq 1}$ reveals that the curse of dimensionality can be avoided in the approximation of relevant parametric PDEs.

1.5. The n-widths of the solution manifold $\mathcal{M}$

In both the uniform or mean-square ways of measuring error, as described above, the success of reduced modelling can be proved if $d_n(\mathcal{M})_V$ converges to zero sufficiently fast as $n \to +\infty$. Thus, the study of these widths constitutes a major subject in the theoretical justification of reduced modelling.

A common way to measure the widths of compact sets $\mathcal{K}$ in classical spaces is to embed $\mathcal{K}$ into an appropriate smoothness space such as a Sobolev or Besov space. For example, in our model parametric elliptic equation (1.5), the space V is $H_0^1(D)$, and this approach would lead us to examine the H^m Sobolev smoothness of the individual functions $u(a)$ for $m > 1$. Classical elliptic regularity theory says that for smooth domains the smoothness of $u(a)$ can be inferred from the smoothness of the right-hand side f. However,

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

for general domains, there are severe limits on this regularity due to the irregularity of the boundary. Therefore, bounding the decay of widths of $\mathcal{M}$ by means of such regularity results will generally prove only slow decay for $d_n(\mathcal{M})_V$ and is therefore not useful for obtaining the fast decay rates we seek. Indeed, recall that classical smoothness spaces, such as Sobolev or Besov spaces of order m in d variables, have widths that decay, at best, like $n^{-(m-r)/d}$ as $n \to \infty$, when these widths are measured in a $W^{r,p}$-norm for some $r < m$. For example, it is known that if $\mathcal{K}$ is the unit ball of $C^m([-1,1]^d)$, then its n-width in L^∞ satisfies

$$c_d n^{-m/d} \leq d_n(\mathcal{K})_{L^\infty} \leq C_d n^{-m/d}, \quad n \geq 1. \tag{1.54}$$

Likewise, if $\mathcal{K}$ is the unit ball of $H^m([-1,1]^d)$, then its n-width in H^1 satisfies

$$c_d n^{-(m-1)/d} \leq d_n(\mathcal{K})_{H^1} \leq C_d n^{-(m-1)/d}, \quad n \geq 1. \tag{1.55}$$

Also note that the regularity of $u(a)$, as measured by membership in Sobolev and Besov spaces, is closely related to the performance of piecewise polynomial approximation such as that used in finite element methods, and is the reason why these algorithms have rather slow convergence. So the fast decay of $d_n(\mathcal{M})_V$ to zero cannot be obtained by such an approach.

The widths $d_n(\mathcal{M})_V$ go to zero fast not because the individual elements in $\mathcal{M}$ are smooth in the physical variable, but rather because $\mathcal{M}$ is the image of the solution map u which is, as previously discussed, smooth and anisotropic in the parametric variable. However, let us remark that it is by no means trivial to deduce fast decay for $d_n(\mathcal{M})_V$ from this fact alone, because of what is called *the curse of dimensionality*. Namely, approximation rates for a given target function are generally proved by showing that the target function, in our case the function u, has sufficiently high regularity. But, in high dimensions, regularity by itself is usually not enough. Indeed, returning to the bounds (1.54), we see that the large dimension d affects the approximation rate in two detrimental ways. The exponent in the smoothness rate is divided by d and the constants C_d are known to grow exponentially with increasing d. In our case d can even be infinite, and this makes the derivation of approximation rates a subtle problem.

In Section 4 we discuss general principles for estimating the n-widths $d_n(\mathcal{M})_V$ of the solution manifold $\mathcal{M}$ by the above. One immediate consequence of the approximation results in Section 3 is that, for specific classes of parametric PDEs, when Assumption A holds and $(\|\psi_j\|_X)_{j\geq 0} \in \ell^p$, then these widths decay at least like n^{-s} with s given by (1.53). We extend this analysis to the general case when Assumption A does not necessarily hold. Since the manifold $\mathcal{M}$ is not directly accessible, one would like to understand what properties of $\mathcal{A}$, which is assumed to be completely known to us, imply decay rates for $d_n(\mathcal{M})$. We show that the asymptotic decay of the n-width of $\mathcal{M}$ in V is related to that of the n-width of $\mathcal{A}$ in X. This follows

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

from general results on widths of images of compact sets under holomorphic maps. Namely, if $u : X \to V$ is holomorphic in a neighbourhood of a general compact set $\mathcal{A} \subset X$, then we prove the following result on the n-width of the image $\mathcal{M} = u(\mathcal{A})$ in V:

$$\sup_{n\geq 1} n^r d_n(\mathcal{A})_X < \infty \Rightarrow \sup_{n\geq 1} n^s d_n(\mathcal{M})_V < \infty, \quad s < r - 1. \tag{1.56}$$

This result shows that from the viewpoint of preserving n-widths, holomorphic maps behave almost as well as linear maps, up to a loss of 1 in the convergence rate. One open problem is to understand whether this loss is sharp or could be improved.

1.6. Numerical methods for reduced modelling

The above-mentioned results on holomorphic extensions, sparse expansions for u, and n-widths of the manifold $\mathcal{M}$, can be thought of as theoretical justifications for the role of reduced modelling in solving parametric and stochastic equations. They provide evidence that reduced modelling numerical methods should yield significant computational savings over traditional methods such as finite element solvers for parametric problems or Monte Carlo methods for stochastic problems. However, they do not constitute actual numerical methods.

Part II of our article turns to the construction of numerical algorithms motivated by these theoretical results. Such algorithms compute specific separable approximations of the form (1.32) or (1.33) for a given value of n at an affordable computational cost. Our objective, in this regard, is not to give an exhaustive description of numerical reduced modelling methods and their numerous variants. Rather, our main focus is to introduce some important representative examples of those methods for which an *a priori* analysis quantifies the gains in numerical performance.

One important distinction is between *non-adaptive* and *adaptive* methods. In the former, the choice of functions $\{\phi_1, \ldots, \phi_n\}$ or $\{v_1, \ldots, v_n\}$ used in (1.32) or (1.33), for a given value of n, is made in an *a priori* manner, if possible using the available information on the problem. In the latter, the computations executed for lower values of n are exploited in order to monitor the choice at stage n. One desirable feature of an adaptive algorithm is that it should be *incremental* or *greedy*: only one new function ϕ_n or v_n is added at each stage to the $n-1$ previously selected functions which are left unchanged. Adaptive algorithms are known to often perform better than their non-adaptive counterparts, but their convergence analysis is usually more delicate.

A second important distinction between the various numerical methods is whether they are *non-intrusive* or *intrusive*. A non-intrusive algorithm builds on an existing exact or approximate solver for the PDE, which may

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

be computationally expensive. It derives approximations of the form (1.32) or (1.33) by choosing instances $a^1, \dots, a^n \in \mathcal{A}$ or $y^1, \dots, y^n \in U_{\mathcal{A}}$ and using the values $u(a^i)$ or $u(y^i)$ computed by the solver. Non-intrusive algorithms may be implemented even when this solver is a black box, and therefore with possibly limited knowledge of the exact PDE model. An intrusive algorithm, on the other hand, directly exploits the precise form of the PDE for computing the approximation (1.32) or (1.33), and therefore requires full knowledge of the PDE model for its implementation.

It should be noted that instances $u(a^i)$ as well as the functions $\{v_1, \dots, v_n\}$ used in (1.32) or (1.33) can only be computed with a certain level of spatial discretization, for example resulting from a finite element solver. In such a case they belong to a finite-dimensional space $V_h \subset V$. If the same finite element space V_h is used to discretize all instances $u(a)$ up to a precision that is satisfactory for the user, this means that we are actually trying to capture the *approximate solution maps*

$$a \mapsto u_h(a) \in V_h, \tag{1.57}$$

or

$$y \mapsto u_h(y) \in V_h. \tag{1.58}$$

The analysis of the performance of numerical reduced modelling methods needs to incorporate the additional error produced by this discretization.

1.7. Sparse polynomial approximation algorithms

One first class of methods that we analyse consists in finding a numerically computable polynomial approximation of the form (1.50). There are two major tasks in constructing such a numerical approximation: (i) find good truncation sets $(\Lambda_n)_{n\geq 1}$, and (ii) numerically compute an approximation to the coefficients v_ν for each $\nu \in \Lambda_n$. By far the most significant issue in numerical methods based on polynomial expansions is to find a good choice for the sets $(\Lambda_n)_{n\geq 1}$. If everything was known to us, we would simply take for Λ_n the set of indices ν corresponding to the n largest V-norms of the coefficients in (1.51) or (1.52). However, finding such an optimal Λ_n would require in principle that we compute the coefficients for all values $\nu \in \mathcal{F}$, which is obviously out of reach. In addition, the structure of the optimal set Λ_n can be quite complicated. One saving grace is that our analysis in Section 3 gives *a priori* bounds on the size of these coefficients. These bounds can be used in order to make an *a priori* selection of the sets Λ_n. This is a non-adaptive approach, which generally gives sub-optimal performance due to the possible lack of sharpness in the *a priori* bounds. This leads one to try to enhance performance by combining the *a priori* bounds together with an adaptive selection of the sets $(\Lambda_n)_{n\geq 1}$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The numerical methods are facilitated by imposing that the selected index sets Λ_n are *downward closed* (or *lower sets*), that is, they satisfy the following property:

$$\nu \in \Lambda_n \quad \text{and} \quad \tilde{\nu} \leq \nu \Rightarrow \tilde{\nu} \in \Lambda_n, \tag{1.59}$$

where $\tilde{\nu} \leq \nu$ means that $\tilde{\nu}_j \leq \nu_j$ for all j.

In Section 6 we discuss algorithms which compute the polynomial approximation by an interpolation process. These algorithms are non-intrusive and apply to a broad scope of problems. One key issue is the choice of interpolation points, which is facilitated by the following result: if $\{z_k\}_{k\geq 0}$ is a sequence of pairwise distinct points in $[-1,1]$ and if $z_\nu := (z_{\nu_j})_{j\geq 1} \in U$, then for any downward closed set Λ_n, any polynomial of the form (1.50) is uniquely characterized by its value on the grid $\{z_\nu : \nu \in \Lambda_n\}$. This allows us to construct the interpolation in a hierarchical manner, by simultaneously incrementing the polynomial space and the interpolation grid. The sets Λ_n can either be *a priori* chosen based on the bounds on the coefficients established in Section 3 or adaptively generated. These sets generally differ from the ideal sets corresponding to the n largest coefficients (which may not fulfil the downward closedness property). We show that certain choices of the univariate sequence $\{z_k\}_{k\geq 0}$, known as Leja or $\mathfrak{R}$-Leja points, lead to stable interpolation processes (in the sense that the Lebesgue constant has moderate growth with the number of points), allowing us to retrieve by interpolation the same algebraic convergence rate n^{-s} proved for the best polynomial approximations.

In Section 7 we discuss another class of algorithms which recursively compute the exact coefficients in the Taylor series of the approximate solution map (1.58). These algorithms are intrusive and they only apply to problems where $\mathcal{P}$ is linear both in u and a up to a constant term, such as the elliptic equation (1.5), which serves a guiding example. The recursive computation is facilitated by imposing that the index sets Λ_n in the truncated Taylor expansion are downward closed. Similar to the interpolation algorithm from Section 6, the sets Λ_n can either be *a priori* chosen, based on the available bounds on the coefficients established in Section 3, or adaptively generated. One main result shows that adaptive algorithms based on a so-called *bulk chasing* procedure have the same convergence rate n^{-s} as the one that is established when using the index sets corresponding to the n largest Taylor coefficients.

1.8. Reduced basis algorithms

The second class of methods that we analyse seeks to find, in an *offline* stage, a set of functions $\{v_1, \ldots, v_n\}$ for which the resulting n-dimensional space $V_n := \operatorname{span}\{v_1, \ldots, v_n\}$ is close to an optimal linear n-dimensional approximation space. For mean-square estimates, one approach, known as *proper*

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

orthogonal decomposition, consists in building the functions $\{v_1, \dots, v_n\}$ based on an approximation of the exact covariance operator (1.44) computed from a sufficiently dense sampling of the random solution $u(a)$. Another approach, which targets uniform estimates, is the *reduced basis method*, which consists in generating V_n by a selection of n particular solution instances $\{u(a^1), \dots, u(a^n)\}$ chosen from a very large set of potential candidates. In both cases, the offline stage is potentially very costly.

Once such a space V_n is chosen, one builds an *online* solver, such that for any given $a \in \mathcal{A}$, the approximate solution $u_n(a)$ is an element from V_n. There are several possible ways to build this online solver. The most prominent of these is to take the Galerkin projection of $u(a)$ onto V_n, which consists of finding $u_n(a)$ by solving the system of equations

$$\langle \mathcal{P}(u_n(a), a), w \rangle = 0, \quad w \in V_n, \tag{1.60}$$

for a suitable duality product $\langle \cdot, \cdot \rangle$. This online computation determines the values $\phi_i(a)$ for each a in (1.32), for $i = 1, \dots, n$. The advantage of using the Galerkin solver, is that, for certain problems such as elliptic ones, it is known to give the best error in approximating $u(a)$ by elements of V_n when error is measured in the norm of V. Its disadvantage lies in its computational cost in finding $u_n(a)$ given the query a. For this reason other projections onto V_n are also studied, some of which are based on interpolation.

The key issue when using reduced basis methods is how to find a good space V_n, that is, how to find good basis functions. In Section 8 we discuss an elementary greedy strategy for the offline selection of the instances $v_i = u(a^i)$, that consists in picking the nth instance which deviates the most from the space V_{n-1} generated from the $n-1$ previously selected instances. The approximation error

$$\sigma_n(\mathcal{M}) := \sup_{v \in \mathcal{M}} \inf_{w \in V_n} \|v - w\| \tag{1.61}$$

produced by such spaces may be significantly larger than the ideal benchmark of the n-width of the solution manifold for a given value of n. However, a striking result is that both are comparable in terms of rate of decay: for any $s > 0$, there is a constant C_s such that

$$\sup_{n \geq 1} n^s \sigma_n(\mathcal{M}) \leq C_s \sup_{n \geq 1} n^s d_n(\mathcal{M})_V. \tag{1.62}$$

Similar results are established for exponential convergence rates.

While both classes of numerical methods aim to construct separable approximations of the form (1.32) or (1.33), there is a significant distinction between them in the way they organize computation. For the first class of polynomial approximation methods, the offline stage fixes the polynomial functions ϕ_i through the selection of the set Λ_n and computes the coefficients v_i. Then the online stage is in some sense trivial since it simply

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

computes $u_n(y)$ through the linear combination (1.50). For the second class, the online stage still requires solution of PDEs approximately in the chosen reduced space V_n. This offline/online splitting makes it difficult to draw a fair comparison between the different methods from the point of view of computational time versus accuracy.

Notation for constants. Numerous multiplicative constants appear in this paper, for example in convergence estimates. We use the generic notation C, which may therefore change value between different formulas, and if necessary we indicate the parameters on which C depends. We use a more specific notation if we want to express the dependence of the constant with respect to a certain parameter (for example the dimension d in (1.54)) or if we want to refer to this specific constant later in the paper.

1.9. Historical orientation

Numerical methods for parametric and stochastic PDEs using polynomials (or other approximation tools) in the parametric variable have been widely studied since the 1990s. We refer in particular to Ghanem and Spanos (1991), Gunzburger, Webster and Zhang (2014), Kleiber and Hien (1992), Knio and Le Maitre (2010) and Xiu (2010) for general introductions to these approaches, and to Andreev, Bieri and Schwab (2006), Ghanem and Spanos (1997), Frauenfelder, Schwab and Todor (2005), Karniadakis and Xiu (2002*a*, 2002*b*) and Schwab and Todor (2003, 2007) for related work prior to the results described in our paper.

The approximation results presented in Section 3 were obtained by the authors and their co-authors in a series of papers (Cohen, DeVore and Schwab 2010, Cohen, DeVore and Schwab 2011, Chkifa, Cohen and Schwab 2015*b*), and the results in Section 4 on the evaluation of n-widths are from Cohen and DeVore (2015). These results establish for the first time convergence rates immune to the curse of dimensionality, in the sense that they hold with infinitely many variables; see also Gittelson and Schwab (2011) for a survey dealing with these issues. In a similar infinite-dimensional framework, and not covered in our paper, let us mention the following related works: (i) similar holomorphy and approximation results are established in Hansen and Schwab (2013*a*, 2013*b*) and Kunoth and Schwab (2013) for specific types of PDEs and control problems; (ii) the approximation of integrals by quadratures is discussed in Kuo, Schwab and Sloan (2012, 2015); (iii) inverse problems are discussed in Schillings and Schwab (2013, 2014) and Schwab and Stuart (2012), following the Bayesian perspective from Stuart (2010); (iv) diffusion problems with lognormal coefficients are treated in Gittelson (2010), Hoang and Schwab (2014) and Graham *et al.* (2015).

The sparse interpolation method presented in Section 6 is introduced and studied in Chkifa, Cohen and Schwab (2014). See also Babuška, Nobile

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and Tempone (2007), Beck, Nobile, Tamellini and Tempone (2012) and Nobile, Tempone and Webster and (2008*a*, 2008*b*) for related work on collocation methods. Other non-intrusive methods are based on least-squares regression as discussed in Chkifa *et al.* (2015*a*), Doostan and Iaccarino (2009), Doostan and Owadi (2011) and Migliorati, Nobile, Tempone and Von Schwerin (2013), or on pseudo-spectral approximations as discussed in Xiu (2007) and Constantine, Eldred and Phipps (2012). The Taylor approximation algorithm presented in Section 7 is introduced and studied in Chkifa, Cohen, DeVore and Schwab (2013). Other intrusive methods based on Galerkin projection are discussed in Beck *et al.* (2012), Beck, Nobile, Tamellini and Tempone (2014), Cohen *et al.* (2010) and Gittelson (2013).

Reduced basis methods have been studied since the 1980s (Noor and Peters 1980). The greedy algorithms presented in Section 8 were introduced and discussed in Veroy, Prud'homme, Rovas and Patera (2003), Machiels *et al.* (2000), Maday, Patera and Turinici (2002*a*, 2002*b*) and Rozza, Huynh and Patera (2008), and their convergence analysis was given in Binev *et al.* (2011) and DeVore, Petrova and Wojtaszczyk (2013). We refer to Kahlbacher and Volkwein (2007) for a general introduction to the related POD method, which is not discussed in our paper.

PART I
Smoothness and approximation results

2. Holomorphic extensions

In this section we discuss smoothness properties of the solution maps $a \mapsto u(a)$ and $y \mapsto u(y)$, which are central to the development of efficient numerical methods that are immune to the curse of dimensionality. We show that, under suitable assumptions on the parametric PDE, these maps admit holomorphic extensions to certain complex domains. Recall that a map F from a complex Banach space X to a second complex Banach space Y is said to be holomorphic on an open set $\mathcal{D} \subset X$ if, for each $x \in \mathcal{D}$, F has a Fréchet derivative $\mathrm{d}F(x)$ at x. Here $\mathrm{d}F(x)$ is a linear operator mapping X to Y such that

$$\|F(x+h) - F(x) - \mathrm{d}F(x)h\|_Y = o(\|h\|_X), \quad h \in X. \tag{2.1}$$

2.1. Extension of $a \mapsto u(a)$ for the model elliptic equation

In order to formulate results on holomorphy, we need to introduce existence–uniqueness theory for solutions to (1.1) in the case when a is complex-valued.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We begin with our guiding example of the elliptic equation (1.5). We now consider

$$(X, V, W) = (L^\infty(D), H_0^1(D), H^{-1}(D)) \tag{2.2}$$

as spaces of complex-valued functions, and extend the standard variational formulation to such spaces in a straightforward manner. For a given $a \in X$, and with $f \in W := V'$, find $u = u(a) \in V$ such that

$$\int_D a\nabla u \cdot \nabla v = \langle f, v\rangle_{W,V}, \quad v \in V, \tag{2.3}$$

where in the left integrand

$$\nabla u \cdot \nabla v := \sum_{i=1}^{m} \partial_{x_i} u \overline{\partial_{x_i} v} \tag{2.4}$$

is the standard Hilbertian inner product, and $\langle f, v\rangle$ is the anti-duality pairing between W and V, which when $f \in L^2(D)$ is given by the Hilbertian inner product

$$\langle f, v\rangle = \int_D f\overline{v}. \tag{2.5}$$

We recall that

$$\|v\|_V := \|\nabla v\|_{L^2(D)}, \tag{2.6}$$

and

$$\|v\|_W := \sup\{\langle v, w\rangle : \|w\|_V \le 1\}. \tag{2.7}$$

This is therefore a particular case of a linear problem with the following general variational formulation. Let $\mathfrak{B}$ denote the set of all sesquilinear forms defined on $V \times V$ and let $W = V^*$ be the set of all antilinear functionals defined on V, that is, W is the anti-dual of V. We define the following norm on $\mathfrak{B}$:

$$\|B\| := \sup_{\|v\|_V \le 1,\ \|w\|_V \le 1} |B(v, w)|. \tag{2.8}$$

Problem. Given $B \in \mathfrak{B}$ and $L \in W$, find $u \in V$ such that

$$B(u, v) = L(v), \quad \text{for all } v \in V. \tag{2.9}$$

The existence–uniqueness theory for such problems can be proved from the complex version of the Lax–Milgram theorem given in Theorem 2.1. This theorem is a particular case of Theorem 2.2 proved below. To formulate this theorem and for later use, we introduce the notation $\mathcal{L}(X, Y)$ for the space of all linear operators T mapping the Banach space X into the Banach space Y with its usual norm

$$\|T\|_{\mathcal{L}(X,Y)} := \sup_{x \in X} \frac{\|Tx\|_Y}{\|x\|_X}. \tag{2.10}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Given $B \in \mathfrak{B}$, we have that $B(u,\cdot)$ is an antilinear functional and hence, for any $u \in V$, there is a $\mathcal{B}u \in W$ such that

$$B(u,v) = \langle \mathcal{B}u, v\rangle_{W,V}, \quad v \in V, \tag{2.11}$$

where $\langle\cdot,\cdot\rangle_{W,V}$ is the anti-duality pairing between W and V. Therefore, $\mathcal{B}$ is a linear operator from V into W and its norm is the same as that of B:

$$\|\mathcal{B}\|_{\mathcal{L}(V,W)} = \|B\|. \tag{2.12}$$

So the operator $\mathcal{B}$ is bounded and hence continuous. Problem (2.9) is equivalent to the equation

$$\mathcal{B}u = L, \tag{2.13}$$

set in W. With this notation and remarks in hand, we can now state the complex version of the Lax–Milgram theorem.

Theorem 2.1. Assume that $B \in \mathfrak{B}$ is a sesquilinear form on $V \times V$ such that

$$|B(u,u)| \geq \alpha \|u\|_V^2, \quad u \in V, \tag{2.14}$$

for some $\alpha > 0$. Then, $\mathcal{B}$ is invertible and its inverse satisfies

$$\|\mathcal{B}^{-1}\|_{\mathcal{L}(W,V)} \leq \frac{1}{\alpha}. \tag{2.15}$$

Thus, for each $L \in W$, problem (2.9) has a unique solution $u_L = \mathcal{B}^{-1}(L)$ which satisfies the *a priori* estimate

$$\|u_L\|_V \leq \frac{\|L\|_W}{\alpha}. \tag{2.16}$$

For the particular B and L given by the left- and right-hand sides of (2.3), the ellipticity condition (2.14) holds with $\alpha = r$ under the assumption

$$\Re(a(x)) \geq r, \quad x \in D, \tag{2.17}$$

since the latter implies, for all $v \in V$,

$$|B(v,v)| \geq \Re(B(v,v)) = \int_D \Re(a)|\nabla v|^2 \geq r\|\nabla v\|_{L^2}^2 = r\|v\|_V^2. \tag{2.18}$$

Therefore, for any $r > 0$ we may extend the solution map $a \mapsto u(a)$ of the elliptic problem (1.5) to the complex domain

$$\mathcal{D}_r := \{a \in X : \Re(a) \geq r\}, \tag{2.19}$$

with the uniform bound

$$\|u\|_{L^\infty(\mathcal{D}_r,V)} = \sup\{\|u(a)\|_V : a \in \mathcal{D}_r\} \leq \frac{\|f\|_W}{r}. \tag{2.20}$$

This extension is therefore defined on the open set $\mathcal{D} := \cup_{r>0}\mathcal{D}_r$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The fact that this map is holomorphic immediately follows by viewing it as a chain of holomorphic maps: introducing for any a the operator $\mathcal{B}(a) : v \mapsto -\mathrm{div}(a\nabla v)$ acting from V into W, we can decompose $a \mapsto u(a)$ into the chain of maps

$$a \mapsto \mathcal{B}(a) \mapsto \mathcal{B}(a)^{-1} \mapsto \mathcal{B}(a)^{-1} f = u(a). \tag{2.21}$$

The first and third maps are continuous linear and therefore holomorphic, from X into $\mathcal{L}(V,W)$ and from $\mathcal{L}(W,V)$ onto V respectively. The second map is the operator inversion, which is holomorphic at any invertible $\mathcal{B} \in \mathcal{L}(V,W)$.

For further purposes, it is interesting to compute the Fréchet complex derivative $\mathrm{d}u(a) \in \mathcal{L}(X,V)$ for the elliptic problem (1.5). Fix an $a \in \mathcal{D}_r$ and let $h \in X$ be such that $\|h\|_X \leq r/2$. Then, the solution map is also defined at $a + h \in \mathcal{D}_{r/2}$. Subtracting the variational formulations (2.3) for $u(a+h)$ and $u(a)$, we find that

$$\int_D a\nabla(u(a+h) - u(a)) \cdot \nabla v = -\int_D h\nabla u(a+h) \cdot \nabla v. \tag{2.22}$$

We first use this identity to obtain a Lipschitz-continuity bound: by taking $v = u(a+h) - u(a)$ and taking the real part of both sides, we find that

$$\begin{aligned} r\|u(a+h) - u(a)\|_V^2 &\leq \|h\|_X \|u(a+h)\|_V \|u(a+h) - u(a)\|_V \\ &\leq \frac{2\|f\|_W}{r} \|h\|_X \|u(a+h) - u(a)\|_V, \end{aligned}$$

and therefore

$$\|u(a+h) - u(a)\|_V \leq C\|h\|_X, \quad C := \frac{2\|f\|_W}{r^2}. \tag{2.23}$$

We next show that $\mathrm{d}u(a)h$ can be defined as the solution $w = w(h) \in V$ to the problem

$$\int_D a\nabla w \cdot \nabla v = -\int_D h\nabla u(a) \cdot \nabla v, \quad v \in V, \tag{2.24}$$

which is well-posed in the sense of the above Lax–Milgram theorem. Indeed, on the one hand, the dependence of w on h is linear and it is continuous, because taking $v = w$ we find that

$$\|w\|_V \leq C\|h\|_X, \quad C := \frac{\|u(a)\|_V}{r}. \tag{2.25}$$

On the other hand, the remainder $g = u(a+h) - u(a) - w$ is the solution to

$$\int_D a\nabla g \cdot \nabla v = \int_D h(\nabla u(a) - \nabla u(a+h)) \cdot \nabla v, \quad v \in V, \tag{2.26}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

which by taking $v = g$ and using (2.23) gives the quadratic bound

$$\|g\|_V \leq \frac{1}{r}\|h\|_X\|u(a) - u(a+h)\|_V \leq C\|h\|_X^2, \quad C := \frac{2\|f\|_W}{r^3}. \tag{2.27}$$

This confirms that $\mathrm{d}u(a)h = w(h)$.

2.2. Extensions by the Ladyzhenskaya–Babuška–Brezzi theory

Our next goal is to treat more general linear parametric problems that are not necessarily elliptic. In particular, we have in mind parabolic problems such as the heat equation, or saddle-point problems such as the Stokes equations. In order to formulate this general class of problems, we suppose that V and $\tilde{V}$ are two complex Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_{\tilde{V}}$, respectively. So, for example, for every $v \in V$, $\langle v,\cdot\rangle_V$ is an antilinear functional on V and $\langle\cdot, v\rangle_V$ is a linear functional on V, and the same holds for $\tilde{V}$. We let $W := \tilde{V}^*$ denote the space of all antilinear functionals on $\tilde{V}$, that is, W is the anti-dual space of $\tilde{V}$.

We now denote by $\mathfrak{B} = \mathfrak{B}(V, \tilde{V})$ the set of all such sesquilinear forms on $V \times \tilde{V}$ and introduce the following topology on $\mathfrak{B}$:

$$\|B\| = \max_{\|v\|_V=1, \|w\|_{\tilde{V}}=1} |B(v, w)|. \tag{2.28}$$

As in the previous subsection, if $B \in \mathfrak{B}$ then $B(u,\cdot)$ is an antilinear functional on $\tilde{V}$. Therefore, as before, we can define the linear operator $\mathcal{B}$ that maps V into W by (2.11), and we again have $\|\mathcal{B}\|_{\mathcal{L}(V,W)} = \|B\|$.

We consider the following general problem.

Problem. Given $B \in \mathfrak{B}(V, \tilde{V})$ and $L \in W$, find $u \in V$ such that

$$B(u, v) = L(v), \quad \text{for all } v \in \tilde{V}, \tag{2.29}$$

This problem is again equivalent to the equation $\mathcal{B}u = L$. In order to establish existence and uniqueness of solutions for complex formulations of such problems, we use the following complex-valued version of the Ladyzhenskaya–Babuška–Brezzi theorem.

Theorem 2.2. Assume that $B \in \mathfrak{B}(V, \tilde{V})$ satisfies

$$\inf_{u\in V}\sup_{v\in\tilde{V}} \frac{|B(u,v)|}{\|u\|_V\|v\|_{\tilde{V}}} \geq \alpha \quad \text{and} \quad \inf_{v\in\tilde{V}}\sup_{u\in V} \frac{|B(u,v)|}{\|u\|_V\|v\|_{\tilde{V}}} \geq \alpha, \tag{2.30}$$

for some $\alpha > 0$. Then, the operator $\mathcal{B}$ defined via (2.11) is invertible and its inverse satisfies

$$\|\mathcal{B}^{-1}\|_{\mathcal{L}(W,V)} \leq \frac{1}{\alpha}. \tag{2.31}$$

Hence, for each $L \in W$, problem (2.29) has a unique solution $u_L = \mathcal{B}^{-1}(L)$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

which satisfies the *a priori* estimate

$$\|u_L\|_V \leq \frac{\|L\|_W}{\alpha}. \tag{2.32}$$

Proof. From the first inf-sup condition in (2.30) we obtain that

$$\alpha\|u\|_V \leq \|\mathcal{B}u\|_W, \quad u \in V. \tag{2.33}$$

This shows that $\mathcal{B}$ is injective and that its range $\mathcal{B}(V)$ is closed in W. In order to prove that $\mathcal{B}$ is invertible we need to show that $\mathcal{B}(V)$ is all of W. We prove it by contradiction: if $\mathcal{B}(V)$ was strictly contained in W, we can pick a non-trivial $w \in W$ which is orthogonal in W to all elements of $\mathcal{B}(V)$. Then, we define $v = v(w)$ in the anti-dual of W, that is, $v \in \tilde{V}$ by setting

$$v : e \mapsto \overline{\langle e, w\rangle}_W = \langle e, v\rangle_{W,\tilde{V}}. \tag{2.34}$$

It follows that

$$B(u, v) = \langle \mathcal{B}u, v\rangle_{W,\tilde{V}} = v(\mathcal{B}u) = \overline{\langle \mathcal{B}u, w\rangle}_W = 0, \tag{2.35}$$

for all $u \in V$. This contradicts the second inf-sup condition. Hence $\mathcal{B}$ is invertible and the bound (2.31) on its inverse follows from (2.33). □

Remark 2.3. One particular case of Theorem 2.2 is Theorem 2.1, since (2.14) implies (2.30) in the case $\tilde{V} = V$.

The argument, centering on (2.21), which justified the holomorphy of the solution map for our canonical elliptic setting, may be generalized to any linear problem of the form

$$\mathcal{B}(a)u = f(a), \tag{2.36}$$

where $a \mapsto \mathcal{B}(a)$ and $a \mapsto f(a)$ are holomorphic maps from an open set $\mathcal{D} \subset X$ into $\mathcal{L}(V, W)$ and W respectively, where (X, V, W) are complex Banach spaces. Namely, if $\mathcal{B}(a)$ is invertible for all $a \in \mathcal{D}$, we find by

$$a \mapsto \mathcal{B}(a) \mapsto \mathcal{B}(a)^{-1} \mapsto \mathcal{B}(a)^{-1} f(a) = u(a) \tag{2.37}$$

that the solution map is holomorphic over $\mathcal{D}$.

In particular, we may consider the following parametric linear problems of the form (2.29) for a pair of Hilbert spaces $(V, \tilde{V})$: for a given $a \in X$, find $u(a) \in V$ such that

$$B(u(a), v; a) = L(v; a), \quad v \in \tilde{V}, \tag{2.38}$$

where $B(\cdot, \cdot; a)$ and $L(\cdot; a)$ are continuous sesquilinear and antilinear forms over $V \times \tilde{V}$ and $\tilde{V}$ respectively, which depend on $a \in X$.

Corollary 2.4. If the assumptions of Theorem 2.2 are satisfied for the problem (2.38) for each a in a set $\mathcal{D} \subset X$, then the operator $\mathcal{B}(a)$ defined by $B(u, v; a) = \langle \mathcal{B}(a)u, v\rangle_{W,\tilde{V}}$ is well defined and invertible from V to $W = \tilde{V}^*$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

for all $a \in \mathcal{D}$ and the solution map $a \mapsto u(a)$ is well defined from $\mathcal{D}$ into V. If the constant $\alpha > 0$ can be chosen independent of $a \in \mathcal{D}$ and if $\sup_{a\in\mathcal{D}} \|L(\cdot, a)\|_W =: M < \infty$, then the solution map is uniformly bounded on $\mathcal{D}$ with

$$\|u\|_{L^\infty(\mathcal{D},V)} = \sup_{a\in\mathcal{D}} \|u(a)\|_V \le \frac{M}{\alpha}. \tag{2.39}$$

In addition, if $\mathcal{D}$ is an open set and if the maps $a \mapsto L(\cdot; a)$ and $a \mapsto B(\cdot, \cdot; a)$ are holomorphic from $\mathcal{D}$ into W and $\mathcal{D}$ into $\mathfrak{B} = \mathfrak{B}(V, \tilde{V})$, respectively, then the solution map $a \mapsto u(a)$ is holomorphic over $\mathcal{D}$.

We next give two examples that fall into this general framework. The first one is a linear parabolic equation with parametrized evolution operator. As a simple model, we consider the heat equation parametrized by its diffusion coefficient,

$$\partial_t u = \operatorname{div}(a\nabla u) + f, \quad \text{in }]0, T[\times D, \tag{2.40}$$

where $D \subset \mathbb{R}^m$ is a Lipschitz domain, $f \in L^2(]0, T[; H^{-1}(D))$, and the initial and boundary value conditions are

$$u_{|t=0} = u_0 \in L^2(D) \quad \text{and} \quad u_{|\partial D} = 0. \tag{2.41}$$

It is well known that a solution space for this PDE is

$$V := L^2(]0, T[; H^1_0(D)) \cap H^1(]0, T[; H^{-1}(D)). \tag{2.42}$$

We obtain a space–time variational formulation of the type (2.38) by introducing the auxiliary space

$$\tilde{V} := L^2(]0, T[; H^1_0(D)) \times L^2(D), \tag{2.43}$$

and defining for $a \in X := L^\infty$, $u \in V$ and $v = (v_1, v_2) \in \tilde{V}$,

$$\begin{aligned} B(u, v; a) := &\int_0^T \int_D \big(\partial_t u(x,t)\overline{v_1(x,t)} + a\nabla u(x,t) \cdot \nabla v_1(x,t)\big)\, \mathrm{d}x\, \mathrm{d}t \\ &+ \int_D u(\cdot, 0)\overline{v_2(x)}\, \mathrm{d}x, \end{aligned} \tag{2.44}$$

and

$$L(v; a) := \int_0^T \langle f(\cdot, t), v_1(\cdot, t)\rangle\, \mathrm{d}t + \int_D u_0(x)\overline{v_2(x)}\, \mathrm{d}x, \tag{2.45}$$

where $\langle \cdot, \cdot \rangle$ is the anti-duality pairing between $H^{-1}(D)$ and $H^1_0(D)$.

The fact that these are bounded sesquilinear and antilinear forms follows readily from the choice of spaces X, V and $\tilde{V}$. By using the general arguments from Schwab and Stevenson (2009), one can show that whenever the diffusion coefficient comes from the uniform ellipticity class $\mathcal{D}_r$ of (2.17), then the inf-sup condition (2.30) holds, with the values of α in

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

(2.30) depending on that of r. Therefore, from Theorem 2.2, the solution map $a \mapsto u(a)$ is defined on $\mathcal{D}_r$ with a uniform bound

$$\|u\|_{L^\infty(\mathcal{D}_r,V)} = \sup\{\|u(a)\|_V : a \in \mathcal{D}_r\} \le C_r. \tag{2.46}$$

Since $r > 0$ is arbitrary, the solution map is therefore defined on the open set $\mathcal{D} := \cup_{r>0}\mathcal{D}_r$, and its holomorphy follows from Corollary 2.4 since the sesquilinear form $B(\cdot,\cdot;a)$ depends on a in an affine manner.

The second example is a linear elliptic PDE parametrized by the shape of the physical domain. As a simple model we consider the Laplace equation

$$-\Delta w = 1, \tag{2.47}$$

set on a domain $D_a \subset \mathbb{R}^2$ with homogeneous Dirichlet boundary conditions $w_{|\partial D_a} = 0$. Here a describes the shape of the domain D_a in polar coordinates, according to

$$D_a := \{x = (\rho\cos\theta, \rho\sin\theta) : 0 \le \rho < a(\theta)\}. \tag{2.48}$$

In order to obtain a Lipschitz domain, we take $a \in X$ with

$$X := W^1_{\mathrm{per}}(L^\infty([0,2\pi[), \tag{2.49}$$

the space of 2π-periodic Lipschitz-continuous functions, which is equipped with the norm

$$\|a\|_X := \|a\|_{L^\infty([0,2\pi[)} + \|a'\|_{L^\infty([0,2\pi[)}. \tag{2.50}$$

If, in addition, for some $r > 0$ we have

$$a(\theta) \ge r, \quad \theta \in [0,2\pi[, \tag{2.51}$$

then D_a is a Lipschitz domain, and thus there exists a unique solution $w = w(a) \in H^1_0(D_a)$ to (2.47) in the sense of the variational formulation

$$\int_{D_a} \nabla w \cdot \nabla v = \int_{D_a} v, \quad v \in H^1_0(D_a). \tag{2.52}$$

Note that (2.51) implies that D_a is star-shaped with respect to a ball of sufficiently small radius centred at the origin. In order to study the parametric smoothness of $a \mapsto w(a)$, we need to represent the solution in a function space V which does not change with a. One way to do this is to utilize the pullback of the solution to a reference domain D under a suitable transformation F_a which maps D into D_a. A natural choice is to take for D the unit disc of $\mathbb{R}^2$ centred at the origin, and use the transformation

$$F_a(\rho\cos\theta, \rho\sin\theta) := (a(\theta)\rho\cos\theta, a(\theta)\rho\sin\theta). \tag{2.53}$$

For $x = (\rho\cos\theta, \rho\sin\theta)$, the Jacobian matrix of $F_a(x)$ is given by

$$\mathrm{d}F_a(x) = \begin{pmatrix} a(\theta)\cos\theta & a'(\theta)\cos\theta - a(\theta)\sin\theta \\ a(\theta)\sin\theta & a'(\theta)\sin\theta + a(\theta)\cos\theta \end{pmatrix}, \tag{2.54}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and its determinant by

$$J_a(\hat{x}) := a(\theta)^2 \geq r^2 > 0. \tag{2.55}$$

We let

$$u = w \circ F_a \in V := H_0^1(D) \tag{2.56}$$

denote the pullback solution, and study the solution map $a \mapsto u(a)$. Using F_a as a change of variable in (2.52), we find that u satisfies

$$\int_D M_a \nabla u \cdot \nabla v = \int_D J_a v, \quad v \in V, \tag{2.57}$$

that is, the variational formulation of the equation

$$-\operatorname{div}(M_a \nabla u) = J_a, \tag{2.58}$$

set over the domain D with homogeneous Dirichlet boundary conditions, where

$$M_a(x) := J_a(\hat{x})\, \mathrm{d}F_a^{-1}(x)\, \mathrm{d}F_a^{-t}(x) = \begin{pmatrix} 1+b(\theta)^2 & -b(\theta) \\ -b(\theta) & 1 \end{pmatrix}, \quad b(\theta) := \frac{a'(\theta)}{a(\theta)}. \tag{2.59}$$

We define the complex extension as a solution of (2.38), with $V = \tilde{V} = H_0^1(D)$, and the forms B and L given by

$$B(u, v; a) := \int_D M_a \nabla u \cdot \nabla v \quad \text{and} \quad L(v; a) := \int_D J_a \overline{v}. \tag{2.60}$$

In view of the expressions of M_a and J_a, it is readily seen that $a \mapsto L(\cdot; a)$ is holomorphic from X onto $W = V^*$ and that $a \mapsto B(\cdot, \cdot; a)$ is holomorphic from the open set in X of nowhere vanishing functions into $\mathfrak{B} = \mathfrak{B}(V, V)$. It remains to understand for which $a \in X$ the problem has a solution. Introducing the real symmetric matrix $R_a(x) = \Re(M_a(x))$ and denoting by $\lambda_{\min}(a, x)$ its smallest eigenvalue, we have for any $u \in V$

$$\Re(B(u, u; a)) = \int_D R_a \nabla u \cdot \nabla u \geq \lambda_{\min}(a) \|u\|_V^2, \quad \lambda_{\min}(a) := \min_{x \in D} \lambda_{\min}(a, x). \tag{2.61}$$

Therefore the coercivity condition (2.14) holds if $\lambda_{\min}(a) > 0$. A straightforward computation shows that

$$\det(R_a(x)) = 1 - \Im(b(\theta))^2 \quad \text{and} \quad \operatorname{tr}(R_a(x)) = 1 + \Re(b(\theta))^2 - \Im(b(\theta))^2. \tag{2.62}$$

We are thus ensured of the existence of the solution $u(a) \in X$ for those $a \in X$ which are nowhere vanishing and such that

$$|\Im(b(\theta))| < 1, \quad \theta \in [0, 2\pi[, \quad b(\theta) := \frac{a'(\theta)}{a(\theta)}. \tag{2.63}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

By application of Corollary 2.4, the solution map has a holomorphic extension onto the open domain $\mathcal{D} \in X$ consisting of those $a \in X$ which are nowhere vanishing and such that (2.63) holds.

One important observation for this last example is the following: if the parameter domain $\mathcal{A}$ is a compact set of real-valued functions in X such that (2.51) holds for all $a \in \mathcal{A}$, then there exists an open neighbourhood $\mathcal{O}$ of $\mathcal{A}$ in the complex-valued X such that the holomorphic extension of the solution map is uniformly bounded over $\mathcal{O}$. Indeed, in view of the above remarks, for every $a \in \mathcal{A}$, there exists $\varepsilon = \varepsilon(a) > 0$ such that

$$\mathring{B}(a, \varepsilon) := \{\tilde{a} : \|\tilde{a} - a\|_X < \varepsilon\} \subset \mathcal{D}, \tag{2.64}$$

and such that the assumptions of Theorem 2.2 are satisfied for the problem (2.38) with constants α and C_L that are uniform over $\mathring{B}(a, \varepsilon)$. By compactness of $\mathcal{A}$, we may define $\mathcal{O}$ as a finite cover of $\mathcal{A}$ of the form

$$\mathcal{O} = \bigcup_{i=1}^{M} \mathring{B}(a_i, \varepsilon(a_i)), \tag{2.65}$$

for $\{a_1, \ldots, a_M\} \in \mathcal{A}$, so that $a \mapsto u(a)$ is holomorphic and uniformly bounded over $\mathcal{O}$.

2.3. Extensions by the implicit function theorem

In this subsection we consider a further generalization of problems of the form (1.1) for which we can prove holomorphy of the solution map on certain subsets of the complex Banach space X. In particular, this generalization can be applied to certain nonlinear PDEs. As a simple example, to motivate what follows, we consider the nonlinear elliptic equation

$$u^3 - \operatorname{div}(a \nabla u) = f, \tag{2.66}$$

set on a bounded Lipschitz domain $D \subset \mathbb{R}^m$ where $m = 2$ or 3, with homogeneous Dirichlet boundary conditions $u_{|\partial D} = 0$, parametrized by the diffusion coefficient a. Similar to the linear equation (1.5), we set

$$(X, V, W) = (L^\infty(D), H_0^1(D), H^{-1}(D)), \tag{2.67}$$

and consider for $f \in W$ and $a \in X$ the variational formulation

$$\int_D u^3 v + \int_D a \nabla u \nabla v = \langle f, v \rangle_{W,V}, \quad v \in V. \tag{2.68}$$

By the theory of monotone operators (see for example Theorem 1 in Chapter 6 of Runst and Sickel 1996), and using the Sobolev embedding $H_0^1(D) \subset L^4(D)$, one can easily check that for any real-valued $a \in X$ such that $a \geq r$ for some $r > 0$, there exists a unique solution $u(a)$ to (2.68) which satisfies

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

the *a priori* estimate

$$\|u(a)\|_V \leq \frac{\|f\|_W}{r}. \tag{2.69}$$

However, we cannot use monotone operator theory to derive a solution $u(a)$ to (2.68) for complex-valued a since in this form the monotonicity is lost. One could consider an alternative extension of (2.68) to complex-valued functions given by

$$|u|^2 u - \operatorname{div}(a\nabla u) = f. \tag{2.70}$$

For this equation, one can now apply monotone operator theory to the real and imaginary part of the equation in order to show that the problem is well posed when $\Re(a) \geq r$ for some $r > 0$. However, this extension is not holomorphic in the variable a due to the presence of the modulus in (2.70). We thus want to adhere to the original problem (2.66) for complex-valued a, but find an alternative to using monotone operator theory. This alternative is provided by the following general theorem, which is based on the holomorphic version of the implicit function theorem in Banach spaces.

Theorem 2.5. Let $\mathcal{P} : V \times X \to W$, where X, V and W are complex Banach spaces, and let $\mathcal{A} \subset X$ be a compact set such that, for each $a \in \mathcal{A}$, there exists a unique solution $u(a) \in V$ to (1.1). Assume, in addition, that there exists an open set $\mathcal{D}$ of X containing $\mathcal{A}$ for which

(i) $\mathcal{P}$ is a holomorphic map from $V \times \mathcal{D}$ to W,

(ii) for each $a \in \mathcal{A}$, the partial differential $\partial_u \mathcal{P}(u(a), a)$ is an isomorphism from V to W.

Then there exists an open set $\mathcal{O} \subset X$ containing $\mathcal{A}$, such that u has a holomorphic extension to $\mathcal{O}$ which takes values in V and which is uniformly bounded:

$$\|u\|_{L^\infty(\mathcal{O},V)} = \sup_{a\in\mathcal{O}} \|u(a)\|_V < \infty. \tag{2.71}$$

Proof. Let $a \in \mathcal{A}$. Assumptions (i) and (ii) allow us to apply the holomorphic version of the implicit function theorem on complex Banach spaces (see Dieudonné 1969, Theorem 10.2.1), and conclude that there exists an $\varepsilon = \varepsilon(a) > 0$, and a unique holomorphic extension of u from the open ball $\mathring{B}(a,\varepsilon)$ of X, with centre a and radius ε into V, such that $\mathcal{P}(u(b), b) = 0$ for any $b \in \mathring{B}(a,\varepsilon)$. In addition, the map u is uniformly bounded and holomorphic on $\mathring{B}(a,\varepsilon)$ with

$$\mathrm{d}u(b) = -\big(\partial_u \mathcal{P}(u(b), b)\big)^{-1} \circ \partial\mathcal{P}_b(u(b), b), \quad b \in \mathring{B}(a,\varepsilon). \tag{2.72}$$

From the compactness of $\mathcal{A}$, we may define $\mathcal{O}$ as a finite cover of $\mathcal{A}$ of the

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

form

$$\mathcal{O} = \bigcup_{i=1}^{M} \mathring{B}(a_i, \varepsilon(a_i)), \tag{2.73}$$

for $\{a_1, \dots, a_M\} \in \mathcal{A}$. Therefore u has a uniformly bounded holomorphic extension to $\mathcal{O}$. □

There are many settings where Theorem 2.5 can be applied, including nonlinear equations. As an example, we return to (2.66), where the operator $\mathcal{P}$ is given by

$$\mathcal{P}(u, a) = u^3 - \operatorname{div}(a\nabla u) - f, \tag{2.74}$$

or in variational form by

$$\langle \mathcal{P}(u, a), v\rangle_{W,V} = \int_D u^3 v + \int_D a\nabla u \nabla v - \langle f, v\rangle_{W,V}. \tag{2.75}$$

Using the fact that $H^1_0(D)$ is continuously embedded into $L^4(D)$, it is easily seen that $\mathcal{P}$ acts as a holomorphic map from $V \times X$ to W, and therefore assumption (i) holds.

We now take for $\mathcal{A}$ any compact set of X contained in the set of real-valued functions $a \in X$ such that $a \geq r$, where $r > 0$ is fixed, so that there exists a unique solution $u(a) \in V$ for each $a \in \mathcal{A}$. We observe that, for $a \in \mathcal{A}$,

$$\partial_u \mathcal{P}(u(a), a)(w) = 3u(a)^2 w - \operatorname{div}(a\nabla w). \tag{2.76}$$

The operator $\partial_u \mathcal{P}(u(a), a)$ is associated to the sesquilinear form

$$A(v, w; a) = \langle \partial_u \mathcal{P}(u(a), a)(v), w\rangle_{V',V} = \int_D 3u(a)^2 v\overline{w} + \int_D a\nabla v \cdot \overline{\nabla w}, \tag{2.77}$$

which is continuous over $V \times V$ (by the continuous embedding of $H^1_0(D)$ into $L^4(D)$) and satisfies the coercivity condition

$$|A(v, v; a)| \geq \Re(A(v, v; a)) \geq r\|v\|^2_V, \quad v \in V. \tag{2.78}$$

By the complex version of the Lax–Milgram theorem, $\partial_u \mathcal{P}(u(a), a)$ is thus an isomorphism from V onto W, and therefore assumption (ii) holds. We therefore conclude from Theorem 2.5 that there exists an open set $\mathcal{O} \subset X$ containing $\mathcal{A}$, such that the solution map has a uniformly bounded holomorphic extension over $\mathcal{O}$.

Remark 2.6. Theorem 2.5 may also be applied to treat linear parametric problems such as the previously discussed elliptic, parabolic or domain dependent elliptic equations, which are already covered by the LBB theory. Its weakness, however, is that it does not give an explicit description of the domain where the holomorphic extension is defined, in contrast to the explicit

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

conditions on a that can be established for these specific problems using Theorem 2.2. Nevertheless, as will be seen further, the sole existence of a holomorphic extension on an open neighbourhood of the compact parameter domain $\mathcal{A}$ turns out to be sufficient for deriving approximation results for the solution map which are immune to the curse of dimensionality.

2.4. The uniform ellipticity assumption

For the remainder of this section, and all of Section 3, we assume that the parameter space X is the complex Banach space L^∞ and that the parameter set $\mathcal{A} \subset X$ has an affine scalar representation of the form (1.15). This allows us to view the solution map as $y \mapsto u(y) := u(a(y))$. The focus of the remainder of this section is to show that this map has a holomorphic extension to certain complex domains.

Here and in Section 3 we assume that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$, which means that the solution map $a \mapsto u(u)$ is well defined over

$$a(U) := \left\{ a(y) = \overline{a} + \sum_{j\geq 1} y_j \psi_j : y = (y_j)_{j\geq 0} \in U \right\}, \tag{2.79}$$

where $U = [-1,1]^{\mathbb{N}}$, so that the map $y \mapsto u(y)$ is well defined from U to V. We also assume that

$$(\|\psi_j\|_X)_{j\geq 1} \in \ell^1(\mathbb{N}), \tag{2.80}$$

which implies that the series in (1.15) converges absolutely for all $y \in U$. In addition, this assumption guarantees the compactness of the set $a(U)$ defined by (2.79), as shown by the following result.

Lemma 2.7. Under the assumption (2.80), the set $a(U)$ defined by (2.79) is compact in X.

Proof. Let $(a_n)_{n\geq 1}$ be a sequence in $a(U)$. Since $(\|\psi_j\|_X)_{j\geq 1} \in \ell^1(\mathbb{N})$, the sequence $(a_n)_{n\geq 1}$ is bounded in X. Each a_n is of the form

$$a_n = \sum_{j\geq 1} y_{n,j} \psi_j.$$

Using a Cantor diagonal argument, we infer that there exists $y^* = (y_j)_{j\geq 1} \in U$ such that

$$\lim_{n\to+\infty} y_{\sigma(n),j} = y_j^*, \quad j \geq 1, \tag{2.81}$$

where $(\sigma(n))_{n\geq 1}$ is a monotone sequence of positive integers. Defining

$$a^* := \sum_{j\geq 1} y_j^* \psi_j \in a(U),$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

we may write for any $k \geq 1$

$$\|a_{\sigma(n)} - a^*\|_X \leq \left\| \sum_{j=1}^{k} (y_j^* - y_{\sigma(n),j}) \psi_j \right\|_X + 2 \sum_{j \geq k+1} \|\psi_j\|_X. \tag{2.82}$$

It follows that $a_{\sigma(n)}$ converges to a^* in X, and therefore $a(U)$ is compact. □

Let us recall the four previously discussed examples of parametric PDEs, that is, equations (1.5), (2.40), (2.58), and (2.66). For these problems, we have seen that the solution map is defined at any real-valued $a \in X$ that satisfies

$$a(x) \geq r, \quad x \in D, \tag{2.83}$$

for some $r > 0$. Here, the physical domain D is replaced by the angular domain $[0, 2\pi[$ in the case of (2.58). When $\mathcal{A}$ is a compact set of the form (2.79), this condition is met for all $a \in \mathcal{A}$ if and only if

$$a(x, y) = \overline{a}(x) + \sum_{j \geq 1} y_j \psi_j(x) \geq r, \quad x \in D,\ y \in U. \tag{2.84}$$

By taking the particular choice $y_j = -\text{sign}(\psi_j(x))$, we find that the above inequality is equivalent to

$$\sum_{j \geq 1} |\psi_j(x)| \leq \overline{a}(x) - r, \quad x \in D. \tag{2.85}$$

We refer to (2.84) or (2.85) as the *uniform ellipticity assumption* of constant r, or UEA(r). In the case of (2.58), the physical domain D is replaced by the angular domain $[0, 2\pi[$, and UEA(r) thus means that, for all $a \in \mathcal{A}$, the domains D_a are star-shaped with respect to a ball of sufficiently small radius centred at the origin.

Our previous analysis showed that the assumption UEA(r) ensures that in all four examples the solution map $y \mapsto u(y)$ is well defined from U to V. We now want to build an extension $z \mapsto u(z)$ by setting

$$a(z) = \overline{a} + \sum_{j \geq 1} z_j \psi_j \tag{2.86}$$

for suitable $z = (z_j)_{j \geq 1} \in \mathbb{C}^N$, and defining

$$u(z) := u(a(z)). \tag{2.87}$$

This only makes sense for those $z \in \mathbb{C}^{\mathbb{N}}$ for which $a(z)$ is well defined and falls inside the domain $\mathcal{D} \subset X$, where $a \mapsto u(a)$ admits its holomorphic extension. At such a z, the chain rule ensures that the resulting map $z \mapsto u(z)$ is holomorphic in each variable z_j, with partial derivatives given by

$$\partial_{z_j} u(z) = \mathrm{d}u(a(z))\psi_j. \tag{2.88}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Our next objective is to describe some relevant domains of $\mathbb{C}^{\mathbb{N}}$ on which the holomorphic extension $z \mapsto u(z)$ exists and is uniformly bounded.

2.5. Holomorphic extensions of $y \mapsto u(y)$ on polydiscs

We first consider the elliptic problem (1.5) and the parabolic problem (2.40). For such problems, we have seen that the solution map $a \mapsto u(a)$ admits a holomorphic extension on the complex domain $\mathcal{D}_r$ defined by the condition $\Re(a) \geq r$, with uniform bound

$$\|u\|_{L^\infty(\mathcal{D}_r,V)} \leq C_r. \tag{2.89}$$

If UEA(r) holds, then

$$\Re(a(x,z)) = \overline{a}(x) + \sum_{j\geq 1} \Re(z_j)\psi_j(x) \geq r, \quad x \in D, \tag{2.90}$$

holds for all $z \in \mathbb{C}^N$ such that $|\Re(z_j)| \leq 1$, and in particular for all $z \in \mathcal{U}$, where $\mathcal{U}$ is the unit *polydisc*

$$\mathcal{U} := \{z = (z_j)_{j\geq 1} : |z_j| \leq 1\} = \otimes_{j\geq 1}\{|z_j| \leq 1\}. \tag{2.91}$$

This shows that the set

$$a(\mathcal{U}) := \left\{a(z) = \overline{a} + \sum_{j\geq 1} z_j\psi_j : z \in \mathcal{U}\right\} \tag{2.92}$$

is contained in $\mathcal{D}_r$. In turn, the map $z \mapsto u(z)$ is holomorphic in each variable z_j over $\mathcal{U}$ with the uniform bound

$$\sup_{z\in\mathcal{U}} \|u(z)\|_V \leq C_r. \tag{2.93}$$

We next consider general polydiscs of the form

$$\mathcal{U}_\rho := \{z = (z_j)_{j\geq 1} : |z_j| \leq \rho_j\} = \otimes_{j\geq 1}\{|z_j| \leq \rho_j\}, \tag{2.94}$$

where $\rho = (\rho_j)_{j\geq 1}$ is a sequence of positive numbers. Then, for any $t > 0$ and any positive sequence ρ that satisfies the constraint

$$\sum_{j\geq 1} \rho_j|\psi_j(x)| \leq \overline{a}(x) - t, \quad x \in D, \tag{2.95}$$

we find that

$$z \in \mathcal{U}_\rho \quad \Rightarrow \quad \Re(a(x,z)) \geq t, \quad x \in D, \tag{2.96}$$

which shows that $a(\mathcal{U}_\rho) \subset \mathcal{D}_t$. Therefore, the map $z \mapsto u(z)$ is holomorphic in each variable z_j over $\mathcal{U}_\rho$ and the uniform bound (2.93) now holds for some constant $C_t > 0$.

If UEA(r) holds and if $0 < t < r$, we can find sequences ρ which satisfy (2.95) and such that $\rho_j \geq 1$ for all $j \geq 1$, so that the polydisc $\mathcal{U}_\rho$ contains $\mathcal{U}$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

In particular, let $\varepsilon := r - t > 0$ and $\rho = (\rho_j)_{j\geq 1}$ be any sequence of numbers such that $\rho_j \geq 1$ for all $j \geq 1$ and that satisfies the constraint

$$\sum_{j\geq 1}(\rho_j - 1)\|\psi_j\|_X \leq \varepsilon. \tag{2.97}$$

Then, using UEA(r), we have

$$\sum_{j\geq 1}\rho_j|\psi_j(x)| \leq \sum_{j\geq 1}|\psi_j(x)| + \sum_{j\geq 1}(\rho_j - 1)\|\psi_j\|_X \leq \overline{a}(x) - r + \varepsilon = \overline{a}(x) - t \tag{2.98}$$

for $x \in D$. Therefore the map $z \mapsto u(z)$ is holomorphic in each variable z_j over $\mathcal{U}_\rho$ again with a uniform bound

$$\sup_{z\in\mathcal{U}_\rho}\|u(z)\|_V \leq C_t. \tag{2.99}$$

We shall make further use of the following observation: if ρ satisfies one the above constraints (2.95) or (2.97), then for each $j \geq 1$ there is an open set $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ that contains the disc $\{|z_j| \leq \rho_j\}$ and such that the map $z \mapsto u(z)$ is holomorphic in each variable z_j over the tensorized set

$$\mathcal{O}_\rho := \otimes_{j\geq 1}\mathcal{O}_{\rho_j}. \tag{2.100}$$

One possible choice is to take for $\mathcal{O}_{\rho_j}$ the open disc

$$\mathcal{O}_{\rho_j} := \{|z_j| < \tilde{\rho}_j\}, \quad \tilde{\rho}_j := \rho_j + \frac{\tilde{t}}{\sum_{j\geq 1}\|\psi_j\|_X}, \tag{2.101}$$

for some $0 < \tilde{t} < t$, since we then have

$$\sum_{j\geq 1}\tilde{\rho}_j|\psi_j(x)| \leq \sum_{j\geq 1}\rho_j|\psi_j(x)| + \tilde{t} \leq \overline{a}(x) - (t - \tilde{t}), \quad x \in D, \tag{2.102}$$

which shows that $a(\mathcal{O}_\rho) \subset \mathcal{D}_{t-\tilde{t}}$.

In Section 3 we exploit these domains of bounded holomorphy in order to derive convergence results for polynomial approximations of the type (1.50), which are obtained by truncation of the Taylor development of $u(z)$ on suitable sets Λ_n. For now, let us observe that the varying radii ρ_j in each variable of the polydiscs $\mathcal{U}_\rho$ reflect the *anisotropy* of the solution map $z \mapsto u(z)$. Let us also note that the above discussion does not identify one particular polydisc $\mathcal{U}_\rho$. Instead, it shows that bounded holomorphy holds on all of the polydiscs $\mathcal{U}_\rho$ associated to any of the sequences ρ which satisfy the constraint (2.97) or (2.95).

Let us next observe that the above procedure of extending the solution map to polydiscs is not restricted to the problems of the first and second examples. More generally, we can obtain bounded holomorphic extensions

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

on similar polydiscs with constrained radii, for any parametric PDE that satisfies the assumptions of the following theorem.

Theorem 2.8. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. Assume that $(\|\psi_j\|_X)_{j\geq 1} \in \ell^1(\mathbb{N})$ and that the solution map $a \mapsto u(a)$ admits a holomorphic extension to an open set $\mathcal{O} \subset X$ which contains the set $a(\mathcal{U})$ defined by (2.92), with uniform bound

$$\sup_{a\in\mathcal{O}} \|u(a)\|_V \leq C. \tag{2.103}$$

Then there exists $\varepsilon > 0$ such that, for any sequence $\rho = (\rho_j)_{j\geq 1}$ of numbers larger than or equal to 1 which satisfies the constraint (2.97), the following holds. For all $j \geq 1$, there exists an open set $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ that contains the disc $\{|z_j| \leq \rho_j\}$ for which the map $y \mapsto u(y)$ admits an extension to the set $\mathcal{O}_\rho$ defined by (2.100), and this extension is holomorphic in each variable z_j with uniform bound

$$\sup_{z\in\mathcal{O}_\rho} \|u(z)\|_V \leq C, \tag{2.104}$$

with the same value of C as in (2.103)

Proof. We first observe that there exists $\delta > 0$ sufficiently small that the δ-neighbourhood of $a(\mathcal{U})$ is contained in $\mathcal{O}$, that is,

$$\bigcup_{a\in a(\mathcal{U})} B(a,\delta) \subset \mathcal{O}. \tag{2.105}$$

To see this, we observe that, by the same argument as used for $a(U)$ in the proof of Lemma 2.7, the set $a(\mathcal{U})$ is compact. The distance function

$$a \mapsto \operatorname{dist}(a,\mathcal{O}^c) := \inf\{\|a-b\|_X : b \notin \mathcal{O}\} \tag{2.106}$$

is continuous and strictly positive over $a(\mathcal{U})$. By compactness of $a(\mathcal{U})$, it attains a strictly positive minimal value, and therefore (2.105) holds by taking $\delta > 0$ strictly smaller than this minimal value.

Next, we take $\varepsilon > 0$ strictly smaller than δ and define

$$\mathcal{O}_{\rho_j} := \{|z_j| < \tilde{\rho}_j\}, \quad \tilde{\rho}_j := \rho_j + \frac{\delta-\varepsilon}{\sum_{j\geq 1}\|\psi_j\|_X}, \tag{2.107}$$

so that by (2.97) we have

$$\sum_{j\geq 1}(\tilde{\rho}_j - 1)\|\psi_j\|_X = \sum_{j\geq 1}(\rho_j-1)\|\psi_j\|_X + \delta - \varepsilon \leq \delta. \tag{2.108}$$

For any $z \in \mathcal{O}_\rho$, we define $\tilde{z}_j := z_j \min\{1, |z_j|^{-1}\}$, which gives that $\tilde{z} :=$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$(\tilde{z}_j)_{j\geq 1}$ is in $\mathcal{U}$ and

$$\begin{aligned} a(z) &= \overline{a} + \sum_{j\geq 1} z_j \psi_j \\ &= \overline{a} + \sum_{j\geq 1} \tilde{z}_j \psi_j + \sum_{j\geq 1} (z_j - \tilde{z}_j)\psi_j \\ &= a(\tilde{z}) + r(z). \end{aligned}$$

Since

$$\|r(z)\|_X \leq \sum_{j\geq 1} |z_j - \tilde{z}_j| \|\psi_j\|_X \leq \sum_{j\geq 1} (\tilde{\rho}_j - 1)\|\psi_j\|_X \leq \delta, \tag{2.109}$$

it follows from (2.105) that $a(z) \in \mathcal{O}$. Therefore $y \mapsto u(y)$ admits a holomorphic extension over $\mathcal{O}_\rho$ with at least the same uniform bound. □

2.6. Holomorphic extensions of $y \mapsto u(y)$ on polyellipses

The reader should note that the results of the previous subsection on extensions of the solution map to polydiscs were not applied to two of our main examples: the parametrized domain problem (2.58) and the nonlinear problem (2.66). For such problems, in contrast to the elliptic and parabolic problems (1.5) and (2.40), the solution map $a \mapsto u(a)$ might not admit a holomorphic extension to the whole domain $\mathcal{D}_r$ defined by the condition $\Re(a) \geq r$. In turn, while the uniform ellipticity assumption UEA(r) ensures that the map $y \mapsto u(y)$ is well defined over U, it does not allow us to define its holomorphic extension on the polydisc $\mathcal{U}$, or more generally on polydiscs $\mathcal{U}_\rho$ for sequences ρ which satisfy (2.95).

On the other hand, for both problems (2.58) and (2.66), we have seen that if $\mathcal{A}$ is any compact set of real-valued functions in X such that $a \geq r$ for all $a \in \mathcal{A}$, there exists an open set $\mathcal{O} \subset X$ which contains $\mathcal{A}$ and such that the solution map $a \mapsto u(a)$ admits a holomorphic extension on $\mathcal{O}$. In particular, if UEA(r) holds and $(\|\psi_j\|_X)_{j\geq 1} \in \ell^1(\mathbb{N})$, then such an open set exists for $\mathcal{A} = a(U)$ defined by (2.79). This allows us to define a bounded holomorphic extension to $y \mapsto u(y)$ on complex domains that contain U, but with shorter extensions in the imaginary axes than the polydiscs $\mathcal{U}_\rho$. In Section 3 we exploit these domains of bounded holomorphy in order to derive convergence results for polynomial approximations of the type (1.50) that are obtained by truncation of the expansion of $u(z)$ into orthogonal Legendre series.

To formulate the extensions we seek, we introduce some standard concepts from complex analysis. For any real number $s > 1$, we define in $\mathbb{C}$ the so-called *Bernstein ellipse*,

$$\mathcal{E}_s := \left\{ \frac{z + z^{-1}}{2} : |z| = s \right\}, \tag{2.110}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

which has semi-axes of length $(s+s^{-1})/2$ in the real axis and $(s-s^{-1})/2$ in the imaginary axis. Note that in the limit $s \to 1$, we obtain $\mathcal{E}_1 = [-1,1]$. The convex hull of $\mathcal{E}_s$ is given by the filled-in ellipse

$$\mathcal{H}_s := \left\{ \frac{z+z^{-1}}{2} : 1 \le |z| \le s \right\}. \tag{2.111}$$

Note that

$$[-1,1] \subset \mathcal{H}_s \subset \{|z| \le s\}. \tag{2.112}$$

Therefore, defining for any sequence $(\rho_j)_{j\ge 1}$ of numbers strictly larger than 1 the polyellipse

$$\mathcal{E}_\rho := \otimes_{j\ge 1} \mathcal{E}_{\rho_j} \tag{2.113}$$

and the filled-in polyellipse

$$\mathcal{H}_\rho := \otimes_{j\ge 1} \mathcal{H}_{\rho_j}, \tag{2.114}$$

we find that

$$U \subset \mathcal{H}_\rho \subset \mathcal{U}_\rho. \tag{2.115}$$

However, the set $\mathcal{H}_\rho$ has much shorter extension than $\mathcal{U}_\rho$ in the imaginary axis for the coordinates j for which ρ_j is close to 1. This allows us to derive bounded holomorphic extensions on such domains for the solution map $z \mapsto u(z)$ in the case of problems (2.58) and (2.66), and more generally for any parametric PDE satisfying the assumptions of the following result.

Theorem 2.9. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\ge 1}$. Assume that $(\|\psi_j\|_X)_{j\ge 1} \in \ell^1(\mathbb{N})$ and that the solution map $u \mapsto u(a)$ admits a holomorphic extension to an open set $\mathcal{O} \subset X$ which contains the set $a(U)$ defined by (2.79), with uniform bound

$$\sup_{a\in\mathcal{O}} \|u(a)\|_V \le C. \tag{2.116}$$

Then, there exists $\varepsilon > 0$ such that, for any sequence $\rho = (\rho_j)_{j\ge 1}$ of numbers strictly larger than 1 that satisfies the constraint (2.97), the following holds. For all $j \ge 1$, there exists an open set $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ that contains the filled-in ellipse $\mathcal{H}_{\rho_j}$ and such that the map $y \mapsto u(y)$ admits an extension over the set $\mathcal{O}_\rho$ defined by (2.100), which is holomorphic in each variable z_j with uniform bound

$$\sup_{z\in\mathcal{O}_\rho} \|u(z)\|_V \le C, \tag{2.117}$$

with the same value of C as in (2.116).

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Proof. By the same compactness argument as used for $a(\mathcal{U})$ in the proof of Theorem 2.8, there exists $\delta > 0$ sufficiently small that the δ-neighbourhood of $a(U)$ is contained in $\mathcal{O}$, that is,

$$\bigcup_{a \in a(U)} B(a, \delta) \subset \mathcal{O}. \tag{2.118}$$

We now define $\varepsilon = \delta$ and set $\mathcal{O}_{\rho_j}$ to be the oval-shaped domain

$$\mathcal{O}_{\rho_j} := \Big\{ z \in \mathbb{C} : \operatorname{dist}(z, [-1,1]) := \min_{y \in [-1,1]} |z - y| < \rho_j - 1 \Big\}, \tag{2.119}$$

for which it is easily checked that $\mathcal{H}_{\rho_j} \subset \mathcal{O}_{\rho_j}$. For any $z = (z_j)_{j \geq 1} \in \mathcal{O}_\rho$, there exists a $y = (y_j)_{j \geq 1} \in U$ such that

$$|y_j - z_j| \leq \rho_j - 1, \quad j \geq 1. \tag{2.120}$$

We may therefore write

$$a(z) = \overline{a} + \sum_{j \geq 1} z_j \psi_j = a(y) + r(z), \tag{2.121}$$

where

$$a(y) = \overline{a} + \sum_{j \geq 1} y_j \psi_j \in a(U), \tag{2.122}$$

and

$$\|r(z)\|_X \leq \sum_{j \geq 1} |z_j - y_j| \|\psi_j\|_X \leq \sum_{j \geq 1} (\rho_j - 1) \|\psi_j\|_X \leq \varepsilon = \delta. \tag{2.123}$$

It follows from (2.105) that $a(z) \in \mathcal{O}$. Therefore $z \mapsto u(z)$ admits a uniformly bounded extension over $\mathcal{O}_\rho$. □

Remark 2.10. A general setting in which the existence of $\mathcal{O}$ in the above result is satisfied is provided by Theorem 2.5.

Remark 2.11. The assumptions of Theorem 2.9 are obviously weaker than those of Theorem 2.8, since $a(\mathcal{U})$ is replaced by $a(U)$.

Remark 2.12. Theorems 2.8 and 2.9 can be formulated for a general map u from $\mathcal{A}$ to V that is not necessarily the solution map of a parametric PDE. Indeed, the only assumptions on u which are used in the proof of these results is that it admits a bounded holomorphic extension in neighbourhoods of $a(\mathcal{U})$ or $a(U)$. In other words, the same conclusions in these theorems hold for any map u which admits a bounded holomorphic extension on an open set containing $a(\mathcal{U})$ or $a(U)$. On the other hand, as seen in Sections 2.1, 2.2 and 2.3, the fact that u is the solution map to a parametric PDE can be utilized to prove the validity of these assumptions.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

3. Best n-term polynomial approximations

In this section we place ourselves in the same framework as in Section 2.4: we consider a parametric problem (1.1), and assume that Assumption A holds for a suitable affine scalar representation (1.15). The solution map $y \mapsto u(y) := u(a(y))$ is then well defined from U to V. Our goal is now to establish convergence rates for specific separable approximations of this map which are polynomials in the y variable. We construct these approximations by suitable truncations of infinite expansions.

3.1. Approximation by n-term truncated expansions

Let us begin with some general remarks concerning the convergence to u of separable expansions of the form

$$\sum_{\nu\in\mathcal{F}} u_\nu \phi_\nu, \tag{3.1}$$

where ϕ_ν acts from U to $\mathbb{R}$ and $u_\nu \in V$, for some countable index set $\mathcal{F}$.

Definition 3.1. A sequence $(\Lambda_n)_{n\geq 1}$ of finite subsets of $\mathcal{F}$ is called an exhaustion of $\mathcal{F}$ if and only if, for any $\nu \in \mathcal{F}$, there exists n_0 such that $\nu \in \Lambda_n$ for all $n \geq n_0$. Here we do not impose that $\#(\Lambda_n) = n$.

Definition 3.2. The series (3.1) is said to converge *conditionally* with limit u in a given norm $\|\cdot\|$ if and only if there exists an exhaustion $(\Lambda_n)_{n\geq 1}$ of $\mathcal{F}$ such that

$$\lim_{n\to+\infty} \left\| u - \sum_{\nu\in\Lambda_n} u_\nu \phi_\nu \right\| = 0. \tag{3.2}$$

The series (3.1) is said to converge *unconditionally* to u in the same norm, if and only if (3.2) holds for every exhaustion $(\Lambda_n)_{n\geq 1}$ of $\mathcal{F}$.

We are interested both in establishing unconditional convergence and providing estimates for the approximation error. One first instance where this is feasible is when $(\phi_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis, as indicated by the following result, which simply gathers well-known facts from Hilbert space theory.

Theorem 3.3. Let $(\phi_\nu)_{\nu\in\mathcal{F}}$ be an orthonormal basis of $L^2(U,\mu)$ for some given measure μ on U, and let $u \in L^2(U,V,\mu)$. Then, the inner products

$$u_\nu := \int_U u(y)\phi_\nu(y)\,\mathrm{d}\mu(y), \quad \nu \in \mathcal{F}, \tag{3.3}$$

are elements of V, and the series (3.1) converges unconditionally to u in

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$L^2(U,V,\mu)$, with the error given by

$$\left\| u - \sum_{\nu\in\Lambda_n} u_\nu \phi_\nu \right\|_{L^2(U,V,\mu)} = \left(\sum_{\nu\notin\Lambda_n} \|u_\nu\|_V^2 \right)^{1/2}, \tag{3.4}$$

for any exhaustion $(\Lambda_n)_{n\geq 1}$.

Let us observe that if μ is any probability measure, the $L^\infty(U,V)$-norm controls the $L^2(U,V,\mu)$-norm. In the previous section we gave various examples for which u is uniformly bounded over U, and we may therefore apply the above result whenever μ is a probability measure.

We next give a general result which can be used to establish convergence and give error bounds in the L^∞-norms for truncating the expansion (3.1).

Theorem 3.4. Consider an expansion (3.1) for which the following hold:

(i) the functions $\phi_\nu : U \mapsto \mathbb{R}$ are such that $\|\phi_\nu\|_{L^\infty(U)} = 1$, for all $\nu \in \mathcal{F}$,

(ii) the functions $u_\nu \in V$ are such that $(\|u_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^1(\mathcal{F})$.

Then, whenever the expansion (3.1) converges conditionally to a function u in $L^\infty(U,V)$, it also converges unconditionally to u in $L^\infty(U,V)$, and for any exhaustion $(\Lambda_n)_{n\geq 1}$, we have the error estimate

$$\left\| u - \sum_{\nu\in\Lambda_n} u_\nu \phi_\nu \right\|_{L^\infty(U,V)} \leq \sum_{\nu\notin\Lambda_n} \|u_\nu\|_V. \tag{3.5}$$

Proof. Let $(\Lambda_n)_{n\geq 1}$ be any given exhaustion of $\mathcal{F}$ and suppose that $\varepsilon > 0$ is arbitrary. We know that there exists an exhaustion $(\Lambda_n^*)_{n\geq 1}$ and an n_0 such that

$$\left\| u - \sum_{\nu\in\Lambda_n^*} u_\nu \phi_\nu \right\|_{L^\infty(U,V)} \leq \varepsilon, \quad n \geq n_0. \tag{3.6}$$

In addition, there exists m larger than n_0 such that

$$\sum_{\nu\notin\Lambda_m^*} \|u_\nu\|_V \leq \varepsilon. \tag{3.7}$$

Since $(\Lambda_n)_{n\geq 1}$ is an exhaustion, there exists n_1 such that $\Lambda_m^* \subset \Lambda_n$ for all $n \geq n_1$, and therefore

$$\left\| u - \sum_{\nu\in\Lambda_n} u_\nu \phi_\nu \right\|_{L^\infty(U,V)} \leq \left\| u - \sum_{\nu\in\Lambda_m^*} u_\nu \phi_\nu \right\|_{L^\infty(U,V)} + \sum_{\nu\notin\Lambda_m^*} \|u_\nu\|_V \leq 2\varepsilon \tag{3.8}$$

for $n \geq n_1$. This confirms the unconditional convergence. The estimate (3.5) follows by an application of the triangle inequality. □

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

In the particular case where $(\phi_\nu)_{\nu\in\mathcal{F}}$ is an orthogonal basis normalized in L^∞, the next theorem shows that the same result holds without the need to assume conditional convergence.

Theorem 3.5. Let $(\phi_\nu)_{\nu\in\mathcal{F}}$ be an orthogonal basis of $L^2(U,\mu)$ for some given probability measure μ on U, normalized so that $\|\phi_\nu\|_{L^\infty(U)} = 1$ for all $\nu\in\mathcal{F}$. If $u\in L^2(U,V,\mu)$ and the inner products

$$u_\nu := \frac{1}{\|\phi_\nu\|^2_{L^2(U,V,\mu)}}\int_U u(y)\phi_\nu(y)\,\mathrm{d}\mu(y), \quad \nu\in\mathcal{F}, \tag{3.9}$$

satisfy $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}\in\ell^1(\mathcal{F})$, then $u\in L^\infty(U,V)$ and the series (3.1) converges unconditionally to u in $L^\infty(U,V)$ and the estimate (3.5) holds.

Proof. The summability of $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ ensures that (3.1) converges to a limit in $L^\infty(U,V)$ and hence also $L^2(U,V,\mu)$. On the other hand, we know from Theorem 3.3 that it converges to $u\in L^2(U,V,\mu)$. Therefore, its limit in $L^\infty(U,V)$ is also u. □

If the expansion (3.1) converges unconditionally to u in some given norm $\|\cdot\|$, then a crucial issue is the choice of sets Λ_n that we decide to use to truncate (3.1) and define an n-term approximation. Since n measures the complexity of this approximation, we would like to find the set Λ_n which minimizes the truncation error in some given norm $\|\cdot\|$ among all sets of cardinality n, that is,

$$\Lambda_n := \arg\min\left\{\left\|u-\sum_{\nu\in\Lambda}u_\nu\phi_\nu\right\| : \#(\Lambda)=n\right\}, \tag{3.10}$$

provided that such a set exists. This is an instance of *best n-term* approximation, which is itself an instance of *nonlinear approximation*. We refer to DeVore (1998) for a general survey on nonlinear approximation.

In the case where the error is measured in $L^2(U,V,\mu)$, and if $(\phi_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis of $L^2(U,\mu)$ and $(u_\nu)_{\nu\in\mathcal{F}}$ are the coefficients of u in this basis, (3.4) shows that the optimal Λ_n is the set of indices corresponding to the n largest $\|u_\nu\|_V$. Note that such a set is not necessarily unique, in which case any realization of Λ_n is optimal.

In the case where the error is measured in $L^\infty(U,V)$, there is generally no simple description of the optimal set Λ_n. However, when the functions ϕ_ν are normalized in $L^\infty(U)$, the right-hand side in the estimate (3.5) provides a bound for the error of n-term approximation. This upper bound is minimized by again defining Λ_n as the set of indices corresponding to the n largest $\|u_\nu\|_V$. The only difference is that the error is measured by the ℓ^1 tail of the sequence $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$, in contrast to the ℓ^2 tail which appears in (3.4). Let us emphasize that this procedure gives only a bound for the

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

error of best n-term approximation in $L^\infty(U,V)$ but is not guaranteed to be the best error.

There is a good understanding of the properties of a given sequence $(c_\nu)_{\nu\in\mathcal{F}}$ of real or complex numbers, which ensure a certain rate of decay n^{-s} of its ℓ^q tail after one retains its n largest entries. The following result, originally due to Stechkin in the particular case $q=2$, shows that this rate of decay is related to the ℓ^p-summability of the sequence for values of p smaller than q.

Lemma 3.6. Let $0<p<q<\infty$ and let $(c_\nu)_{\nu\in\mathcal{F}}\in\ell^p(\mathcal{F})$ be a sequence of positive numbers. Then, if Λ_n is a set of indices which corresponds to the n largest c_ν, we have

$$\left(\sum_{\nu\notin\Lambda_n} c_\nu^q\right)^{1/q} \leq C(n+1)^{-s}, \quad C:=\|(c_\nu)_{\nu\in\mathcal{F}}\|_{\ell^p}, \quad s:=\frac{1}{q}-\frac{1}{p}. \tag{3.11}$$

Proof. Let $(c_k)_{k\geq 1}$ be the decreasing rearrangement of the sequence $(c_\nu)_{\nu\in\mathcal{F}}$. From the definition of Λ_n, we have

$$\sum_{\nu\notin\Lambda_n} c_\nu^q = \sum_{k\geq n+1} c_k^q \leq c_{n+1}^{q-p} \sum_{k\geq n+1} c_k^p \leq C^p c_{n+1}^{q-p}. \tag{3.12}$$

On the other hand, we also have

$$(n+1)c_{n+1}^p \leq \sum_{k=0}^{n+1} c_k^p \leq C^p. \tag{3.13}$$

Combining both estimates gives

$$\sum_{\nu\notin\Lambda_n} c_\nu^q \leq C^q (n+1)^{\frac{q}{p}-1}, \tag{3.14}$$

which is (3.11). □

Combining the above result with either (3.4) or (3.5) shows that a suitable ℓ^p-summability of the sequence $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ is a sufficient condition to guarantee a convergence rate n^{-s} when retaining the terms corresponding to the n largest $\|u_\nu\|_V$ in (3.1). For the $L^2(U,V,\mu)$ error, and when $(\phi_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis of $L^2(U,\mu)$, we obtain the rate $s=\frac{1}{p}-\frac{1}{2}$ if $p<2$. For the $L^\infty(U,V)$ error, and when the ϕ_ν are normalized in $L^\infty(U)$, we obtain the rate $s=\frac{1}{p}-1$ if $p<1$.

Remark 3.7. Lemma 3.6 shows that ℓ^p-summability implies that the ℓ^q tail of $(c_\nu)_{\nu\in\mathcal{F}}$, after retaining the best n-terms, decays with rate n^{-s}, where $s:=\frac{1}{p}-\frac{1}{q}$. It is in fact possible to *exactly* characterize the properties which govern this rate of decay via weaker summability properties. Let us recall that for $0<p<\infty$ the space $w\ell^p(\mathcal{F})$ consists of those sequences $(c_\nu)_{\nu\in\mathcal{F}}$ of

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

real or complex numbers such that, for a finite constant $C \geq 0$,

$$\#\{\nu : |c_\nu| \geq \eta\} \leq C^p \eta^{-p}, \quad \eta > 0, \tag{3.15}$$

or equivalently such that, for a finite constant $C \geq 0$,

$$c_k \leq C k^{-1/p}, \quad k \geq 1, \tag{3.16}$$

where $(c_k)_{k\geq 1}$ is the decreasing rearrangement of $(|c_\nu|)_{\nu\in\mathcal{F}}$. The quasi-norm $\|(c_\nu)_{\nu\in\mathcal{F}}\|_{w\ell^p(\mathcal{F})}$ can be defined as the smallest C for which either one of these inequalities holds. Then, for $0 < p < q \leq \infty$ one can check that the ℓ^q tail of $(c_\nu)_{\nu\in\mathcal{F}}$ after retaining the largest n-terms decays with rate n^{-s} where $s := \frac{1}{p} - \frac{1}{q}$, if and only if $(c_\nu)_{\nu\in\mathcal{F}} \in w\ell^p(\mathcal{F})$; see DeVore (1998).

3.2. Convergence of n-term truncated polynomial expansions

We now restrict our attention to polynomial series. This corresponds to particular choices of the functions ϕ_ν as polynomials. For the remainder of this section we take $\mathcal{F}$ to be the set of all sequences $\nu = (\nu_j)_{j\geq 1}$ of non-negative integers which are finitely supported. For $\nu \in \mathcal{F}$, we use the notation

$$\|\nu\|_0 := \#(\mathrm{supp}(\nu)) < \infty, \quad \mathrm{supp}(\nu) := \{j \geq 1 : \nu_j \neq 0\}, \tag{3.17}$$

as well as

$$|\nu| := \|\nu\|_1 = \sum_{j\geq 1} \nu_j < \infty. \tag{3.18}$$

For any $z = (z_j) \in \mathbb{C}^{\mathbb{N}}$, and $\nu \in \mathcal{F}$, we define

$$z^\nu := \prod_{j\geq 1} z_j^{\nu_j}. \tag{3.19}$$

We consider three types of polynomial series.

- Taylor (or power) series of the form

$$\sum_{\nu\in\mathcal{F}} t_\nu y^\nu, \tag{3.20}$$

  where

$$t_\nu := \frac{1}{\nu!}\partial^\nu u(y=0), \quad \nu! := \prod_{j\geq 1} \nu_j!, \tag{3.21}$$

  with the convention that $0! = 1$.

- Legendre series of the form

$$\sum_{\nu\in\mathcal{F}} v_\nu L_\nu(y), \quad L_\nu(y) = \prod_{j\geq 1} L_{\nu_j}(y_j), \tag{3.22}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

where $(L_k)_{k\geq 0}$ is the sequence of Legendre polynomials on $[-1,1]$ normalized with respect to the uniform measure, that is, such that

$$\int_{-1}^{1} |L_k(t)|^2 \frac{\mathrm{d}t}{2} = 1. \tag{3.23}$$

It follows that $(L_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis of $L^2(U,\mu)$, with

$$\mu = \otimes_{j\geq 1} \frac{\mathrm{d}y_j}{2} \tag{3.24}$$

the uniform measure over U. The coefficients v_ν are therefore given by

$$v_\nu := \int_U u(y) L_\nu(y)\, \mathrm{d}\mu(y). \tag{3.25}$$

- Renormalized Legendre series of the form

$$\sum_{\nu\in\mathcal{F}} w_\nu P_\nu(y), \quad P_\nu(y) = \prod_{j\geq 1} P_{\nu_j}(y_j), \tag{3.26}$$

where $(P_k)_{k\geq 0}$ is the sequence of Legendre polynomials on $[-1,1]$ with the standard normalization

$$\|P_k\|_{L^\infty([-1,1])} = P_k(0) = 1. \tag{3.27}$$

We have $L_k = \sqrt{1+2k}P_k$, and therefore the coefficients w_ν are given by

$$w_\nu := \left(\prod_{j\geq 1}(1+2\nu_j)\right)^{1/2} v_\nu, \tag{3.28}$$

where v_ν is defined by (3.25).

In the case of the Taylor series, the following result shows that the assumptions in Theorem 2.8 ensure the conditional convergence of (3.20) to u in $L^\infty(U,V)$.

Theorem 3.8. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. If the assumptions of Theorem 2.8 are satisfied, then the Taylor expansion (3.20) converges conditionally to u in $L^\infty(U,V)$.

Proof. Under the assumptions of Theorem 2.8, the Fréchet derivative of the solution map $a \mapsto u(a)$ is uniformly bounded over $a(\mathcal{U})$, and therefore

$$M := \max_{a\in a(\mathcal{U})} \|\mathrm{d}u(a)\|_{\mathcal{L}(X,V)} < \infty. \tag{3.29}$$

From the assumption that $(\|\psi_j\|_X)_{j\geq 1} \in \ell^1(\mathcal{F})$, for any $n \geq 1$, there exists $J = J(n)$ such that

$$\sum_{j\geq J+1} \|\psi_j\|_X \leq \frac{1}{2nM}. \tag{3.30}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Increasing the value of J decreases the left-hand side, so we may assume that $J(n) \geq n$.

We know from Theorem 2.8 that the map $y \mapsto u(y)$ admits a holomorphic extension $z \mapsto u(z)$ to domains $\mathcal{U}_\rho$ that contain $\mathcal{U}$. For any $z = (z_j)_{j\geq 1} \in \mathcal{U}$, we define its truncation

$$T_J z := (z_1, \dots, z_J, 0, 0, \dots), \tag{3.31}$$

and the map

$$v(z) := u(T_J z) = u(a(T_J z)) = u\left(\overline{a} + \sum_{j=1}^{J} z_j \psi_j\right). \tag{3.32}$$

Since, for $z \in \mathcal{U}$, we have

$$\|a(z) - a(T_J z)\|_X \leq \sum_{j\geq J+1} \|\psi_J\|_X \leq \frac{1}{2nM}, \tag{3.33}$$

it follows from (3.29) that

$$\|u - v\|_{L^\infty(\mathcal{U},V)} \leq \frac{1}{2n}. \tag{3.34}$$

Now we can write

$$v(z) = w(z_1, \dots, z_J), \tag{3.35}$$

where the finite-dimensional map w is bounded and holomorphic in each variable z_j on an open neighbourhood of the unit polydisc

$$\mathcal{U}_J := \otimes_{j=1}^{J} \{|z_j| \leq 1\}.$$

It follows that w has a Taylor expansion that converges on $\mathcal{U}_J$. Its Taylor coefficients are given by the t_ν for all ν of the form $(\nu_1, \dots, \nu_J, 0, 0, \dots)$. Therefore, there exists $K = K(n) \geq n$ such that, for

$$\Lambda_n := \{\nu \in \mathcal{F} : \operatorname{supp}(\nu) \subset \{1, \dots, J\} \text{ and } |\nu| \leq K\}, \tag{3.36}$$

we have

$$\sup_{z\in\mathcal{U}} \left\| v(z) - \sum_{\nu\in\Lambda_n} t_\nu z^\nu \right\|_V \leq \frac{1}{2n}, \tag{3.37}$$

and therefore

$$\sup_{z\in\mathcal{U}} \left\| u(z) - \sum_{\nu\in\Lambda_n} t_\nu z^\nu \right\|_V \leq \frac{1}{n}, \tag{3.38}$$

Since both $K(n)$ and $J(n)$ tend to infinity with n, the sequence of sets $(\Lambda_n)_{n\geq 0}$ is a exhaustion of $\mathcal{F}$. We have thus proved the conditional convergence of (3.20) to u in $L^\infty(\mathcal{U}, V)$, and hence also in $L^\infty(U, V)$. □

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We are now in a position to state our main result, which gives simple conditions that guarantee the ℓ^p-summability, $0 < p < 1$, of the sequence $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$, where u_ν is either t_ν, v_ν or w_ν. These conditions are expressed in terms of the ℓ^p-summability of the sequence $(\|\psi_j\|_X)_{j\geq 1}$ for the *same value of* p, and the assumptions in Theorem 2.8 in the case of t_ν or in Theorem 2.9 in the case of v_ν or w_ν.

Theorem 3.9. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. Then, the following summability results hold.

(i) If the assumptions of Theorem 2.8 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then $(\|t_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$ for the same value of p.

(ii) If the assumptions of Theorem 2.9 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then $(\|v_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$ and $(\|w_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$ for the same value of p.

The proof of this result is given in Section 3.7. For now, we use this theorem together with the previous results of this subsection to obtain corollaries on the rate of convergence of n-term approximations obtained by truncation of Taylor or Legendre series.

Corollary 3.10. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. If the assumptions of Theorem 2.8 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then the Taylor series (3.20) converges unconditionally to u in $L^\infty(U, V)$. Moreover, for any set Λ_n of indices corresponding to the n largest of $\|t_\nu\|_V$, we have

$$\sup_{y\in U}\left\|u(y) - \sum_{\nu\in\Lambda_n} t_\nu y^\nu\right\|_V \leq C(n+1)^{-s}, \quad s = \frac{1}{p} - 1, \tag{3.39}$$

where $C := \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p} < \infty$.

Proof. Using the first part of Theorem 3.9, we see that

$$(\|t_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p(\mathbb{N}).$$

Since, by Theorem 3.8, the series (3.20) converges conditionally, by application of Theorem 3.4, we find that it also converges unconditionally with the error bound

$$\sup_{y\in U}\left\|u(y) - \sum_{\nu\in\Lambda_n} t_\nu y^\nu\right\|_V \leq \sum_{\nu\notin\Lambda_n} \|t_\nu\|_V. \tag{3.40}$$

We now use (3.11) with $c_\nu = \|t_\nu\|_V$ and $q = 1$ to obtain the error bound (3.39). □

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Corollary 3.11. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. If the assumptions of Theorem 2.9 are satisfied, and if in addition

$$(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$$

for some $p < 1$, then the Legendre series (3.22) and (3.26) converges unconditionally to u in $L^\infty(U,V)$ and in $L^2(U,V,\mu)$, where μ is the uniform probability measure. In addition, we have the following error bounds.

- If Λ_n is the set of indices that corresponds to the n largest $\|v_\nu\|_V$, we have

$$\Big\|u - \sum_{\nu\in\Lambda_n} v_\nu L_\nu\Big\|_{L^2(U,V,\mu)} \leq C(n+1)^{-s}, \quad s = \frac{1}{p} - \frac{1}{2}, \tag{3.41}$$

  where $C := \|(\|v_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p} < \infty$.

- If Λ_n is the set of indices that corresponds to the n largest $\|w_\nu\|_V$, we have

$$\Big\|u - \sum_{\nu\in\Lambda_n} w_\nu P_\nu\Big\|_{L^\infty(U,V)} \leq C(n+1)^{-s}, \quad s = \frac{1}{p} - 1, \tag{3.42}$$

  where $C := \|(\|w_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p} < \infty$.

Proof. Using the second part of Theorem 3.9, we deduce that $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|w_\nu\|_V)_{\nu\in\mathcal{F}}$ belong to $\ell^p(\mathcal{F})$. The unconditional convergence claims in the theorem are ensured by Theorems 3.3 and 3.5. These latter two theorems also give the estimates

$$\|u - \sum_{\nu\in\Lambda_n} v_\nu L_\nu\|_{L^2(U,V,\mu)} = \Big(\sum_{\nu\notin\Lambda_n} \|v_\nu\|_V^2\Big)^{1/2}, \tag{3.43}$$

and

$$\|u - \sum_{\nu\in\Lambda_n} w_\nu P_\nu\|_{L^\infty(U,V,\mu)} = \sum_{\nu\notin\Lambda_n} \|w_\nu\|_V. \tag{3.44}$$

The application of (3.11) with $c_\nu = \|v_\nu\|_V$ and $q = 2$, or with $\|w_\nu\|_V$ and $q = 1$, give the error bounds (3.41) and (3.42). □

Remark 3.12. Note that since we have

$$v_\nu L_\nu = w_\nu P_\nu, \quad \nu \in \mathcal{F}, \tag{3.45}$$

the terms in the series (3.22) and (3.26) are actually identical. However the sets Λ_n defined by the n largest $\|v_\nu\|_V$ or the n largest $\|w_\nu\|_V$, which are used to define the truncations for L^2 or L^∞ estimates in the previous result, generally differ from each other.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The above corollaries show that the curse of dimensionality can be broken for certain classes of parametric PDEs: although the solution map $y \mapsto u(y)$ has infinitely many variables, it can be approximated in various norms with an algebraic rate n^{-s}, where n is the number of terms in the separable expansion. The exponent s can be large if $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for a small value of p. Several critical ingredients have been used in order to reach this conclusion:

- the holomorphic extension of the solution map $a \mapsto u(a)$,
- the anisotropy of the solution map with respect to the different variables y_j,
- the use of best n-term polynomial approximations.

The fact that anisotropic smoothness may allow certain numerical methods to break the curse of dimensionality, in the sense that approximation results are immune to the growth in the number of variables, has also been studied in information-based complexity, using certain weight sequences in order to quantify anisotropy; see Kuo, Sloan, Wasilkowski and Wozniakowski (2010).

Remark 3.13. Theorem 3.9 and its corollaries can be formulated for a general map u from $\mathcal{A}$ to V that is not necessarily the solution map of a parametric PDE, since, as observed in Remark 2.12, Theorems 2.8 and 2.9 hold in this more general framework.

3.3. Estimates of Taylor coefficients

In this subsection, as well as the two that follow, we establish upper estimates for the V-norms of the Taylor coefficients t_ν and Legendre coefficients v_ν and w_ν, which are instrumental in the proof of Theorem 3.9. These estimates are derived from the results on holomorphic extensions of the map $y \mapsto u(y)$ established in Theorems 2.8 and 2.9, namely by an application of the Cauchy integral formula in the different complex variables z_j.

We recall that the Cauchy formula states that if φ is a function from $\mathbb{C}$ to a Banach space V which is holomorphic on a simply connected open set $\mathcal{O} \subset \mathbb{C}$, and if Γ is a closed rectifiable curve contained in $\mathcal{O}$, then for any $\tilde{z}$ contained in the bounded domain delimited by Γ,

$$\varphi(\tilde{z}) := \frac{1}{2\pi \mathrm{i}} \int_\Gamma \frac{\varphi(z)}{\tilde{z} - z} \, \mathrm{d}z, \tag{3.46}$$

where the fraction in the integrand stands for the scalar multiplication of $\varphi(z) \in V$ by $(\tilde{z} - z)^{-1} \in \mathbb{C}$ and the curve Γ is positively oriented in the integral; see for instance Theorem 2.1.2 of Hervé (1989).

We begin with the estimates on Taylor coefficients which are based on the bounded holomorphic extensions onto polydiscs $\mathcal{U}_\rho$ that were established in Theorem 2.8.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Lemma 3.14. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. If the assumptions of Theorem 2.8 are satisfied, then there exists an $\varepsilon > 0$ and a $C > 0$ such that the estimates

$$\|t_\nu\|_V \leq C\rho^{-\nu} = C\prod_{j\geq 1}\rho_j^{-\nu_j}, \quad \nu \in \mathcal{F}, \tag{3.47}$$

hold for any sequence $\rho = (\rho_j)_{j\geq 0}$ of numbers larger than or equal to 1, for which

$$\sum_{j\geq 1}(\rho_j - 1)\|\psi_j\|_X \leq \varepsilon. \tag{3.48}$$

Proof. Let $\varepsilon > 0$ and C be as in Theorem 2.8, and let $\rho = (\rho_j)_{j\geq 1}$ of numbers larger than or equal to 1 satisfying the constraint (3.48). For each $j \geq 1$, let $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ be the open set that contains the disc $\{|z_j| \leq \rho_j\}$ given in Theorem 2.8. Then, we know that the map $y \mapsto u(y)$ admits an extension $z \mapsto u(z)$ onto the set $\mathcal{O}_\rho$ defined by (2.100), which is holomorphic in each variable z_j with uniform bound

$$\sup_{z\in\mathcal{O}_\rho}\|u(z)\|_V \leq C. \tag{3.49}$$

For any given $\nu \in \mathcal{F}$, we define

$$J := J(\nu) := \max\{j : \nu_j \neq 0\}. \tag{3.50}$$

Similar to (3.35), we introduce the finite-dimensional function w defined by

$$w(z_1,\ldots,z_J) = u(T_J z), \quad T_J z = (z_1,\ldots,z_J,0,0,\ldots), \tag{3.51}$$

so that we have for this particular ν

$$\partial^\nu u(0) = \frac{\partial^{|\nu|}w}{\partial z_1^{\nu_1}\cdots\partial z_J^{\nu_J}}(0,\ldots,0). \tag{3.52}$$

We know that w is holomorphic on the set

$$\mathcal{O}_{\rho,J} := \otimes_{1\leq j\leq J}\,\mathcal{O}\rho_j, \tag{3.53}$$

which is an open neighbourhood of the J-dimensional polydisc

$$\mathcal{U}_{\rho,J} := \otimes_{1\leq j\leq J}\,\mathcal{U}\rho_j. \tag{3.54}$$

In addition, we have

$$\sup_{(z_1,\ldots,z_J)\in\mathcal{U}_{\rho,J}}\|w(z_1,\ldots,z_J)\|_V \leq C. \tag{3.55}$$

We may thus apply the Cauchy formula (3.46) recursively in each variable z_j, and obtain for any $(\tilde{z}_1,\ldots,\tilde{z}_J)$ in the interior of $\mathcal{U}_{\rho,J}$ a representation of

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$w(\tilde{z}_1, \ldots, \tilde{z}_J)$ as a multiple integral

$$w(\tilde{z}_1, \ldots, \tilde{z}_J) = (2\pi \mathrm{i})^{-J} \int_{|z_1|=\rho_1} \cdots \int_{|z_J|=\rho_J} \frac{w(z_1, \ldots, z_J)}{(\tilde{z}_1 - z_1) \cdots (\tilde{z}_J - z_J)} \, \mathrm{d}z_1 \cdots \mathrm{d}z_J. \tag{3.56}$$

By differentiation, this yields

$$\frac{\partial^{|\nu|}}{\partial z_1^{\nu_1} \cdots \partial z_J^{\nu_J}} w(0, \ldots, 0) = \nu! (2\pi \mathrm{i})^{-J} \int_{|z_1|=\rho_1} \cdots \int_{|z_J|=\rho_J} \frac{w(z_1, \ldots, z_J)}{z_1^{\nu_1+1} \cdots z_J^{\nu_J+1}} \, \mathrm{d}z_1 \cdots \mathrm{d}z_J, \tag{3.57}$$

and therefore, using (3.55), we obtain the estimate

$$\|\partial^\nu u(0)\|_V = \left\| \frac{\partial^{|\nu|} w}{\partial z_1^{\nu_1} \cdots \partial z_J^{\nu_J}}(0, \ldots, 0) \right\|_V \le C \nu! \prod_{j \le J} \rho_j^{-\nu_j}, \tag{3.58}$$

which is equivalent to (3.47). □

Let us comment on the estimate (3.47). Since we may take any sequence ρ on the right-hand side, as long as it satisfies the constraint (3.48), we have the estimate

$$\|t_\nu\|_V \le C \min\left\{ \rho^{-\nu} : \sum_{j \ge 1} (\rho_j - 1) \|\psi_j\|_X \le \varepsilon \text{ and } \rho_j \ge 1, \ j \ge 1 \right\}. \tag{3.59}$$

It is possible to characterize the sequence ρ^* for which the minimum in the above right-hand side is attained. An important observation is that this minimizing sequence *depends on* ν.

To find ρ^*, we observe that for a given ν, this minimization problem is in fact finite-dimensional since $\rho^{-\nu}$ is not influenced by the values of ρ_j for those j such that $\nu_j = 0$. Since $\rho^{-\nu}$ is monotone non-increasing with ρ_j for the other values of j, and in view of the constraint (3.48), we should thus set

$$\rho_j^* = 1, \quad j \notin \operatorname{supp}(\nu). \tag{3.60}$$

It remains to solve the finite-dimensional problem

$$\min\left\{ \prod_{j \in \operatorname{supp}(\nu)} \rho_j^{-\nu_j} : \sum_{j \in \operatorname{supp}(\nu)} (\rho_j - 1) \|\psi_j\|_X \le \varepsilon \text{ and } \rho_j \ge 1, \ j \in \operatorname{supp}(\nu) \right\}, \tag{3.61}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

or equivalently

$$\max\Big\{ \sum_{j\in\mathrm{supp}(\nu)} \nu_j \log(\rho_j) : \sum_{j\in\mathrm{supp}(\nu)} (\rho_j - 1)\|\psi_j\|_X \le \varepsilon \text{ and } \rho_j \ge 1,\ j \in \mathrm{supp}(\nu)\Big\}, \tag{3.62}$$

which admits a unique solution since we minimize a strictly concave function over a convex set. The solution necessarily satisfies the equality constraint

$$\sum_{j\in\mathrm{supp}(\nu)} (\rho_j^* - 1)\|\psi_j\|_X = \varepsilon. \tag{3.63}$$

For the optimal solution ρ^*, if $E \subset \mathrm{supp}(\nu)$ is the subset of those $j \in \mathrm{supp}(\nu)$ such that $\rho_j^* > 1$, there exists a Lagrange multiplier $\lambda \in \mathbb{R}$ such that

$$\frac{\nu_j}{\rho_j^*} = \lambda\|\psi_j\|_X, \quad j \in E. \tag{3.64}$$

For any index $\nu = (\nu_j)_{j\ge1} \in \mathcal{F}$ and $E \subset \mathbb{N}$, we use the notation

$$\nu_E := (\tilde\nu_j)_{j\ge1}, \quad \tilde\nu_j = \nu_j \quad \text{if } j \in E, \quad \tilde\nu_j = 0 \quad \text{otherwise.} \tag{3.65}$$

Combining (3.64) and (3.63), we thus find that

$$\lambda = \frac{|\nu_E|}{\sigma_E + \varepsilon} \quad \text{where } \sigma_E := \sum_{j\in E} \|\psi_j\|_X. \tag{3.66}$$

Therefore, the solution $\rho^* = \rho^*(\nu) = (\rho_j^*)_{j\ge1}$, has the form

$$\rho_j^* = \frac{\nu_j(\sigma_E + \varepsilon)}{|\nu_E|\,\|\psi_j\|_X} \quad \text{if } j \in E, \quad \rho_j^* = 1 \quad \text{if } j \notin E. \tag{3.67}$$

This characterization is not satisfactory since the set E is not explicitly given. However, given any set E, we can define λ by (3.66) and define a corresponding sequence (ρ_j) as in (3.67). Therefore, the minimum we seek is the same as

$$\min_E \left(\frac{|\nu_E|}{\sigma_E + \varepsilon}\right)^{|\nu_E|} \prod_{j\ge1} \left(\frac{\|\psi_j\|_X}{\nu_j}\right)^{\nu_j}, \tag{3.68}$$

over all sets $E \subset \mathrm{supp}(\nu)$ for which the corresponding ρ_j given as in (3.67) are strictly larger than 1 for all $j \in E$. The optimal set E is the one for which this minimum is reached. This is a combinatorial problem, which is not easy to solve except for those $\nu \in \mathcal{F}$ of small support. For this reason, we do not make further use of the above optimal sequence $\rho^*(\nu)$ in bounding $\|t_\nu\|_V$. Instead, we use in Section 3.6 certain sub-optimal choices $\rho(\nu)$ which have an explicit expression inspired by (3.67).

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

3.4. Refined estimates for elliptic and parabolic PDEs

The estimate (3.47) can be refined in the particular case of the elliptic and parabolic problems (1.5) and (2.40). We recall that for each of these problems, the parameter a is taken in

$$X = L^\infty(D), \tag{3.69}$$

and that the uniform boundedness and holomorphy of the solution map is ensured under a condition of the form $\mathfrak{R}(a) \geq t$ for some $t > 0$. In such a case, we have seen in Section 2.5 that when the sequence $\rho = (\rho_j)_{j\geq 1}$ fulfils the constraints

$$\sum_{j\geq 1} \rho_j |\psi_j(x)| \leq \overline{a}(x) - t, \quad x \in D, \tag{3.70}$$

for some $t > 0$, the holomorphic extension is defined over the polydisc $\mathcal{U}_\rho$ with uniform bound

$$\sup_{z\in\mathcal{U}_\rho} \|u(z)\|_V \leq C_t. \tag{3.71}$$

By recursive application of Cauchy's formula, as in the proof of Lemma 3.14, we now obtain, for any fixed $t > 0$, the estimate

$$\|t_\nu\|_V \leq C_t \min\left\{\rho^{-\nu} : \sum_{j\geq 1} \rho_j |\psi_j(x)| \leq \overline{a}(x) - t, \quad x \in D\right\}. \tag{3.72}$$

It is not clear how to give a simple characterization of the above minimization problem, due to the form of the constraints (3.70) which need to be fulfilled for every $x \in D$. There are, however, two particular instances of affine decompositions where such a simple characterization exists.

The first of these is when the ψ_j have *disjoint supports*, by which we mean that

$$|\mathrm{supp}(\psi_i) \cap \mathrm{supp}(\psi_j)| = 0, \quad i \neq j. \tag{3.73}$$

In this case, the uniform ellipticity assumption UEA(r) holds if and only if

$$|\psi_j(x)| \leq \overline{a}(x) - r, \quad x \in D, \quad j \geq 1. \tag{3.74}$$

This instance is sometimes referred to as the model of *disjoint inclusions*. One particular example is when a is piecewise constant over a finite or infinite partition $(D_j)_{j\geq 1}$ of D, in which case $\overline{a}$ is a strictly positive constant and $\psi_j = c_j \chi_{D_j}$ for some positive numbers c_j, each of them smaller than $\overline{a} - r$.

In the general case of disjoint supports of the ψ_j, the constraint (3.70) can be decoupled, so that the minimization problem on the right-hand side of (3.72) is equivalent to

$$\min\{\rho^{-\nu} : \rho_j |\psi_j(x)| \leq \overline{a}(x) - t, \ x \in D, \ j \geq 1\}. \tag{3.75}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The optimal solution ρ^* to this problem is obviously given by

$$\rho_j^* = \inf_{x\in D} \frac{\overline{a}(x) - t}{|\psi_j(x)|}. \tag{3.76}$$

Let us note that in that case ρ^* does not depend on ν. This leads us to the estimate

$$\|t_\nu\|_V \le C_t \prod_{j\ge 1} \left(\sup_{x\in D} \frac{|\psi_j(x)|}{\overline{a}(x) - t} \right)^{\nu_j}. \tag{3.77}$$

If UEA(r) holds, we see that we can take each ρ_j^* strictly larger than 1 if we take $0 < t < r$, for example by setting $t = r/2$. In such a case, we indeed have

$$\frac{\overline{a}(x) - t}{|\psi_j(x)|} \ge \frac{\overline{a}(x) - t}{\overline{a}(x) - r} \ge 1 + \frac{r - t}{\overline{a}(x) - r} \ge 1 + \frac{r}{2\|\overline{a}\|_X}, \tag{3.78}$$

which shows that $\rho_j^* > 1$. Note that the values ρ_j^* increase as t decrease, which in principle results in a better bound for $\|t_\nu\|_V$. However, the constant C_t tends to $+\infty$ as $t \to 0$. One may in principle search for an optimal value of t, but we do not enter into this discussion.

The second instance is when the ψ_j are functions of constant moduli, such as complex exponentials. In this case, the uniform ellipticity assumption UEA(r) holds if and only if

$$\sum_{j\ge 1} \|\psi_j\|_X \le \overline{a}_{\min} - r, \quad \overline{a}_{\min} := \min_{x\in D} \overline{a}(x), \tag{3.79}$$

and the minimization problem on the right-hand side of (3.72) is equivalent to

$$\min \left\{ \rho^{-\nu} : \sum_{j\ge 1} \rho_j \|\psi_j\|_X \le \overline{a}_{\min} - t \right\}. \tag{3.80}$$

By the same Lagrange multiplier approach used above for the characterization of the minimizer in (3.59), we find that the above minimum is attained for $\rho^* = \rho^*(\nu) = (\rho_j^*)_{j\ge 1}$ given by

$$\rho_j^* = \frac{\nu_j(\overline{a}_{\min} - t)}{|\nu|\|\psi_j\|_X}. \tag{3.81}$$

This leads us to the estimate

$$\|t_\nu\|_V \le C_t \left(\frac{|\nu|}{\overline{a}_{\min} - t} \right)^{|\nu|} \prod_{j\ge 1} \left(\frac{\|\psi_j\|_X}{\nu_j} \right)^{\nu_j}. \tag{3.82}$$

3.5. Estimates of Legendre coefficients

Returning to general parametric PDEs of the form (1.1) with an affine representation (1.15), our next objective is to establish similar estimates

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

for the Legendre coefficients $\|v_\nu\|_V$ and $\|w_\nu\|_V$. Let us recall that these coefficients are given by

$$v_\nu = \int_U u(y) L_\nu(y)\, \mathrm{d}\mu(y), \tag{3.83}$$

and

$$w_\nu = \prod_{j\geq 1} (2\nu_j + 1) \int_U u(y) P_\nu(y)\, \mathrm{d}\mu(y). \tag{3.84}$$

They are linked by the relation

$$w_\nu = \left(\prod_{j\geq 1} (1 + 2\nu_j) \right)^{1/2} v_\nu. \tag{3.85}$$

We introduce the function

$$t \mapsto \theta(t) := \frac{\pi t}{2(t-1)}, \tag{3.86}$$

which is monotone non-increasing over $]1, +\infty[$.

The following result establishes estimates on the Legendre coefficients, based on the bounded holomorphic extension of u onto the polyellipses $\mathcal{H}_\rho$ and their neighbourhood $\mathcal{O}_\rho$ established in Theorem 2.9.

Lemma 3.15. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. If the assumptions of Theorem 2.9 are satisfied, then there exists $\varepsilon > 0$ and $C > 0$ such that the estimates

$$\|v_\nu\|_V \leq C \prod_{j\in \mathrm{supp}(\nu)} \theta(\rho_j)(1 + 2\nu_j)^{1/2} \rho_j^{-\nu_j} \tag{3.87}$$

and

$$\|w_\nu\|_V \leq C \prod_{j\in \mathrm{supp}(\nu)} \theta(\rho_j)(1 + 2\nu_j) \rho_j^{-\nu_j} \tag{3.88}$$

hold for any sequence $\rho = (\rho_j)_{j\geq 0}$ of numbers strictly larger than 1, which satisfies the constraint (3.48).

Proof. We only need to prove (3.87), since (3.88) then follows by (3.85). Let $\varepsilon > 0$ and C be as in Theorem 2.9, and let $\rho = (\rho_j)_{j\geq 1}$ be a sequence of numbers strictly larger than 1, which satisfies the constraint (3.48). We know that for each $j \geq 1$ there exists an open set $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ that contains the filled-in ellipse $\mathcal{H}_{\rho_j}$ and such that the map $y \mapsto u(y)$ admits an extension $z \mapsto u(z)$ over the set $\mathcal{O}_\rho$ defined by (2.100), which is holomorphic in each variable z_j with uniform bound

$$\sup_{z\in \mathcal{O}_\rho} \|u(z)\|_V \leq C. \tag{3.89}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We observe that $U \subset \mathcal{O}_\rho$.

In the case $\nu = 0$, the estimate (3.87) is immediate since

$$\|w_0\|_V = \left\| \int_U u(y)\, \mathrm{d}\mu(y) \right\|_V \leq \sup_{y \in U} \|u(y)\|_V \leq C, \tag{3.90}$$

where we have used the fact that μ is a probability measure. We now assume that $\nu \neq 0$. Up to a reordering of $(\psi_j)_{j \geq 1}$, we may assume without loss of generality that $\nu_j \neq 0$ for $j \leq J$ and $\nu_j = 0$ for $j > J$ for $J = |\mathrm{supp}(\nu)| \geq 1$. We partition the variable y into

$$y = (y_1, \ldots, y_J, y'), \quad y' := (y_{J+1}, y_{J+2}, \ldots) \in [-1,1]^{\mathbb{N}} = U, \tag{3.91}$$

and rewrite (3.84) as

$$w_\nu = \prod_{j=1}^{J} (2\nu_j + 1) \int_U v(y')\, \mathrm{d}\mu(y'), \tag{3.92}$$

where

$$v(y') := \int_{[-1,1]^J} u(y_1, \ldots, y_J, y') \left(\prod_{j=1}^{J} P_{\nu_j}(y_j) \right) \frac{\mathrm{d}y_1}{2} \cdots \frac{\mathrm{d}y_J}{2}. \tag{3.93}$$

For a fixed $y' \in U$, we use the holomorphy of the finite-dimensional map $(z_1, \ldots, z_J) \mapsto u(z_1, \ldots, z_J, y')$ in order to evaluate $\|v(y')\|_V$. For this purpose, we introduce for any integer $n \geq 1$ the following function of a single complex variable z,

$$Q_n(z) := \int_{-1}^{1} \frac{P_n(y)}{z - y}\, \mathrm{d}y, \tag{3.94}$$

and the corresponding multivariate functions

$$Q_\nu(z_1, \ldots, z_J) := \prod_{j=1}^{J} Q_{\nu_j}(z_j), \tag{3.95}$$

which are well defined as long as $|z_j| > 1$ for $j = 1, \ldots, J$. For our given ρ, we introduce the J-dimensional polyellipse

$$\mathcal{E}_{\rho,J} := \otimes_{1 \leq j \leq J} \mathcal{E}_{\rho_j}. \tag{3.96}$$

Since $\rho_j > 1$, $1 \leq j \leq J$, the unit interval $[-1,1]$ is contained in the interior of each filled-in ellipse $\mathcal{H}_{\rho_j}$. Therefore, we may recursively apply Cauchy's integral formula on each ellipse $\mathcal{E}_{\rho_j}$ for each of the variables z_j, $j = 1, \ldots, J$, and obtain

$$u(y_1, \ldots, y_J, y') = \frac{1}{(2\pi \mathrm{i})^J} \int_{\mathcal{E}_{\rho_1}} \cdots \int_{\mathcal{E}_{\rho_J}} \frac{u(z_1, \ldots, z_J, y')}{(y_1 - z_1) \cdots (y_J - z_J)}\, \mathrm{d}z_1 \cdots \mathrm{d}z_J, \tag{3.97}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

for any $(y_1, \dots, y_J) \in [-1,1]^J$ and any $y' \in U$. Multiplying by $\prod_{j=1}^J P_{\nu_j}(y_j)$ and integrating over $[-1,1]^J$ with respect to $(\mathrm{d}y_1/2)\cdots(\mathrm{d}y_J/2)$, we therefore obtain

$$v(y') = \left(\frac{\mathrm{i}}{4\pi}\right)^J \int_{\mathcal{E}_{\rho_1}} \cdots \int_{\mathcal{E}_{\rho_J}} u(z_1, \dots, z_J, y') Q_\nu(z_1, \dots, z_J)\, \mathrm{d}z_1 \cdots \mathrm{d}z_J. \tag{3.98}$$

From the uniform bound (3.89) we know that

$$\begin{aligned}(z_1, \dots, z_J) \in \mathcal{E}_{\rho,J} \text{ and } y' \in U &\Rightarrow (z_1, \dots, z_J, y') \in \mathcal{O}_\rho \\ &\Rightarrow \|u(z_1, \dots, z_J, y')\|_V \le C. \end{aligned} \tag{3.99}$$

Injecting this bound in the above integral yields

$$\|v(y')\|_V \le C \left(\prod_{j=1}^J \frac{\rho_j}{2}\right) \max_{(z_1,\dots,z_J)\in\mathcal{E}_{\rho,J}} |Q_\nu(z_1, \dots, z_J)|, \quad y' \in U, \tag{3.100}$$

where we have used the fact the perimeter of $\mathcal{E}_{\rho_j}$ has length smaller than $2\pi\rho_j$. We now use the estimate

$$\max_{z\in\mathcal{E}_t} |Q_n(z)| \le \frac{\pi\ t^{-n}}{t-1} \tag{3.101}$$

(see Davis 1963, p. 313), which yields

$$\max_{(z_1,\dots,z_J)\in\mathcal{E}_{\rho,J}} |Q_\nu(z_1, \dots, z_J)| \le \prod_{j=1}^J \frac{\pi\ \rho_j^{-\nu_j}}{\rho_j - 1}, \tag{3.102}$$

and therefore

$$\|v(y')\|_V \le C \prod_{j=1}^J \theta(\rho_j) \rho_j^{-\nu_j}, \quad y' \in U. \tag{3.103}$$

Combining this estimate with (3.92), we obtain

$$\|w_\nu\|_V \le \prod_{j=1}^J (1+2\nu_j) \sup_{y'\in U} \|v(y')\|_V \le C \prod_{j\in\mathrm{supp}(\nu)} \theta(\rho_j)(1+2\nu_j)\rho_j^{-\nu_j}, \tag{3.104}$$

which is (3.87). □

Estimates (3.88) and (3.87) are very similar to the estimate (3.47) obtained in Lemma 3.14 for the Taylor coefficients, but with two noticeable differences.

- On the one hand, the estimates for the Legendre coefficients are a bit more pessimistic, due to the presence of the additional factors $\theta(\rho_j)$ and $(1 + 2\nu_j)$. Intuitively, these factors are absorbed by the decay

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

of the factor $\rho_j^{-\nu_j}$ when ρ_j or ν_j become large. The analysis in the next subsection confirms that they do not affect the ℓ^p-summability properties of the estimate.

- On the other hand, these estimates are obtained under much weaker conditions than those of Lemma 3.14. Indeed Theorem 2.9 only requires the existence of a holomorphic extension of the solution map $a \mapsto u(a)$ in a neighbourhood of $a(U)$, in contrast to Theorem 2.8 which requires a neighbourhood of $a(\mathcal{U})$. In particular, for problems such as (2.58) or (2.66), we know that the conditions of Theorem 2.9 are met but not those of Theorem 2.8.

Similar to the estimate for Taylor coefficients, we can use the fact that (3.88) and (3.87) hold for any sequence ρ satisfying the prescribed constraints, in order to obtain the estimates

$$\|v_\nu\|_V \leq C \inf\left\{ \prod_{j\in\mathrm{supp}(\nu)} \theta(\rho_j)(1+2\nu_j)^{1/2}\rho_j^{-\nu_j} \right\}, \tag{3.105}$$

and

$$\|w_\nu\|_V \leq C \inf\left\{ \prod_{j\in\mathrm{supp}(\nu)} \theta(\rho_j)(1+2\nu_j)\rho_j^{-\nu_j} \right\}, \tag{3.106}$$

where the infima are taken over all sequences ρ of numbers strictly larger than 1, such that

$$\sum_{j\geq 1}(\rho_j - 1)\|\psi_j\|_X \leq \varepsilon.$$

Remark 3.16. The values of ρ_j enter the above estimates only for $j \in \mathrm{supp}(\nu)$. This implies that we can consider the above infima over all sequences ρ of numbers larger than or equal to 1 with $\rho_j > 1$ if $j \in \mathrm{supp}(\nu)$ and such that

$$\sum_{j\in\mathrm{supp}(\nu)} (\rho_j - 1)\|\psi_j\|_X \leq \varepsilon,$$

which amounts to taking $\rho_j = 1$ if $j \notin \mathrm{supp}(\nu)$.

3.6. Summability of multi-indexed sequences

We want to use the upper estimates obtained for $\|t_\nu\|_V$, $\|v_\nu\|_V$ and $\|w_\nu\|_V$ derived in the previous subsections in order to prove Theorem 3.9. As a preliminary step, in this subsection we establish several results concerning the ℓ^p-summability of certain multi-indexed sequences, which appear in the proof of Theorem 3.9 below.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We begin by considering sequences of the form $(b^\nu)_{\nu\in\mathcal{F}}$, where $b=(b_j)_{j\geq 1}$ is a given sequence of positive numbers. For such sequences we have the following elementary result.

Lemma 3.17. For any $0<p<\infty$, the sequence $(b^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $b\in\ell^p(\mathbb{N})$ and $\|b\|_{\ell^\infty}<1$. Moreover

$$\|(b^\nu)_{\nu\in\mathcal{F}}\|_{\ell^p}\leq \exp\left(c_p\frac{\|b\|_{\ell^p}^p}{p}\right),\quad c_p:=\frac{1}{1-\|b\|_{\ell^\infty}^p}. \tag{3.107}$$

Proof. For any positive integer J, let $\mathcal{F}_J$ denote the set of those $\nu\in\mathcal{F}$ such that $\operatorname{supp}(\nu)\subset\{1,\dots,J\}$. Now, if $\|b\|_{\ell^\infty}<1$, we can write

$$\sum_{\nu\in\mathcal{F}_J} b^{p\nu}=\prod_{1\leq j\leq J}\sum_{n\geq 0} b_j^{pn}=\prod_{1\leq j\leq J}\frac{1}{1-b_j^p},\quad J=1,2,\dots. \tag{3.108}$$

If $b\in\ell^p(\mathbb{N})$, we can let J tend to $+\infty$ and obtain

$$\sum_{\nu\in\mathcal{F}} b^{p\nu}=\prod_{j\geq 1}\frac{1}{1-b_j^p}<\infty. \tag{3.109}$$

This proves the one implication of the theorem. Since

$$\begin{aligned}\prod_{j\geq 1}\frac{1}{1-b_j^p}&=\prod_{j\geq 1}\left(1+\frac{b_j^p}{1-b_j^p}\right)\leq\prod_{j\geq 1}\exp\left(\frac{b_j^p}{1-b_j^p}\right)\\&\leq\prod_{j\geq 1}\exp(c_p b_j^p)=\exp(c_p\|b\|_{\ell^p}^p),\end{aligned} \tag{3.110}$$

we also have the bound (3.107).

For the other implication, we observe that the sequences $(b_j)_{j\geq 1}$ and $(b_j^n)_{n\geq 0}$, for any $j\geq 1$, are subsequences of $(b^\nu)_{\nu\in\mathcal{F}}$ corresponding to particular selections of indices ν. This shows that the ℓ^p-summability of $(b^\nu)_{\nu\in\mathcal{F}}$ implies both that $b\in\ell^p(\mathbb{N})$ and $\|b\|_{\ell^\infty}<1$. □

One immediate application of the above lemma concerns the ℓ^p-summability of the Taylor coefficients for the elliptic or parabolic problems in the model of disjoint inclusions discussed in Section 3.4. In this case, estimate (3.77) has the form

$$\|t_\nu\|_V\leq C_t b^\nu,\quad \text{where } b=(b_j)_{j\geq 1}\quad \text{with } b_j:=\sup_{x\in D}\frac{|\psi_j(x)|}{\overline{a}(x)-t}. \tag{3.111}$$

Given UEA(r) and taking $t=r/2$, we know from (3.78) that for $X:=L^\infty(D)$,

$$\|b\|_{\ell^\infty}\leq\frac{2\|\overline{a}\|_X}{2\|\overline{a}\|_X+r}<1. \tag{3.112}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Since in addition

$$b_j \leq \frac{2\|\psi_j\|_X}{r}, \tag{3.113}$$

this shows that $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ implies $b \in \ell^p(\mathbb{N})$. Combining these observations with Lemma 3.17, we thus find that if UEA(r) holds and if $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$, then the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$, which is a particular case of Theorem 3.9.

Remark 3.18. We have mentioned in Remark 3.7 that the convergence rate n^{-s} of best n-term approximation in ℓ^q spaces is equivalent to the property of weak ℓ^p-summability with $s = \frac{1}{p} - \frac{1}{q}$. Therefore, a relevant question is whether the above Lemma 3.17 is valid with ℓ^p replaced by $w\ell^p$. Surprisingly, the answer is negative, and closely related to classical results in number theory. Indeed, fix any $0 < p < 1$ and consider the prototype sequence $b \in w\ell^p(\mathbb{N})$ given by

$$b_j = (j+1)^{-1/p}. \tag{3.114}$$

This sequence also satisfies $\|b\|_{\ell^\infty} < 1$. If we were to have $(b^\nu)_{\nu\in\mathcal{F}} \in w\ell^p(\mathcal{F})$ then there would be a constant C such that, for any $\eta > 0$, we have

$$\#\{\nu \in \mathcal{F} : b^\nu \geq \eta\} \leq C\eta^{-p}, \tag{3.115}$$

or equivalently, such that, for any $A \geq 2$,

$$t(A) := \#\left\{\nu \in \mathcal{F} : \prod_{j\geq 2} j^{\nu_j} \leq A\right\} \leq CA. \tag{3.116}$$

The left-hand side can be rewritten as

$$t(A) = \sum_{n=2}^{\lfloor A\rfloor} f(n), \tag{3.117}$$

where $f(n)$ is the number of possible multiplicative partitions of n. The problem of counting multiplicative partitions of natural numbers, sometimes referred to as *factorisatio numerorum*, has been extensively studied in number theory; see in particular Canfield, Erdős and Pomerance (1983), who give a sharp asymptotic bound for $f(n)$. Luca, Mukhopadhyay and Srinivas (2010) prove that the total number of multiplicative partitions $t(A)$ has the asymptotic behaviour

$$\frac{t(A)}{A} \sim \exp\left\{\frac{4\sqrt{\log(A)}}{\sqrt{2e}\log(\log(A))}(1+o(1))\right\} \to +\infty \tag{3.118}$$

as $A \to +\infty$. This shows that (3.116) does not hold, and thus that $(b^\nu)_{\nu\in\mathcal{F}}$ does not belong to $w\ell^p(\mathcal{F})$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We make further use of a slightly more general version of Lemma 3.17, where we incorporate additional algebraic factors into the sequence b^ν.

Lemma 3.19. For a given sequence $b = (b_j)_{j\geq 1}$ of positive numbers, and for non-negative numbers c and r, let $(b_\nu)_{\nu\in\mathcal{F}}$ be defined by

$$b_\nu := b^\nu \prod_{j\geq 1}(1 + c\nu_j^r) = \prod_{j\geq 1}(1 + c\nu_j^r)b_j^{\nu_j}. \tag{3.119}$$

For any $0 < p < \infty$, this sequence belongs to $\ell^p(\mathcal{F})$ if and only if $b \in \ell^p(\mathbb{N})$ and $\|b\|_{\ell^\infty} < 1$.

Proof. Since $b_\nu \geq b^\nu$, the 'only if' part follows from Lemma 3.17 and therefore we only need to prove the 'if' part. With $\mathcal{F}_J$ as in the proof of Lemma 3.17, we write

$$\sum_{\nu\in\mathcal{F}_J} b_\nu^p = \prod_{1\leq j\leq J}\sum_{n\geq 0}(1 + cn^r)^p b_j^{pn}, \tag{3.120}$$

Since $\|b\|_{\ell^\infty} \leq 1$ we find that

$$\sum_{n\geq 0}(1 + cn^r)^p b_j^{pn} \leq 1 + Cb_j^p, \tag{3.121}$$

where the constant C depends on c, r, p and $\|b\|_{\ell^\infty}$. Since $b \in \ell^p(\mathbb{N})$, this shows that the product on the right-hand side of (3.120) converges as $J \to \infty$. Therefore $(b_\nu)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$. □

The estimates obtained for $\|t_\nu\|_V$, $\|v_\nu\|_V$ and $\|w_\nu\|_V$ also involve quantities of the form

$$\frac{|\nu|^{|\nu|}}{\prod_{j\geq 1}\nu_j^{\nu_j}} d^\nu, \tag{3.122}$$

for sequences $d = (d_j)_{j\geq 1}$ of positive numbers. In view of the Stirling inequalities

$$n! \leq n^n \leq n!\,\mathrm{e}^n, \tag{3.123}$$

we may write

$$\frac{|\nu|^{|\nu|}}{\prod_{j\geq 1}\nu_j^{\nu_j}} d^\nu \leq \mathrm{e}^{|\nu|}\frac{|\nu|!}{\nu!} d^\nu = \frac{|\nu|!}{\nu!} b^\nu, \tag{3.124}$$

where

$$b = (b_j)_{j\geq 0}, \quad b_j = ed_j. \tag{3.125}$$

This suggests we should study the ℓ^p-summability of sequences of the form $\left(\frac{|\nu|!}{\nu!}b^\nu\right)_{\nu\in\mathcal{F}}$. Due to the presence of the multinomial factor $\frac{|\nu|!}{\nu!}$, which can be

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

much larger than 1, we expect the conditions for ℓ^p-summability to be more stringent than for the sequence $(b^\nu)_{\nu\in\mathcal{F}}$. This is confirmed by the following result.

Lemma 3.20. For any $0 < p < 1$, a sequence $\left(\frac{|\nu|!}{\nu!}b^\nu\right)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $b \in \ell^p(\mathbb{N})$ and $\|b\|_{\ell^1} < 1$.

Proof. We first observe that whenever $b \in \ell^1(\mathbb{N})$, the multinomial formula gives

$$\sum_{|\nu|=k} \frac{|\nu|!}{\nu!} b^\nu = \left(\sum_{j\geq 1} b_j\right)^k. \tag{3.126}$$

Summing over k, we see that $\left(\frac{|\nu|!}{\nu!}b^\nu\right)_{\nu\in\mathcal{F}}$ is in $\ell^1(\mathcal{F})$ if and only if $b \in \ell^1(\mathbb{N})$ and $\|b\|_{\ell^1(\mathbb{N})} < 1$. Moreover,

$$\left\|\left(\frac{|\nu|!}{\nu!}b^\nu\right)_{\nu\in\mathcal{F}}\right\|_{\ell^1(\mathcal{F})} = \sum_{\nu\in\mathcal{F}} \frac{|\nu|!}{\nu!} b^\nu = \frac{1}{1-\|b\|_{\ell^1}}, \tag{3.127}$$

Now suppose that

$$\left(\frac{|\nu|!}{\nu!}b^\nu\right)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$$

for some $p \leq 1$. Then b is in $\ell^p(\mathbb{N})$ since it is a subsequence of $\bar{b}$ corresponding to a particular selection of indices ν. Also, $\bar{b}$ is in $\ell^1(\mathcal{F})$, so b must be in $\ell^1(\mathcal{N})$ with norm smaller than one.

Conversely, assume that $b \in \ell^p(\mathbb{N})$ and $\|b\|_{\ell^1} < 1$. We claim that there exist two positive sequences $c = (c_j)_{j\geq 1}$ and $d = (d_j)_{j\geq 1}$ with the following properties:

(i) $b_j = c_j d_j$ for all $j \geq 1$,

(ii) $c \in \ell^1(\mathbb{N})$ with $\|c\|_{\ell^1} < 1$,

(iii) $d \in \ell^q(\mathbb{N})$ with $\frac{1}{q} = \frac{1}{p} - 1$, or equivalently $q = \frac{p}{1-p}$, and $\|d\|_{\ell^\infty} < 1$.

Before proving this claim, let us show that it implies the ℓ^p-summability of $\left(\frac{|\nu|!}{\nu!}b^\nu\right)_{\nu\in\mathcal{F}}$. Indeed, from Hölder's inequality we have

$$\begin{aligned}\sum_{\nu\in\mathcal{F}}\left(\frac{|\nu|!}{\nu!}b^\nu\right)^p &= \sum_{\nu\in\mathcal{F}}\left(\frac{|\nu|!}{\nu!}c^\nu\right)^p d^{p\nu} \\ &\leq \left(\sum_{\nu\in\mathcal{F}}\frac{|\nu|!}{\nu!}c^\nu\right)^p \left(\sum_{\nu\in\mathcal{F}} d^{q\nu}\right)^{1-p}.\end{aligned}$$

As observed previously in (3.127), the first factor is finite due to the fact that $\|c\|_{\ell^1} < 1$. The second factor is finite by application of Lemma 3.17.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

It remains to prove the claim by constructing specific sequences c and d having the prescribed properties. With $\delta := 1 - \|b\|_{\ell^1} > 0$, we define

$$\eta := \frac{\delta}{3}, \tag{3.128}$$

and take J sufficiently large that

$$\sum_{j>J} b_j^p \le \frac{\delta}{3}. \tag{3.129}$$

We then define c and d by

$$c_j = (1+\eta)b_j \quad \text{and} \quad d_j = \frac{1}{1+\eta}, \quad j \le J, \tag{3.130}$$

and

$$c_j = b_j^p \quad \text{and} \quad d_j = b_j^{1-p}, \quad j > J. \tag{3.131}$$

By construction, we have $c_j d_j = b_j$ for all $j \ge 1$. For the sequence c, we have

$$\|c\|_{\ell^1} \le (1+\eta)\|b\|_{\ell^1} + \sum_{j>J} b_j^p \le \left(1+\frac{\delta}{3}\right)(1-\delta) + \frac{\delta}{3} \le 1 - \frac{\delta}{3}, \tag{3.132}$$

We next bound $\|d\|_{\ell^\infty}$. For $1 \le j \le J$ we have $d_j = \frac{1}{1+\eta} < 1$, and for $j > J$ we have

$$d_j = \left(b_j^p\right)^{(1-p)/p} \le \left(\frac{\delta}{3}\right)^{(1-p)/p} < 1. \tag{3.133}$$

Therefore, we have $\|d\|_{\ell^\infty} < 1$. Finally, since $d_j^q = b_j^p$ for $j > J$, we find that $d \in \ell^q(\mathbb{N})$, which completes the confirmation of the claim. □

Similar to Lemma 3.19, the following result shows that ℓ^p-summability is maintained if we incorporate additional algebraic factors.

Lemma 3.21. For a given sequence $b = (b_j)_{j\ge1}$ of positive numbers, and for non-negative numbers c and r, let $(b_\nu)_{\nu\in\mathcal{F}}$ be defined by

$$b_\nu := \frac{|\nu|!}{\nu!} b^\nu \prod_{j\ge1} (1 + c\nu_j^r). \tag{3.134}$$

For any $0 < p < 1$, this sequence belongs to $\ell^p(\mathcal{F})$ if and only if $b \in \ell^p(\mathbb{N})$ and $\|b\|_{\ell^1} < 1$.

Proof. Since $b_\nu \ge \frac{|\nu|!}{\nu!} b^\nu$, the 'only if' part follows from Lemma 3.20 and we only need to prove the 'if' part.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Using the same sequences c and d as in the proof of Lemma 3.20, and introducing

$$d_\nu = d^\nu \prod_{j\geq 1}(1+c\nu_j^r), \tag{3.135}$$

we write

$$\begin{aligned}\sum_{\nu\in\mathcal{F}} b_\nu^p &= \sum_{\nu\in\mathcal{F}}\left(\frac{|\nu|!}{\nu!}c^\nu\right)^p d_\nu^p \\ &\leq \left(\sum_{\nu\in\mathcal{F}}\frac{|\nu|!}{\nu!}c^\nu\right)^p\left(\sum_{\nu\in\mathcal{F}} d_\nu^q\right)^{1-p},\end{aligned}$$

and conclude similarly that both factors are finite, using Lemma 3.19 for the second factor. □

3.7. Proof of Theorem 3.9

In order to prove Theorem 3.9, we use the estimates (3.47), (3.88) and (3.87) for $\|t_\nu\|_V$, $\|v_\nu\|_V$ and $\|w_\nu\|_V$, respectively. The right-hand side of these estimates has a general form $Cr(\nu,\rho)$ for any sequence ρ of numbers larger than 1 that satisfy the constraint (3.48). Our objective is to build for each ν such a sequence $\rho = \rho(\nu)$, and show that, for $0 < p < 1$ the resulting quantities

$$r_\nu := r(\nu,\rho(\nu)) \tag{3.136}$$

are ℓ^p-summable provided that $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$. Obviously, it is sufficient to treat the case when

$$r(\nu,\rho) := \prod_{j\in\operatorname{supp}(\nu)} \theta(\rho_j)(1+2\nu_j)\rho_j^{-\nu_j}, \tag{3.137}$$

which appears on the right-hand side of (3.87), since it is the largest estimate.

We fix any $\nu \in \mathcal{F}$ and describe our choice for the sequence ρ to be inserted into the above expression. In what follows, we use the notation

$$b = (b_j)_{j\geq 1}, \quad \text{where } b_j := \|\psi_j\|_X, \quad j \geq 1. \tag{3.138}$$

For $J \geq 1$ to be fixed further, we split $\mathbb{N}$ into

$$E := \{1,\dots,J\} \quad \text{and} \quad F := \{J+1, J+2, \dots\}, \tag{3.139}$$

and use the notation $\nu_E = (\nu_1,\dots,\nu_J) \in \mathbb{N}^J$ and $\nu_F = (\nu_{J+1},\nu_{J+2},\dots) \in \mathcal{F}$. In view of Remark 3.16, we may take

$$\rho_j = 1, \quad j \notin \operatorname{supp}(\nu). \tag{3.140}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

With ε the right-hand side of the constraint (3.48), we then take

$$\rho_j = \kappa := 1 + \frac{\varepsilon}{2\|b\|_{\ell^1}}, \quad j \in E \cap \operatorname{supp}(\nu), \tag{3.141}$$

and

$$\rho_j = \kappa + \frac{\varepsilon \nu_j}{2 b_j |\nu_F|}, \quad j \in F \cap \operatorname{supp}(\nu). \tag{3.142}$$

Therefore $\rho_j > 1$ when $j \in \operatorname{supp}(\nu)$, and in addition

$$\sum_{j \geq 1} (\rho_j - 1) b_j \leq \frac{\varepsilon \sum_{j \leq J} b_j}{2\|b\|_{\ell^1}} + \sum_{j > J} \left(\frac{\varepsilon b_j}{2\|b\|_{\ell^1}} + \frac{\varepsilon \nu_j}{2|\nu_F|} \right) \leq \varepsilon, \tag{3.143}$$

which shows that the constraint (3.48) is satisfied.

When using this choice for the sequence ρ, the resulting estimate may be written as

$$r_\nu = r(\nu, \rho(\nu)) = r_E(\nu) r_F(\nu), \tag{3.144}$$

where

$$\begin{aligned} r_E(\nu) &:= \theta(\kappa)^J \prod_{j \in E} (1 + 2\nu_j) \kappa^{-\nu_j}, \\ r_F(\nu) &:= \prod_{j \in F \cap \operatorname{supp}(\nu)} \theta(\rho_j)(1 + 2\nu_j) \rho_j^{-\nu_j}. \end{aligned} \tag{3.145}$$

Letting $\mathcal{F}_E$ and $\mathcal{F}_F$ denote the multi-indices in $\mathcal{F}$ supported on E and F, respectively, we may then write

$$\sum_{\nu \in \mathcal{F}} r_\nu^p = \Sigma_E \Sigma_F, \tag{3.146}$$

where

$$\Sigma_E := \sum_{\nu \in \mathcal{F}_E} r_E(\nu)^p \quad \text{and} \quad \Sigma_F := \sum_{\nu \in \mathcal{F}_F} r_F(\nu)^p, \tag{3.147}$$

provided that both sums converge.

The first sum Σ_E is estimated by

$$\begin{aligned} \Sigma_E &= \theta(\kappa)^{pJ} \sum_{\nu \in \mathbb{N}^J} \prod_{j=1}^{J} (1 + 2\nu_j)^p \kappa^{-p\nu_j} \\ &= \theta(\kappa)^{pJ} \left(\sum_{n \geq 0} (1 + 2n)^p \kappa^{-pn} \right)^J < \infty, \end{aligned}$$

For the second sum Σ_F, we first notice that for each $\nu \in \mathcal{F}_F$,

$$r_F(\nu) \leq \prod_{j \in F \cap \operatorname{supp}(\nu)} \theta(\kappa)(1 + 2\nu_j) \left(\frac{\varepsilon \nu_j}{2 b_j |\nu_F|} \right)^{-\nu_j}, \tag{3.148}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

where we have used the fact that $\theta(\kappa) = \max_{t\geq\kappa}\theta(t) \geq \theta(\rho_j)$ for $j \in F$. Therefore, with $c := 3\theta(\kappa)$, we find that

$$r_F(\nu) \leq |\nu_F|^{|\nu_F|}\prod_{j\in F}\frac{(1+c\nu_j)\left(\frac{2b_j}{\varepsilon}\right)^{\nu_j}}{\nu_j^{\nu_j}}$$
$$\leq \frac{|\nu_F|!}{\nu_F!}\prod_{j\in F}(1+c\nu_j)\left(\frac{2eb_j}{\varepsilon}\right)^{\nu_j},$$

where we have used (3.123). Introducing the sequence $d = (d_j)_{j\geq 1}$ defined by

$$d_j = \frac{2eb_{j+J}}{\varepsilon}, \tag{3.149}$$

we thus find that

$$\Sigma_F \leq \sum_{\nu\in\mathcal{F}} d_\nu^p \quad \text{where } d_\nu := \frac{|\nu|!}{\nu!}d^\nu\prod_{j\geq 1}(1+c\nu_j). \tag{3.150}$$

We now choose J sufficiently large that

$$\|d\|_{\ell^1} = \sum_{j>J}\frac{2eb_j}{\varepsilon} < 1. \tag{3.151}$$

Since our assumption $b \in \ell^p(\mathbb{N})$ implies that $d \in \ell^p(\mathbb{N})$, we may apply Lemma 3.21 to conclude that Σ_F is finite. The proof of Theorem 3.9 is complete.

Remark 3.22. One defect in the proof of Theorem 3.9 is that, while it establishes the ℓ^p-summability of the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|w_\nu\|_V)_{\nu\in\mathcal{F}}$, it does not provide us with a simple bound of the ℓ^p-norms of these sequences in terms of the ℓ^p-norm of the sequence $(\|\psi_j\|_X)_{j\geq 1}$.

3.8. Approximation using downward closed sets

Theorem 3.9 has implications for the rate convergence of polynomial approximations obtained by retaining the terms in Taylor and Legendre series corresponding to the n largest coefficients. Corollaries 3.10 and 3.11 show that these approximations converge with the rates n^{-s}, where $s = \frac{1}{p} - 1$ for uniform convergence and $s = \frac{1}{p} - \frac{1}{2}$ for convergence in $L^2(U, V, \mu)$.

These results should be viewed as a theoretical justification that reduced modelling methods based on polynomial approximations may perform well for parametric PDEs which satisfy the assumptions of Theorem 3.9. However, they in no way constitute a numerical algorithm, since finding the optimal sets $(\Lambda_n)_{n\geq 1}$ is out of reach in practice, as is the exact computation of the Taylor and Legendre coefficients.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Practical algorithms for the computation of polynomial approximations are discussed below in Sections 5 and 6. The implementation and analysis of the algorithms presented there benefit from imposing additional structure on the index sets Λ_n used to define the polynomial approximation. To define this structure, we first recall that $\mathcal{F}$ has a partial ordering: for $\nu, \tilde{\nu} \in \mathcal{F}$, we write $\tilde{\nu} \leq \nu$ if and only if $\tilde{\nu}_j \leq \nu_j$ for all $j \geq 1$. We also write $\tilde{\nu} < \nu$ if and only if $\tilde{\nu} \leq \nu$ and $\tilde{\nu}_j < \nu_j$ for at least one value of j.

Definition 3.23. A set $\Lambda \subset \mathcal{F}$ is called *downward closed* or a *lower set* if and only if

$$\nu \in \Lambda \quad \text{and} \quad \tilde{\nu} \leq \nu \text{ implies } \tilde{\nu} \in \Lambda. \tag{3.152}$$

When considering polynomial spaces

$$\mathbb{P}_\Lambda := \operatorname{span}\{y \mapsto y^\nu : \nu \in \Lambda\}, \tag{3.153}$$

it is quite natural to make the assumption that Λ is a downward closed set. In particular, this assumption allows us to describe $\mathbb{P}_\Lambda$ in terms of any tensorized polynomial basis of the form

$$\phi_\nu(y) = \prod_{j \geq 1} \phi_{\nu_j}(y_j), \tag{3.154}$$

where $(\phi_k)_{k \geq 0}$ is any family of univariate polynomials such that $\phi_0 = 1$ and ϕ_k has degree exactly k. This includes in particular the tensorized Legendre polynomials L_ν. By expressing each monomial $y \mapsto y^k$ as a linear combination of the ϕ_l for $0 \leq l \leq k$, we find that $\mathbb{P}_\Lambda$ is equivalently defined by

$$\mathbb{P}_\Lambda := \operatorname{span}\{\phi_\nu : \nu \in \Lambda\}. \tag{3.155}$$

Polynomial spaces associated to downward closed sets were introduced by Kuntzman (1959), in dimension $d = 2$, and referred to as *polynômes pleins*. Later, these notions were studied in general dimension d by de Boor and Ron (1992) and Lorentz and Lorentz (1986). Note that in dimension $d = 1$, a downward closed set is simply of the form $\{0, 1, \ldots, n\}$.

The index sets Λ_n corresponding to the n largest $\|t_\nu\|_V$, $\|v_\nu\|_V$ or $\|w_\nu\|_V$ are generally *not* downward closed sets. A legitimate question is therefore whether there exist nested sequences $(\Lambda_n)_{n \geq 0}$ of downward closed sets such that the truncated Taylor or Legendre series using such sets have the same convergence rates as those obtained in Corollaries 3.10 and 3.11, using the n largest $\|t_\nu\|_V$, $\|v_\nu\|_V$ or $\|w_\nu\|_V$. The results of the present subsection give a positive answer to this question.

Let us begin by observing that if a sequence $(c_\nu)_{\nu \in \mathcal{F}}$ of positive numbers is monotone non-increasing, that is, if

$$\nu \leq \tilde{\nu} \Rightarrow c_{\tilde{\nu}} \leq c_\nu, \tag{3.156}$$

then the set Λ_n corresponding to the n largest values of c_ν is downward

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

closed, provided that it is unique. In the case of non-uniqueness, there is at least one realization of such a set which is downward closed. In addition, there exists a sequence $(\Lambda_n)_{n\geq 1}$ of such realizations which is nested. Note that in such a realization, we necessarily have $\Lambda_0 = \{0\}$.

For an arbitrary sequence $c = (c_\nu)_{\nu\in\mathcal{F}} \in \ell^\infty(\mathcal{F})$, we introduce its *monotone majorant*, which is the sequence $\hat{c} = (\hat{c}_\nu)_{\nu\in\mathcal{F}}$ defined by

$$\hat{c}_\nu := \sup_{\tilde{\nu}\geq\nu} |c_{\tilde{\nu}}|. \tag{3.157}$$

This is the smallest monotone non-increasing sequence that dominates c. In order to study best n-term approximations using downward closed sets, we introduce the following sequence spaces.

Definition 3.24. For $0 < p < \infty$, we say that a sequence $c \in \ell^\infty(\mathcal{F})$ belongs to $\ell^p_m(\mathcal{F})$ if and only its monotone majorant $\hat{c}$ belongs to $\ell^p(\mathcal{F})$, and we define

$$\|c\|_{\ell^p_m} := \|\hat{c}\|_{\ell^p}. \tag{3.158}$$

Combining this definition with Lemma 3.6 shows that if $0 < p < q \leq \infty$ and if $(c_\nu)_{\nu\in\mathcal{F}}$ is a positive sequence which belongs to $\ell^p_m(\mathcal{F})$, then we have the tail bound

$$\left(\sum_{\nu\notin\Lambda_n} c_\nu^q\right)^{1/q} \leq Cn^{-s}, \quad C = \|(c_\nu)_{\nu\in\mathcal{F}}\|_{\ell^p_m}, \quad s := \frac{1}{p} - \frac{1}{q}, \tag{3.159}$$

where Λ_n is any downward closed set of indices corresponding to the n largest terms of the monotone majorant $\hat{c}$ of c. We may therefore obtain the same rate n^{-s} as in Lemma 3.6 but now using downward closed sets.

We would therefore like to know under which circumstances the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|w_\nu\|_V)_{\mathcal{F}}$ belong to $\ell^p_m(\mathcal{F})$. The following result, originally proved in Chkifa *et al.* (2013) in the case of elliptic parametric PDEs and in Chkifa *et al.* (2015*b*) for other models, shows that this holds under the exact same assumptions as in Theorem 3.9.

Theorem 3.25. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. Then, the following summability results hold.

- If the assumptions of Theorem 2.8 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then $(\|t_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p_m(\mathcal{F})$ for the same value of p.
- If the assumptions of Theorem 2.9 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then $(\|v_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p_m(\mathcal{F})$ and $(\|w_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p_m(\mathcal{F})$ for the same value of p.

Proof. Similar to the proof of Theorem 3.9, we use the estimates (3.47), (3.88) and (3.87) for the $\|t_\nu\|_V$, $\|v_\nu\|_V$ and $\|w_\nu\|_V$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

In the case of $\|t_\nu\|_V$, the estimate has the form

$$\|t_\nu\|_{\nu\in\mathcal{F}} \le e_\nu := C \inf \rho^{-\nu}, \tag{3.160}$$

where the infimum is taken over the set of sequences ρ of numbers larger than 1 that satisfy the constraint (3.48). Since for any such ρ, the sequence $(\rho^{-\nu})_{\nu\in\mathcal{F}}$ is monotone non-increasing, it follows that the sequence $(e_\nu)_{\nu\in\mathcal{F}}$ is also monotone non-increasing. On the other hand, the proof of Theorem 3.9 shows that $(e_\nu)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$. This implies that $(\|t_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p_m(\mathcal{F})$.

We cannot proceed in the same way for the Legendre coefficients $\|v_\nu\|_V$ and $\|w_\nu\|_W$ since the right-hand side $Cr(\nu,\rho)$ in estimates (3.87) and (3.88) does not have the monotone non-increasing property, due to the presence of the factors $\theta(\rho_j)$ and $(1+2\nu_j)$. Instead we slightly modify the construction of the sequence $\rho = \rho(\nu)$ in the proof of Theorem 3.9, and show that the resulting sequence of estimates

$$r_\nu = r(\nu, \rho(\nu)) \tag{3.161}$$

has a monotone majorant which is ℓ^p-summable over $\mathcal{F}$. Here again, it suffices to work with the estimate (3.87), which is the largest one.

We use the same notation as in Theorem 3.9, in particular $b_j := \|\psi_j\|_X$. For a constant $\beta > 0$ to be fixed later, we take $J \ge 1$ sufficiently large that

$$\sum_{j>J} b_j \le \frac{\varepsilon}{3\beta}, \tag{3.162}$$

where ε is the right-hand side of the constraint (3.48).

We now let $\nu \in \mathcal{F}$ and fix ν, and proceed to define an appropriate sequence $\rho = \rho(\nu)$ for this ν. Namely, using the same splitting of $\mathbb{N}$ into E and F, we take

$$\rho_j = \kappa := 1 + \frac{\varepsilon}{3\|b\|_{\ell^1}}, \quad j \in E \cap \operatorname{supp}(\nu), \tag{3.163}$$

where $b = (b_j)_{j\ge1}$ and

$$\rho_j = \kappa + \beta + \frac{\varepsilon\nu_j}{3b_j|\nu_F|}, \quad j \in F \cap \operatorname{supp}(\nu). \tag{3.164}$$

We again take $\rho_j = 1$ if $j \notin \operatorname{supp}(\nu)$. Therefore $\rho_j > 1$ when $j \in \operatorname{supp}(\nu)$, and in addition

$$\sum_{j\ge1}(\rho_j - 1)b_j \le \frac{\varepsilon \sum_{j\le J} b_j}{3\|b\|_{\ell^1}} + \sum_{j>J}\left(\frac{\varepsilon b_j}{3\|b\|_{\ell^1}} + \beta b_j + \frac{\varepsilon\nu_j}{3|\nu_F|}\right) \le \varepsilon. \tag{3.165}$$

which shows that the constraint (3.48) is satisfied.

For this choice of ρ, the estimate (3.87) may be written

$$\|w_\nu\|_V \le r_\nu = r(\nu,\rho(\nu)) = r_E(\nu) r_F(\nu), \tag{3.166}$$

with $r_E(\nu)$ as in the proof of Theorem 3.9, and a slightly modified $r_F(\nu)$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

that incorporates the new form of ρ_j for $j \in F$. This new $r_F(\nu)$ satisfies

$$r_F(\nu) \le \tilde{r}_F(\nu) := \prod_{j\in F\cap \mathrm{supp}(\nu)} \theta(\kappa)(1+2\nu_j)\left(\beta + \frac{\varepsilon\nu_j}{3b_j|\nu_F|}\right)^{-\nu_j}$$
$$\le \prod_{j\in F\cap \mathrm{supp}(\nu)} \theta(\kappa)(1+2\nu_j)\left(\frac{\varepsilon\nu_j}{3b_j|\nu_F|}\right)^{-\nu_j}.$$

Since $\kappa > 1$, there exists $C_0 = C_0(\kappa) > 0$ such that

$$(1+2n) \le C_0\left(\frac{1+\kappa}{2}\right)^n$$

for any $n \ge 1$, and so we can write

$$r_E(\nu) \le \tilde{r}_E(\nu) := C\prod_{j\in E}\eta^{\nu_j}, \quad \eta := \frac{1+\kappa}{2\kappa} < 1 \quad \text{and} \quad C = (C_0\theta(\kappa))^J. \tag{3.167}$$

The same argument as in the proof of Theorem 3.9 shows that, up to choosing a larger J, the estimates

$$\tilde{r}_\nu := \tilde{r}_E(\nu)\tilde{r}_F(\nu), \tag{3.168}$$

are ℓ^p-summable over $\mathcal{F}$.

We conclude by showing that $(\tilde{r}_\nu)_{\nu\in\mathcal{F}}$ is monotone non-increasing if B has been chosen large enough. On the one hand, since $\eta < 1$, it is readily seen that

$$\nu \le \tilde{\nu} \Rightarrow \tilde{r}_E(\tilde{\nu}) \le \tilde{r}_E(\nu). \tag{3.169}$$

To prove a similar monotonicity property for the second factor $\tilde{r}_F$, it suffices to show that $\tilde{r}_F(\nu)$ is reduced if we increase ν_j by 1 for any $j > J$, that is,

$$\tilde{r}_F(\nu + e_j) \le \tilde{r}_F(\nu), \tag{3.170}$$

where $e_j = (0,\dots,0,1,0,\dots)$ is the Kronecker sequence with 1 at position $j > J$. In the case where $\nu_j \ne 0$, we may write

$$\frac{\tilde{r}_F(\nu+e_j)}{\tilde{r}_F(\nu)}$$
$$= \frac{1+2\nu_j+2}{1+2\nu_j}\frac{\left(\beta+\frac{\varepsilon\nu_j}{3b_j|\nu_F|}\right)^{\nu_j}}{\left(\beta+\frac{\varepsilon(\nu_j+1)}{3b_j(|\nu_F|+1)}\right)^{\nu_j+1}}\prod_{k\in F\cap\mathrm{supp}(\nu)-\{j\}}\left(\frac{\beta+\frac{\varepsilon\nu_k}{3b_k|\nu_F|}}{\beta+\frac{\varepsilon\nu_k}{3b_k(|\nu_F|+1)}}\right)^{\nu_k}$$
$$\le \frac{2}{\beta+\frac{\varepsilon(\nu_j+1)}{3b_j(|\nu_F|+1)}}\prod_{k\in F\cap\mathrm{supp}(\nu)}\left(\frac{\beta+\frac{\varepsilon\nu_k}{3b_k|\nu_F|}}{\beta+\frac{\varepsilon\nu_k}{3b_k(|\nu_F|+1)}}\right)^{\nu_k} \le \frac{2}{\beta}\left(\frac{1+|\nu_F|}{|\nu_F|}\right)^{|\nu_F|},$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and therefore

$$\frac{\tilde{r}_F(\nu+e_j)}{\tilde{r}_F(\nu)} \leq \frac{2e}{\beta}. \tag{3.171}$$

In the case where $\nu_j = 0$, we have

$$\frac{\tilde{r}_F(\nu+e_j)}{\tilde{r}_F(\nu)} = \frac{3c_\kappa}{\beta + \frac{\varepsilon(\nu_j+1)}{3b_j(|\nu_F|+1)}} \prod_{k\in F\cap \mathrm{supp}(\nu)-\{j\}} \left(\frac{\beta + \frac{\varepsilon\nu_k}{3b_k|\nu_F|}}{\beta + \frac{\varepsilon\nu_k}{3b_k(|\nu_F|+1)}}\right)^{\nu_k} \leq \frac{3c_\kappa e}{\beta}. \tag{3.172}$$

We thus find that $(\tilde{r}_\nu)_{\nu\in\mathcal{F}}$ is monotone non-increasing provided that $\beta \geq \max\{2e, 3c_\kappa e\}$. □

Combining the above theorem with (3.159), we obtain the following result.

Corollary 3.26. Corollaries 3.10 and 3.11 remain valid if the sets Λ_n correspond to the n largest terms in the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ or $(\|w_\nu\|_V)_{\nu\in\mathcal{F}}$ replaced by downward closed sets Λ_n corresponding to the n largest terms in the monotone majorants of each of these sequences.

3.9. Exponential approximation rates

The rates of convergence n^{-s} that are established for polynomial approximations in Corollaries 3.10 and 3.11 are of algebraic type. We conclude this study of polynomial approximation with a brief discussion on the circumstances in which faster rates of exponential type can be established. For this, we focus on the finite-dimensional case, that is, when finitely many ψ_j are non-zero in the affine representation (1.15). In such a case, one first obvious observation is that since $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for all values of $p > 0$, Corollaries 3.10 and 3.11 give convergence rates n^{-s} for all $s > 0$. However, a more detailed inspection shows that the multiplicative constant C_s grows very fast to $+\infty$ as $s \to +\infty$. Instead of trying to search for a fast rate by optimizing $C_s n^{-s}$ over s for a given n, we return to the estimates on the polynomial coefficients and use them to obtain exponential convergence rates for the truncated series (3.20), (3.22) or (3.26).

Without loss of generality, we assume that only $\{\psi_1, \ldots, \psi_d\}$ are non-zero, meaning that the scalar parameter vector is now

$$y = (y_1, \ldots, y_d) \in U := [-1, 1]^d, \tag{3.173}$$

and that the solution map $y \mapsto u(y)$ from U to V is finite-dimensional. Polynomial approximations are again based on truncation of the series (3.20), (3.22) or (3.26), now with

$$\mathcal{F} = \mathbb{N}^d. \tag{3.174}$$

For the sake of simplicity, we focus our attention on Taylor series and make some further remarks on the case of Legendre series.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

A particularly simple case for estimates of Taylor coefficients is that of the disjoint inclusion model for the elliptic and parabolic PDEs (1.5) and (2.40) discussed in Section 3.4. In this case, given UEA(r), we explicitly solved (3.75) for any given $0 < t < r$ and obtained the estimate

$$\|t_\nu\|_V \le C\rho^{-\nu} = C\prod_{j=1}^{d}\rho_j^{-\nu_j}, \tag{3.175}$$

with $C = C_t$, and

$$\rho_j = \rho_j^* = \inf_{x\in D}\frac{\overline{a}(x) - t}{|\psi_j(x)|} > 1, \quad j = 1,\dots,d. \tag{3.176}$$

More generally, if we work under the assumptions of Theorem 2.8, we know from Lemma 3.14 that we have an estimate of the form (3.175) for any choice of $\rho_j \ge 1$ such that

$$\sum_{j=1}^{d}(\rho_j - 1)\|\psi_j\|_X \le \varepsilon. \tag{3.177}$$

We may, for instance, take

$$\rho_j := 1 + \frac{\varepsilon}{\|\psi_j\|_X} > 1, \quad j = 1,\dots,J. \tag{3.178}$$

Thus we again obtain (3.175) with a fixed finite vector $(\rho_1,\dots,\rho_d)$ independent of ν and whose coordinates are strictly larger than 1.

Based on such an estimate, a natural choice for the sets Λ_n is to pick the indices ν corresponding to the n largest values of $\rho^{-\nu}$. Equivalently, for any given threshold $\eta > 0$ we define

$$\Lambda_n := \{\nu \in \mathcal{F} : \rho^{-\nu} \ge \eta\}, \quad \text{where } n = n(\eta) := \#\{\nu \in \mathcal{F} : \rho^{-\nu} \ge \eta\}. \tag{3.179}$$

Note that as we vary $\eta > 0$, it may be that not all values of n arise because of possible ties in the values of $\rho^{-\nu}$.

Let us now focus on the particular thresholds

$$\eta = 2^{-k}, \quad k \ge 0, \tag{3.180}$$

we may write, with $n := n(k)$ growing with k,

$$\Lambda_n = S_k := \left\{\nu \in \mathcal{F} : \sum_{j=1}^{d}\lambda_j\nu_j \le k\right\}, \quad \lambda_j := \log_2(\rho_j) > 0. \tag{3.181}$$

Sets of this type consist of all integer lattice points inside the simplex with bounding hyperplanes given by the coordinate hyperplanes together with the hyperplane $\sum_{j=1}^{d} t_j\lambda_j = k$. Note that these sets are downward closed.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The cardinality of the above Λ_n is bounded from above by the volume of the continuous simplex

$$T_k := \left\{(t_1,\dots,t_d)\in\mathbb{R}^d : t_j \geq -1,\ j=1,\dots,d,\ \text{and}\ \sum_{j=1}^d \lambda_j t_j \leq k\right\}$$
$$= \left\{(t_1,\dots,t_d)\in\mathbb{R}^d : t_j \geq -1,\ j=1,\dots,d,\ \text{and}\ \sum_{j=1}^d \lambda_j t_j \leq k\right\}. \tag{3.182}$$

This gives the crude cardinality bound

$$\#(\Lambda_n) = \#(S_k) \leq |T_k| = \frac{1}{d!}\prod_{j=1}^d \left(\frac{k+\sum_{j=1}^d \lambda_j}{\lambda_j}\right) \leq Ck^d, \tag{3.183}$$

where C depends on d and on $(\lambda_1,\dots,\lambda_d)$.

Likewise, we may estimate the approximation error when retaining the n-terms whose indices are in Λ_n by

$$\begin{aligned}\sup_{y\in U}\left\|u(y)-\sum_{\nu\in S_k} t_\nu y^\nu\right\| &\leq \sum_{\nu\notin S_k}\|t_\nu\|_V\\ &\leq C\sum_{\nu\notin S_k}\rho^{-\nu}\\ &\leq C\sum_{l\geq k}2^{-l}\#\{\nu\ :\ 2^{-l-1}\leq\rho^{-\nu}<2^{-k}\}\\ &\leq C\sum_{l\geq k}2^{-l}\#(S_{l+1}).\end{aligned}$$

Using the estimate (3.182) on the asymptotic growth of $\#(S_k)$, we thus find that

$$\sup_{y\in U}\left\|u(y)-\sum_{\nu\in S_k} t_\nu y^\nu\right\| \leq C\sum_{l\geq k}2^{-l}(l+1)^d. \tag{3.184}$$

Combining this estimate with (3.182), we obtain

$$\sup_{y\in U}\left\|u(y)-\sum_{\nu\in S_k} t_\nu y^\nu\right\| \leq C\exp(-ck), \tag{3.185}$$

which is equivalent to the exponential rate

$$\sup_{y\in U}\left\|u(y)-\sum_{\nu\in \Lambda_n} t_\nu y^\nu\right\| \leq C\exp(-cn^{-1/d}), \tag{3.186}$$

with multiplicative constants c and C that depend on d and on $(\lambda_1,\dots,\lambda_d)$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Since this rate is valid for all n of the form $\#(S_k)$ which grow like k^d, it is easily seen that it is also valid for all values of $n \geq 1$, up to a change in the multiplicative constants.

Remark 3.27. We notice that this exponential rate deteriorates as d grows, due to the power $1/d$, as well as to the hidden dependence on d in the constants c and C. However in the case where the ℓ^p-norm of $(\|\psi_j\|_X)_{j=1,\dots,d}$ remains uniformly bounded for some $0 < p < 1$ as we raise d, our analysis of the infinite-dimensional case always ensures the algebraic rate n^{-s} with $s := \frac{1}{p} - 1$.

Remark 3.28. A similar analysis leads to the same exponential rates for approximation by truncated Legendre series, now under the assumptions of Theorem 2.9, based on the estimates (3.87) and (3.88), up to a proper treatment of the algebraic factors $\theta(\nu_j)$ and $(1 + 2\nu_j)$ appearing in these estimates.

4. Estimating the n-widths of solution manifolds

We have already noted that, when approximating the solution map by separable expansions of the form (1.32) or (1.33), the best achievable error in $L^\infty(\mathcal{A}, V)$ or in $L^\infty(U_\mathcal{A}, V)$ is described by the n-width of the solution manifold $\mathcal{M} = u(\mathcal{A})$ in V, that is,

$$d_n(\mathcal{M})_V := \inf_{\dim(V_n)=n} \max_{v\in\mathcal{M}} \min_{w\in V_n} \|v - w\|_V. \tag{4.1}$$

In this section we use the polynomial approximation results established in the previous section to derive *a priori* estimates for the decay of $d_n(\mathcal{M})_V$.

4.1. Estimates of n-width by polynomial approximation

In the case where Assumption A holds, we may use the polynomial approximation results of Section 3 to estimate $d_n(\mathcal{M})_V$ from above. Indeed, if $u_n(y) = \sum_{\nu\in\Lambda_n} c_\nu y^\nu$ is a polynomial approximation to the map $y \mapsto u(y)$ for some set $\Lambda_n \subset \mathcal{F}$ of cardinality n, we define the n-dimensional space

$$V_n := \operatorname{span}\{c_\nu : \nu \in \Lambda_n\} \subset V, \tag{4.2}$$

and observe that

$$d_n(\mathcal{M})_V \leq \max_{v\in\mathcal{M}} \min_{w\in V_n} \|v-w\|_V = \max_{y\in U_\mathcal{A}} \min_{w\in V_n} \|u(y)-w\|_V \leq \|u-u_n\|_{L^\infty(U,V)}. \tag{4.3}$$

Therefore a polynomial approximation bound in $L^\infty(U, V)$ induces an estimate on the n-width of $\mathcal{M}$ in V. Combining this observation with Corollary 3.11, we obtain the following result.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Corollary 4.1. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer (1.15). Assume that the solution map $u \mapsto u(a)$ admits a holomorphic extension over an open set $\mathcal{O} \subset X$ which contains the compact set $a(U)$, and this extension satisfies the uniform bound

$$\sup_{a\in\mathcal{O}} \|u(a)\|_V \leq C. \tag{4.4}$$

If $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$, then

$$d_n(\mathcal{M})_V \leq C(n+1)^{-s}, \quad n \geq 1, \quad s := \frac{1}{p} - 1, \tag{4.5}$$

for a suitable constant C.

Proof. We consider the truncated Legendre expansion

$$u_n = \sum_{\nu\in\Lambda_n} w_\nu P_\nu, \tag{4.6}$$

where Λ_n is the set of indices corresponding to the n largest $\|w_\nu\|_V$. Since the assumptions of Corollary 3.11 are satisfied, we obtain (4.5) with $C := \|(\|w_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p}$, by combining (4.3) and (3.42). □

One drawback of the above result is that it requires Assumption A. For some natural examples of compact sets $\mathcal{A}$ of X, this assumption may not hold. For instance, in the case of the elliptic equation (1.5), we know that the standard compact sets of $X = L^\infty(D)$ are described by a smoothness assumption. A typical example for $\mathcal{A}$ of this type is

$$\mathcal{A} := \{a \in X : a > r,\ \|a\|_{C^\beta} \leq M\}, \tag{4.7}$$

for some $M, \beta, r > 0$, where $C^\beta := C^\beta(D)$ is the Hölder space with smoothness $\beta > 0$, equipped with its usual norm

$$\|a\|_{C^\beta} := \sup_{|\alpha|<m} \|\partial^\alpha a\|_{L^\infty} + \sup_{|\alpha|=m} \sup_{x,x'\in D} |x-x'|^{-(\beta-m)} |\partial^\alpha a(x) - \partial^\alpha a(x')| \tag{4.8}$$

for $m := \lfloor\beta\rfloor$. For such $\mathcal{A}$, there are many ways to choose an $\overline{a} \in \mathcal{A}$ and a properly normalized basis $(\psi_j)_{j\geq 1}$ such that expanding $a - \overline{a}$ in this basis allows us to write

$$\mathcal{A} \subset a(U), \tag{4.9}$$

with $a(U)$ of the form (1.28). However, it will generally not follow that there is an $r' > 0$ such that, for each a in $a(U)$, we have $a > r'$. Therefore, we are not guaranteed to have well-posedness of the PDE for all $u(y) \in U$ and so Assumption A will not hold for this affine representation. We fix this defect in the next subsection with a different approach based on local polynomial approximations.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

4.2. Estimates of n-width by local polynomial approximation

In this subsection we treat parameter sets $\mathcal{A} \in X$ which might not satisfy Assumption A. We assume that $(\psi_j)_{j\geq 1}$ is a complete representer for $\mathcal{A}$ and in addition that $(\|\psi_j\|_X) \in \ell^1(\mathbb{N})$. It follows that, for each $(z_j)_{j\geq 1} \in \mathcal{U}$, the series $\sum_{j\geq 1} z_j\psi_j$ converges in X, and so the set

$$\mathcal{R} := \left\{ \sum_{j\geq 1} z_j\psi_j : z = (z_j)_{j\geq 1} \in \mathcal{U} \right\} \tag{4.10}$$

is well defined. We replace Assumption A with the requirement

$$\mathcal{A} \subset \mathcal{R}. \tag{4.11}$$

Note that, in contrast to $\mathcal{A}$, the set $\mathcal{R}$ might not be contained in the open set $\mathcal{O}$ over which the solution map admits a bounded holomorphic extension. However, we will remedy this problem by using the following covering result.

Lemma 4.2. Let $\mathcal{A}$ be a compact set in a complex Banach space X, and assume that $\mathcal{A} \subset \mathcal{R}$, where $\mathcal{R}$ is of the form (4.10) for a family of functions $(\psi_j)_{j\geq 1}$ such that $(\|\psi_j\|_X)_{j\geq 1} \in \ell^1(\mathbb{N})$. Let $\mathcal{O}$ be any open set of X which contains $\mathcal{A}$. Then there exist $\eta, \varepsilon > 0$, an integer $J \geq 1$ and a finite collection $\{\overline{a}_1, \ldots, \overline{a}_M\} \subset X$ such that, defining

$$\tilde{\psi}_j := \eta\psi_j, \quad j = 1, \ldots, J, \quad \tilde{\psi}_j := \psi_j, \quad j > J, \tag{4.12}$$

and for any sequence $z = (z_j)_{j\geq 1} \in \mathbb{C}^{\mathbb{N}}$,

$$a_i(z) := \overline{a}_i + \sum_{j\geq 1} z_j\tilde{\psi}_j, \quad i = 1, \ldots, M, \tag{4.13}$$

whenever the series on the right converges, the following holds.

(i) The compact set $\mathcal{A}$ admits the following cover:

$$\mathcal{A} \subset \bigcup_{i=1}^{M} \mathcal{A}_i, \quad \mathcal{A}_i := a_i(\mathcal{U}) = \{a_i(z) : z \in \mathcal{U}\}. \tag{4.14}$$

(ii) The compact sets $\mathcal{A}_i$, $i = 1, \ldots, M$, are all contained in $\mathcal{O}$.

(iii) For any sequence $\rho = (\rho_j)_{j\geq 1}$ of numbers, each larger than 1, which satisfies the constraint

$$\sum_{j\geq 1} (\rho_j - 1)\|\tilde{\psi}_j\|_X \leq \varepsilon,$$

there exists, for each $j \geq 1$, an open set $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ which contains the disc $\{|z_j| \leq \rho_j\}$ and for which the set $\mathcal{O}_\rho := \otimes_{j\geq 1}\mathcal{O}_{\rho_j}$ satisfies

$$a_i(\mathcal{O}_\rho) := \{a_i(z) : z \in \mathcal{U}_\rho\} \subset \mathcal{O}. \tag{4.15}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Proof. Similar to the proof Theorem 2.8, we first observe that since $\mathcal{A}$ is compact, there is an $\varepsilon > 0$ such that the 3ε neighbourhood of $\mathcal{A}$ is contained in $\mathcal{O}$, that is,

$$\bigcup_{a\in\mathcal{A}} B(a, 3\varepsilon) \subset \mathcal{O}. \tag{4.16}$$

For this ε, we next choose $J \geq 1$ sufficiently large that

$$\sum_{j>J} \|\psi_j\|_X \leq \frac{\varepsilon}{4}. \tag{4.17}$$

We then define

$$\eta := \frac{\varepsilon}{4\sum_{j=1}^{J} \|\psi_j\|_X}. \tag{4.18}$$

This fixes the ε, η and J in the claim of the theorem. Further, we use the notation

$$\mathcal{U}_J := \{z \in \mathcal{U} : z_j = 0,\ j > J\}. \tag{4.19}$$

Since $\mathcal{A} \subset \mathcal{R}$, for any $a \in \mathcal{A}$ there exists a $z \in \mathcal{U}$ such that

$$a = \sum_{j=1}^{J} z_j\psi_j + \sum_{j>J} z_j\psi_j =: a_J + (a - a_J). \tag{4.20}$$

Note that this decomposition may not be unique – since the ψ_j are not assumed to be linearly independent – but for each $a \in \mathcal{A}$ we assign one such decomposition. We can find a finite set $F \subset \mathcal{U}_J$, such that, for each $z \in \mathcal{U}_J$, there is a $z' \in F$ such that

$$\|z - z'\|_{\ell^\infty(\mathbb{N})} \leq \eta. \tag{4.21}$$

We let $\{\overline{a}_1, \ldots, \overline{a}_M\}$ be the finite set consisting of all elements in X of the form

$$\overline{a}_i = \sum_{j=1}^{J} z'_j\psi_j, \tag{4.22}$$

where $z' \in F$, and in addition there is an

$$a = \sum_{j=1}^{\infty} z_j\psi_j \in \mathcal{A},$$

such that

$$|z_j - z'_j| \leq \eta, \quad j = 1, \ldots, J. \tag{4.23}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Let us now show (i). If $a \in \mathcal{A}$ and $a = \sum_{j=1}^{\infty} z_j \psi_j$, then according to (4.21) and (4.23) there exists $\overline{a}_i$ such that

$$a_J - \overline{a}_i = \sum_{j=1}^{J} c_j \psi_j, \quad |c_j| \le \eta, \tag{4.24}$$

which implies that $a \in \mathcal{A}_i$.

Next, note that (iii) implies (ii). Indeed, take any ρ satisfying the assumptions of (iii). Then $\mathcal{U} \subset \mathcal{O}_\rho$, and hence the validity of (iii) implies $\mathcal{A}_i \subset a_i(\mathcal{O}_\rho) \subset \mathcal{O}$ for each $i = 1, \dots, M$.

We are left to prove (iii). For this, let ρ be any sequence satisfying the constraint in (iii), and define for each $j \ge 1$ the sets

$$\mathcal{O}_{\rho_j} := \{|z_j| < \tilde{\rho}_j\}, \quad \tilde{\rho}_j := \rho_j + \frac{\varepsilon}{\sum_{j \ge 1} \|\tilde{\psi}_j\|_X}. \tag{4.25}$$

We need to check that $a_i(\mathcal{O}_\rho) \subset \mathcal{O}$, $i = 1, 2, \dots, M$. For this, we fix i. We know that $\overline{a}_i = \sum_{j=1}^{J} z'_j \psi_j$, and from (4.23) there is an

$$a^* = \sum_{j=1}^{\infty} z^*_j \psi_j \in \mathcal{A}$$

for which $|z'_j - z^*_j| \le \eta$ for $j = 1, \dots, J$. In view of (4.17) and the definition of η, we have

$$\|\overline{a}_i - a^*\|_X \le \frac{\varepsilon}{2}. \tag{4.26}$$

Now take any $a \in a_i(\mathcal{O}_\rho)$, that is,

$$a = \overline{a}_i + \sum_{j \ge 1} z_j \tilde{\psi}_j, \tag{4.27}$$

with $z = (z_j)_{j \ge 1} \in \mathcal{U}_\rho$. We define

$$\tilde{z}_j = z_j \min\{1, |z_j|^{-1}\}, \quad j \ge 1, \tag{4.28}$$

so that $(\tilde{z}_j)_{j \ge 1}$ is a point in $\mathcal{U}$. We can now estimate

$$\begin{aligned} \|a - a^*\|_X &\le \|a - \overline{a}_i\|_X + \|\overline{a}_i - a^*\|_X \\ &\le \Big\| \sum_{j \ge 1} z_j \tilde{\psi}_j \Big\|_X + \frac{\varepsilon}{2} \\ &\le \Big\| \sum_{j=1}^{J} \tilde{z}_j \tilde{\psi}_j \Big\|_X + \Big\| \sum_{j > J} \tilde{z}_j \tilde{\psi}_j \Big\|_X + \Big\| \sum_{j \ge 1} (\tilde{z}_j - z_j) \tilde{\psi}_j \Big\|_X + \frac{\varepsilon}{2} \end{aligned}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$$\leq \eta \sum_{j=1}^{J} \|\psi_j\|_X + \sum_{j>J} \|\psi_J\|_X + \left\| \sum_{j\geq 1} (\tilde{z}_j - z_j)\tilde{\psi}_j \right\|_X + \frac{\varepsilon}{2}$$
$$\leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \left\| \sum_{j\geq 1} (\tilde{z}_j - z_j)\tilde{\psi}_j \right\|_X + \frac{\varepsilon}{2}.$$

Since

$$|z_j - \tilde{z}_j| \leq (\tilde{\rho}_j - 1) \leq \tilde{\rho}_j - \rho_j + \rho_j - 1, \quad j \geq 1,$$

we obtain

$$\|a - a^*\|_X \leq \varepsilon + \sum_{j\geq 1} (\tilde{\rho}_j - \rho_j)\|\tilde{\psi}_j\|_X + \sum_{j\geq 1} (\rho_j - 1)\|\tilde{\psi}_j\|_X \leq \varepsilon + \varepsilon + \varepsilon = 3\varepsilon,$$

where we have used (4.25) to bound the first sum and the assumption on $(\rho_j)_{j\geq 1}$ in estimating the second sum. This shows that a belongs to the 3ε-neighbourhood of $\mathcal{A}$, which is contained in $\mathcal{O}$. Therefore $a_i(\mathcal{U}_\rho) \subset \mathcal{O}$. □

With the help of the above lemma, we now establish a result which shows that the conclusion of Corollary 4.1 remains valid without the assumption that $\mathcal{A}$ is of the exact form $a(U)$.

Theorem 4.3. For a parametric problem of the form (1.1), assume that the solution map $u \mapsto u(a)$ admits a holomorphic extension over an open set $\mathcal{O}$ of the complex Banach space X which contains $\mathcal{A}$, with uniform bound

$$\sup_{a\in\mathcal{O}} \|u(a)\|_V \leq C. \tag{4.29}$$

Assume in addition that there exist functions $(\psi_j)_{j\geq 1}$ in X such that $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$, and such that $\mathcal{A} \subset \mathcal{R}$, where $\mathcal{R}$ is of the form (4.10). Then there exists $C > 0$ such that

$$d_n(\mathcal{M})_V \leq Cn^{-s}, \quad n \geq 1, \quad s := \frac{1}{p} - 1. \tag{4.30}$$

Proof. Applying Lemma 4.2, we write

$$\mathcal{A} \subset \bigcup_{i=1}^{M} \mathcal{A}_i, \tag{4.31}$$

and therefore

$$\mathcal{M} \subset \bigcup_{i=1}^{M} \mathcal{M}_i, \quad \mathcal{M}_i := u(\mathcal{A}_i). \tag{4.32}$$

It now sufficient to prove that the estimate (4.30) holds for each $\mathcal{M}_i$ in place of $\mathcal{M}$, that is,

$$d_n(\mathcal{M}_i)_V \leq Cn^{-s}, \quad n \geq 1, \quad s := \frac{1}{p} - 1. \tag{4.33}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Indeed, if for each $i = 1, \dots, M$ one can approximate all elements of $\mathcal{M}_i$ with accuracy δ by elements from an n-dimensional space $V_{n,i}$, then one can approximate all elements of $\mathcal{M}$ with the same accuracy by elements from the space $V_{n,1} \oplus \cdots \oplus V_{n,M}$ which has at most dimension nM. This shows that

$$d_{Mn}(\mathcal{M})_V \leq \max_{i=1,\dots,M} d_n(\mathcal{M}_i)_V, \tag{4.34}$$

and therefore (4.33) implies (4.30) up to a change in the constant C.

The proof of (4.33) follows from Corollary 3.10. We fix $i \in \{1, \dots, M\}$. We know that

$$\mathcal{A}_i \subset a_i(\mathcal{U}), \tag{4.35}$$

where

$$a_i(z) := \overline{a}_i + \sum_{j \geq 1} z_j \tilde{\psi}_j.$$

It follows that the assumptions of this corollary are satisfied for $\mathcal{A}_i$ and a_i in place of $\mathcal{A}$ and a. This confirms the estimate (4.33) and therefore concludes the proof. □

Remark 4.4. The above theorem can be formulated for a general map u from $\mathcal{A}$ to V that is not necessarily the solution map of a parametric PDE, following the arguments from Remarks 2.12 and 3.13.

4.3. n-widths under holomorphic maps: a general result

In this subsection we let u be any map from $\mathcal{A}$ to V, not necessarily the solution map to a parametric PDE. In view of Remark 4.4, Theorem 4.3 gives an estimate for the n-widths of $\mathcal{M} = u(\mathcal{A})$ whenever u has a bounded holomorphic extension to a neighbourhood of $\mathcal{A}$, provided that $\mathcal{A}$ is contained in a set $\mathcal{R}$ of the form (4.10) with $(\|\psi_j\|_X)_{j \geq 1} \in \ell^p$, with $0 < p < 1$. Note that the containment and ℓ^p-summability assumptions on $\mathcal{A}$ imply a decay of the n-width of $\mathcal{A}$ in X. Namely, for $s := 1 - \frac{1}{p}$, we can write

$$d_n(\mathcal{A})_X \leq \sup_{a \in \mathcal{A}} \min_{z \in \mathcal{U}} \left\| a - \sum_{j=1}^{n} z_j \psi_j \right\|_X \leq \sum_{j > n} \|\psi_j\|_X \leq C(n+1)^{-s}, \quad n \geq 0, \tag{4.36}$$

where we have used Lemma 3.6. Thus, in a certain sense, the results of the previous subsection can be interpreted as saying that the Kolmogorov widths of $\mathcal{M}$ inherit the decay rate of the widths of $\mathcal{A}$. Since the sets $\mathcal{A}$ are generally more accessible and their n-widths are more readily computed, it is natural to ask whether there is a general principle in effect here. That is, do the n-widths $(d_n(\mathcal{M})_V)_{n \geq 1}$ of an image $\mathcal{M} = u(\mathcal{A})$ of a compact set $\mathcal{A}$ under a general holomorphic map u have the same decay as that of

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$(d_n(\mathcal{A})_X)_{n\geq 1}$. The main goal of this subsection is to show that there is indeed such a general principle in effect, but with a slight loss in the decay rate of the widths of $\mathcal{M}$ when compared with those of $\mathcal{A}$.

Our first step in deriving such comparison results is to show that whenever a compact set $\mathcal{A}$ of a Banach space X has widths $(d_n(\mathcal{A})_X)$ with some prescribed decay, then $\mathcal{A}$ is contained in a set $\mathcal{R}$ of the form (4.10), where the X-norms of the ψ_j defining $\mathcal{R}$ are ℓ^p-summable for certain values of p. For this, we need the following classical result due to Auerbach, the proof of which is given below for completeness.

Lemma 4.5. Let E be an n-dimensional subspace of a complex Banach space X. Then there exists a basis $\{\varphi_1, \dots, \varphi_n\}$ for E and a dual basis $\{\tilde{\varphi}_1, \dots, \tilde{\varphi}_n\}$ in X' such that

$$\langle \tilde{\varphi}_i, \varphi_j \rangle_{X',X} = \delta_{i,j}, \quad i,j = 1, \dots, n, \tag{4.37}$$

and

$$\|\varphi_i\|_X = \|\tilde{\varphi}_i\|_{X'} = 1, \quad i = 1, \dots, n. \tag{4.38}$$

Proof. We start with an arbitrary basis $\psi_1, \dots, \psi_n$ of E and let $\tilde{\psi}_1, \dots, \tilde{\psi}_n$ be its dual basis in E', that is,

$$\langle \tilde{\psi}_i, \psi_j \rangle = \delta_{i,j}, \tag{4.39}$$

where δ is the Kronecker delta and $\langle \cdot, \cdot \rangle := \langle \cdot, \cdot \rangle_{X',X}$ throughout this proof. Then, any $f \in E$ can be uniquely written as

$$f = \sum_{i=1}^{n} \langle \tilde{\psi}_i, f \rangle \psi_i. \tag{4.40}$$

Given any $(g_1, \dots, g_n) \in E^n$, we define

$$J(g_1, \dots, g_n) = |\det(M(g_1, \dots, g_n))|, \quad M := (\langle \tilde{\psi}_i, g_j \rangle)_{i,j=1,\dots,n}. \tag{4.41}$$

We now take $(\varphi_1, \dots, \varphi_n) \in E^n$ such that

$$(\varphi_1, \dots, \varphi_n) := \arg\max J(g_1, \dots, g_n), \tag{4.42}$$

where the maximum is taken over all $(g_1, \dots, g_n) \in E^n$ such that $\|g_i\|_X = 1$, for $i = 1, \dots, n$. This maximum is attained since the function J is continuous and we are maximizing over a compact set. The functions $\varphi_1, \dots, \varphi_n$ are linearly independent since this maximum is positive. Hence they form a basis for E, and any $f \in E$ can be written uniquely as

$$f = \sum_{i=1}^{n} \langle \tilde{\varphi}_i, f \rangle \varphi_i, \tag{4.43}$$

where $\tilde{\varphi}_i$, $i = 1, \dots, n$, is its dual basis. Applying the functional $\tilde{\psi}_i$ to both

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

sides of (4.43), we obtain

$$\sum_{j=1}^{n} \langle \tilde{\varphi}_j, f\rangle \langle \tilde{\psi}_i, \varphi_j\rangle = \langle \tilde{\psi}_i, f\rangle, \quad i = 1, 2, \dots, n. \tag{4.44}$$

From Cramer's rule, it follows that, for any $j \in \{1, \dots, n\}$ and any $f \in E$,

$$|\langle \tilde{\varphi}_j, f\rangle| = \frac{J(\varphi_1, \dots, \varphi_{j-1}, f, \varphi_{j+1}, \dots, \varphi_n)}{J(\varphi_1, \dots, \varphi_n)} \le 1. \tag{4.45}$$

This proves that for each j we have $\|\tilde{\varphi}_j\|_{E'} = 1$. By application of the Hahn–Banach theorem, we can extend $\tilde{\varphi}_j$ over all of X with $\|\tilde{\varphi}_j\|_{X'} = 1$. □

Using Auerbach's lemma, we now show that whenever $\mathcal{A}$ is a compact set of a complex Banach space X and $d_n(\mathcal{A})_X$ has some prescribed rate of decay, then $\mathcal{A}$ is contained in a set $\mathcal{R}$ of the form (4.10) and the X-norms of the ψ_j defining $\mathcal{R}$ have the same rate of decay as $d_n(\mathcal{A})_X$.

Lemma 4.6. Let X be a complex Banach space and $\mathcal{A} \subset X$ be a compact space such that

$$\sup_{n\ge 1} n^s d_n(\mathcal{A})_X < \infty. \tag{4.46}$$

Then there exists a family $(\psi_j)_{j\ge 1}$ of functions from X such that

$$\sup_{j\ge 1} j^s \|\psi_j\|_X < \infty \tag{4.47}$$

and

$$\mathcal{A} \subset \mathcal{R} := \left\{ \sum_{j\ge 1} z_j \psi_j : z = (z_j)_{j\ge 1} \in \mathcal{U} \right\}. \tag{4.48}$$

Proof. From (4.46), we know that there exists a constant $C > 0$ and a sequence of spaces $(V_k)_{k\ge 0}$ with $V_k \subset X$ and $\dim(V_k) = 2^k$ such that

$$\max_{a\in\mathcal{A}} \min_{g\in X_k} \|a - g\|_X \le C 2^{-sk}, \quad k \ge 0. \tag{4.49}$$

By replacing V_k with $V_0 + V_1 + \cdots + V_{k-1}$ and possibly changing the constant C, we may assume that the spaces V_k are nested: $V_{k-1} \subset V_k$, for all $k \ge 1$.

For any $a \in \mathcal{A}$, we let a_k denote a best approximation to a from V_k for $k \ge 0$ and set $a_{-1} := 0$. Then, $g_k := a_k - a_{k-1}$ is in V_k, and we have

$$a = \sum_{k\ge 0} g_k. \tag{4.50}$$

In addition, there exists a constant $C > 0$ such that

$$\|g_k\|_X \le C 2^{-sk}, \quad k \ge 0. \tag{4.51}$$

By Auerbach's lemma, for every $k \ge 0$, there exists a basis $\{\varphi_{k,l}\}_{l=1,\dots,2^k}$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

of the space V_k, and a dual basis $\{\tilde{\varphi}_{k,l}\}_{l=1,\dots,2^k} \subset X'$ such that $\|\varphi_{k,l}\|_X = \|\tilde{\varphi}_{k,l}\|_{X'} = 1$. It follows that any $a \in \mathcal{A}$ can be written as

$$a = \sum_{k\geq 0} \sum_{l=1}^{2^k} z_{k,l}\varphi_{k,l}, \quad |z_{k,l}| \leq C2^{-sk}. \tag{4.52}$$

Each integer $j \geq 1$ can be written uniquely as $j = 2^k + l - 1$ with $l \in \{1, \dots, 2^k\}$. We use this to define

$$\psi_j := C2^{-sk}\varphi_{k,l}, \quad j = 2^k + l - 1. \tag{4.53}$$

This gives that any $a \in \mathcal{A}$ is of the form

$$a = \sum_{j\geq 1} z_j\psi_j, \quad |z_j| \leq 1, \tag{4.54}$$

that is, $\mathcal{A} \subset \mathcal{R}$. In addition, we have

$$\|\psi_j\|_X \leq Cj^{-s}, \tag{4.55}$$

up to a change in the constant C. □

An immediate consequence of the above lemma is that if $d_n(\mathcal{A})_X$ has the rate of decay n^{-s}, then $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for any p such that $sp > 1$. Combining this observation with Theorem 4.3 leads to the following result, which shows that the rate of decay of the n-width is almost preserved under holomorphic maps, up to a loss of 1 in the rate.

Theorem 4.7. For a pair of complex Banach spaces X and V, assume that u is a general holomorphic map from an open set $\mathcal{O} \subset X$ into V with uniform bound

$$\sup_{a\in\mathcal{O}} \|u(a)\|_V \leq C. \tag{4.56}$$

If $\mathcal{A} \subset \mathcal{O}$ is a compact subset of X and $\mathcal{M} = u(\mathcal{A})$, then for any $s > 1$ and $t < s - 1$,

$$\sup_{n\geq 1} n^s d_n(\mathcal{A})_X < \infty \Rightarrow \sup_{n\geq 1} n^t d_n(\mathcal{M})_V < \infty. \tag{4.57}$$

Some comments on this result are in order. If u was a linear map, we could write, for any subspace $X_n \subset X$ of dimension n and any $a \in \mathcal{A}$,

$$\min_{v\in V_n} \|u(a) - v\|_V \leq C \min_{\tilde{a}\in X_n} \|a - \tilde{a}\|_X, \quad C := \|u\|_{\mathcal{L}(X,V)}, \tag{4.58}$$

with $V_n := u(X_n) \subset V$ also of dimension n. Therefore, we would obtain

$$d_n(\mathcal{M})_V \leq Cd_n(\mathcal{A})_X, \tag{4.59}$$

which implies that $d_n(\mathcal{M})_V$ has at least the same rate of decay as $d_n(\mathcal{A})_X$. Theorem 4.7 shows that holomorphic maps behave almost as well as linear

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

maps, except for the loss of 1 in the rate expressed by the inequality $t < s-1$. This loss occurs due to a lack of sharpness in Lemma 4.6: if we start from the conclusion $\|\psi_j\|_X \leq Cj^{-s}$ of this lemma, we may only retrieve that

$$d_n(\mathcal{A})_X \leq d_n(\mathcal{R})_X \leq \sum_{j>n} \|\psi_j\|_X \leq Cn^{1-s}. \tag{4.60}$$

An open question is whether the implication (4.57) in Theorem 4.7 remains valid with $t = s$.

4.4. Towards faster low-rank approximations

We close Part I of this article with some remarks concerning our approximation results. As explained in Section 1, our general interest is in the accuracy of separable approximations of the form (1.32) and (1.33). These approximations can be thought of as the analogue of low-rank approximations for finite-dimensional matrices.

Optimal approximations arise via best optimal n-dimensional spaces V_n either in the sense of n-widths for uniform approximation or Karhunen–Loève decompositions for approximation in the mean square sense. Since these spaces are out of reach, both from a theoretical and a computational point of view, we build sub-optimal approximations $y \mapsto u_n(y)$ based on best n-term truncations of polynomial expansions. This approach leads us to quantitative convergence results such as in Corollaries 3.10 and 3.11, and in turn to estimates for the decay of the n-widths $d_n(\mathcal{M})_V$ of solution manifolds as discussed in Section 4, for example by using the estimate

$$d_n(\mathcal{M})_V \leq \|u - u_n\|_{L^\infty(U,V)} \tag{4.61}$$

in the case when Assumption A holds.

A legitimate question is to evaluate the possible lack of optimality of the convergence rates, obtained by our polynomial approximation approach, in comparison to the rates which could be achieved by using the optimal n-dimensional spaces V_n. Equivalently, we would like to know if the rate of decay of the n-width $d_n(\mathcal{M})_V$ could sometimes be much faster than the rate of decay of the polynomial approximation error on the right-hand side of (4.61).

We can give simple examples which reveal this lack of optimality in the case of the elliptic equation (1.5). Here, we consider the finite-dimensional setting, where

$$a(y) = \overline{a} + \sum_{j=1}^{d} y_j \psi_j. \tag{4.62}$$

In this setting, convergence rates of exponential type, that is,

$$\|u - u_n\|_{L^\infty(U,V)} \leq C\exp(-cn^{1/d}), \tag{4.63}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

are established in Section 3.9, using for u_n the Taylor series truncated with the index set corresponding to the n largest values of $\|t_\nu\|_V$.

Let us now consider the particular case of a piecewise constant diffusion coefficient of the form

$$a(y) = \sum_{j=1}^{d} (1 + \theta y_j) \chi_{D_j}, \tag{4.64}$$

where $\{D_1, \ldots, D_j\}$ is a partition of D, that is, $\overline{a} = 1$ and $\psi_j = \theta \chi_{D_j}$. We assume that $0 < \theta = 1 - r < 1$ so that UEA(r) holds.

We first examine the case where D is a one-dimensional interval partitioned into sub-interval D_j. The problem now reads

$$-(au')' = f, \tag{4.65}$$

with homogeneous Dirichlet boundary conditions at the endpoints of D. Since $a(y)$ is constant on each interval D_j, we find that the restriction of $u(y)$ to this interval is always the sum of an affine function and of a scalar multiple of F such that $F'' = f$. It follows that, for any $y \in [-1, 1]^d$, the solution $u(y)$ belongs to the finite-dimensional space

$$V_{3d} = \text{span}\{\chi_{D_i},\ x\chi_{D_i},\ F\chi_{D_i} : i = 1, \ldots, d\}, \tag{4.66}$$

where x stands for the identity function $x \mapsto x$. Using the fact that $u(y)$ is 0 at the endpoints of D and continuous at the breakpoint between the D_j, we find that it belongs to an even smaller subspace of V_{3d} that has dimension at most $2d - 1$. This implies that

$$d_n(\mathcal{M})_V = 0, \tag{4.67}$$

for $n \geq 2d - 1$, therefore showing that the rate on the right-hand side of (4.63) is not sharp for $d_n(\mathcal{M})_V$.

Let us now examine a less trivial case where D is a domain in higher dimension $m \geq 2$. In such a case, it is not true that $\mathcal{M}$ belongs to a finite-dimensional space, but we can still show that the rate on the right-hand side of (4.63) is not sharp for $d_n(\mathcal{M})_V$. For simplicity, consider the case of a partition into two domains, that is, $d = 2$. Since $\|\psi_1\|_X = \|\psi_2\|_X = \theta$, the sets Λ_n that are used in Section 3.9 to obtain the rate

$$\|u - u_n\|_{L^\infty(U,V)} \leq C \exp(-cn^{1/2}) \tag{4.68}$$

have the simple structure

$$\Lambda_n = \{|\nu| = \nu_1 + \nu_2 \leq k\}, \tag{4.69}$$

for integers $k \geq 0$. Therefore, we use polynomial approximations of total degree k of the form

$$u_n(y) = \sum_{|\nu| \leq k} t_\nu y^\nu, \tag{4.70}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

which have accuracy

$$\|u - u_n\|_{L^\infty(U,V)} \leq C\exp(-ck), \tag{4.71}$$

with

$$n = \frac{k(k+1)}{2} \sim k^2.$$

This truncated power series can be interpreted in a different way by writing the elliptic equation in operator form,

$$\mathcal{B}(y)u(y) = f, \quad \mathcal{B}(y) = \overline{\mathcal{B}} + y_1\mathcal{B}_1 + y_2\mathcal{B}_2, \tag{4.72}$$

where

$$\overline{\mathcal{B}}u := -\Delta u \quad \text{and} \quad \mathcal{B}_j u := -\mathrm{div}(\theta\chi_{D_j}\nabla u). \tag{4.73}$$

With $\overline{\mathcal{B}}^{-1}$ the inverse of $-\Delta$ on D with homogeneous Dirichlet boundary condition, we may then rewrite the equation as

$$(I + y_1\tilde{\mathcal{B}}_1 + y_2\tilde{\mathcal{B}}_2)u(y) = g, \quad g := \overline{\mathcal{B}}^{-1}f, \quad \tilde{\mathcal{B}}_j := \overline{\mathcal{B}}^{-1}\mathcal{B}_j, \ j = 1, 2. \tag{4.74}$$

It is easily seen that the Taylor series of $u(y)$ coincides with the Neumann series

$$u(y) = \sum_{l\geq 0}(-1)^l(y_1\tilde{\mathcal{B}}_1 + y_2\tilde{\mathcal{B}}_2)^l g. \tag{4.75}$$

The convergence of this series can be directly checked by observing that

$$\|y_1\tilde{\mathcal{B}}_1 + y_2\tilde{\mathcal{B}}_2\|_{\mathcal{L}(V,V)} \leq \theta, \quad (y_1, y_2) \in [-1, 1]^2. \tag{4.76}$$

In particular, this confirms the exponential rate

$$\|u - u_n\|_{L^\infty(U,V)} \leq C\theta^k = C\exp(-ck). \tag{4.77}$$

We now observe that, due to the fact that $\chi_{D_1} + \chi_{D_2} = \chi_D$, we have the identity

$$\tilde{\mathcal{B}}_1 + \tilde{\mathcal{B}}_2 = \theta I. \tag{4.78}$$

We may therefore rewrite each term in the Neumann series as

$$\begin{aligned}(-1)^l(y_1\tilde{\mathcal{B}}_1 + y_2\tilde{\mathcal{B}}_2)^l g &= (-1)^l(y_2\theta I + (y_1 - y_2)\tilde{\mathcal{B}}_1)^l g \\ &= (-1)^l\sum_{j=0}^{l}(y_2\theta)^{l-j}\binom{l}{j}(y_1 - y_2)^j\tilde{\mathcal{B}}_1^j g.\end{aligned} \tag{4.79}$$

Therefore, summing the terms in (4.79) from $l = 0$ up to k, we may rewrite $u_n(y)$ as

$$u_n(y) = \sum_{j=0}^{k} v_j\phi_j(y), \tag{4.80}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

with

$$v_j := \tilde{\mathcal{B}}_1^j g \in V, \tag{4.81}$$

and

$$\phi_j(y) := (y_1 - y_2)^j \sum_{l=j}^{k} (-1)^l (y_2\theta)^{l-j} \binom{l}{j}. \tag{4.82}$$

This new representation of $u_n(y)$ shows that it belongs to the $k+1$-dimensional space

$$V_k = \operatorname{span}\{v_0, \ldots, v_k\}. \tag{4.83}$$

We may thus conclude that

$$d_{k+1}(\mathcal{M})_V \leq C\theta^k = C\exp(-ck), \tag{4.84}$$

Since $k \sim \sqrt{n}$, this shows that the rate in the right-hand side of (4.68) is not sharp for $d_n(\mathcal{M})_V$.

These examples reveal that in certain relevant cases, polynomial approximations based on best n-term truncations may be highly sub-optimal in comparison to the n-width spaces. Note, however, that the rank reduction is made possible due to fine properties of the affine representation (1.15), such as the fact that the ψ_j are characteristic functions with disjoint supports. For other affine representations with general functions ψ_j which have overlapping support, numerical computations show that polynomial approximation rates are sometimes close to the optimal rates to be expected from arbitrary separable approximations. The development of alternative strategies for a sharper convergence analysis of separable approximations is thus desirable, and it inevitably requires exploitation of the detailed structure of the affine representation.

PART II
Algorithms for parametric PDEs

5. Towards concrete algorithms

The results exposed in Part I of this paper show that relevant instances of parametric PDEs admit separable approximations u_n of the form (1.32) or (1.33) with error bounds that reflect a certain rate of convergence in terms of the number n of terms that are retained.

However, these approximations are obtained by mathematical techniques which, as such, cannot be implemented via a computational algorithm. For example, in order to compute the best n-term truncation of the Legendre series, we need in principle to be able to compute exactly all Legendre

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

coefficients w_ν and to search for the n largest values of $\|w_\nu\|_V$. This is unfeasible for two reasons: (i) we can only compute the w_ν with limited precision due to spatial discretization, for example through a finite element space V_h of V, and (ii) we cannot perform an exhaustive search over the infinite set $\mathcal{F}$ of multi-indices.

In Part II of the paper we discuss concrete numerical methods which compute separable approximations u_n, still of the form (1.32) or (1.33), but at an affordable computational cost.

5.1. Space discretization and computational cost

Our approach to the computation of such approximations can be viewed as follows.

(i) We develop and analyse strategies for computing separable expansions initially based on a few instances of the exact solution maps $a \mapsto u(a)$ and $y \mapsto u(y)$, or quantities related to these maps such as the Taylor coefficients t_ν.

(ii) We then instead apply these strategies to the *approximate solution maps*

$$a \mapsto u_h(a) \in V_h \quad \text{and} \quad y \mapsto u_h(y) \in V_h, \tag{5.1}$$

which correspond to a certain space discretization process for each instance of the solution map in a fixed discretization space V_h.

Ideally, we would like to obtain error bounds for these approximations which meet the benchmark established in Part I of the paper in terms of their decay as n grows, up to an additional term that reflects the space discretization error.

We assume that space discretization can be performed within a certain finite element space V_h of dimension N_h, through a numerical solver which we may apply for each individual instance of $a \in \mathcal{A}$ or $y \in U_\mathcal{A}$ to compute approximate solutions $u_h(a)$ or $u_h(y)$ from V_h. For simplicity we assume:

(i) a cost C_h for computing $u_h(a)$ or $u_h(y)$ that is independent of a or y;

(ii) an error bound

$$\sup_{a\in\mathcal{A}} \|u(a) - u_h(a)\|_V = \sup_{y\in U_\mathcal{A}} \|u(y) - u_h(y)\|_V \le \varepsilon(h), \tag{5.2}$$

therefore also independent of a or y.

Recall that making $\varepsilon(h)$ small requires us to make N_h large and C_h even larger, which is one of the motivations for reduced modelling.

As an example of such a space discretization, consider the elliptic equation (1.5). The discrete solution can be defined by the standard Galerkin method

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

on V_h, that is, $u_h(a) \in V_h$ is such that

$$\int_D a \nabla u_h(a) \nabla v_h = \int_D f v_h, \quad v_h \in V_h. \tag{5.3}$$

We may then use classical techniques of finite element approximation of elliptic PDEs (see Ciarlet 1978 or Brenner and Scott 2008), in order to obtain an error bound of the form (5.2). First, assuming that $0 < r \leq a \leq R$ for all $a \in \mathcal{A}$, Cea's lemma ensures that

$$\|u(a) - u_h(a)\|_V \leq \sqrt{\frac{R}{r}} \min_{v_h \in V_h} \|u(a) - v_h\|_V. \tag{5.4}$$

Then, if $(V_h)_{h>0}$ are Lagrange finite element spaces of polynomial degree $m \geq 1$ associated to a regular family of conforming simplicial partitions $(\mathcal{T}_h)_{h>0}$ with mesh size $h > 0$, we have for $1 < r \leq m+1$ the classical approximation bound

$$\min_{v_h \in V_h} \|u(a) - v_h\|_V \leq C h^{r-1} \|u(a)\|_{H^r(D)}. \tag{5.5}$$

We therefore obtain an error bound (5.2) with $\varepsilon(h) \sim h^{r-1}$ provided that $u(a)$ is bounded in $H^r(D)$ independently of $a \in \mathcal{A}$.

Remark 5.1. Our approach to space discretization means in particular that, when computing polynomial approximations by truncated expansions, the Taylor or Legendre coefficients are discretized in the *same* finite element space V_h, independently of their index ν. An alternative approach, which we do not embark on here, is to search for space discretizations of these coefficients which *vary* with ν, with the objective of optimizing the total number of degrees of freedom required to reach a given accuracy. This approach is analysed in Cohen *et al.* (2010, 2011) for Legendre and Taylor series. See also Gittelson (2013) for computational approaches based on a global adaptivity both in the parameter and space variable.

When evaluating the total computational cost for computing the separable approximation u_n, we make the distinction between two types of cost.

(i) The *offline cost* refers to the computation of the functions $v_1, \ldots, v_n$ which are used in (1.32) or (1.33), or equivalently of the space $V_n := \operatorname{span}\{v_1, \ldots, v_n\}$ which is used to simultaneously approximate all members of the solution manifold $\mathcal{M}$.

(ii) The *online cost* refers to the computation of the approximate solution $u_n(a)$ or $u_n(y)$ from V_n for any given query $a \in \mathcal{A}$ or $y \in U$.

One can view the offline cost as a 'one time only' fixed cost, while the online cost could be repeated many times in certain applications of reduced modelling.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

5.2. Polynomial approximation algorithms

The first class of numerical methods that we study searches for computable polynomial approximations of the general form (1.50). For any finite set Λ, we define the space

$$V_\Lambda := V \otimes \mathbb{P}_\Lambda, \tag{5.6}$$

of V-valued polynomials associated to Λ, where

$$\mathbb{P}_\Lambda := \operatorname{span}\{y \mapsto y^\nu : \nu \in \Lambda\} \tag{5.7}$$

is the corresponding space of real-valued polynomials. Therefore a polynomial approximation of the form (1.50) belongs to V_{Λ_n}.

There are two main issues in the design of these methods.

(i) Given an index set Λ_n, how do we construct the polynomial approximation (1.50)?

(ii) How do we select the index sets Λ_n?

For treating both of these issues, it is very useful to require the considered sets Λ_n to be *downward closed*, which we assume in going further.

Concerning the first issue, we present two different strategies which illustrate the important distinction between *non-intrusive* and *intrusive* methods, already mentioned in Section 1.6.

The first strategy, discussed in Section 5, is non-intrusive. It computes a polynomial approximation of the form (1.50) by *interpolation* of the solution map at well-chosen points $y^1, \dots, y^n \in U$ by a method introduced in Chkifa *et al.* (2014), along the lines of Nobile, Tempone and Webster (2008*a*, 2008*b*). In particular, it could even be applied in a context where the exact model is not known but only the solver is given. Other important representatives of non-intrusive methods, which we do not discuss in this paper, include *least-squares* projection methods, as developed in Chkifa *et al.* (2015*a*), Doostan and Iaccarino (2009) and Doostan and Owadi (2011), and *pseudo-spectral* methods, as developed in Xiu (2007) and Constantine *et al.* (2012).

The second strategy, discussed in Section 6, performs an explicit computation of the truncated Taylor series, up to the spatial discretization of the coefficients t_ν, by a recursive method introduced in Chkifa *et al.* (2013). In contrast to the previous one, this approach is intrusive. It strongly exploits the particular form of the parametric PDE, and in fact it can only be easily implemented for parametric problems (1.1) where $\mathcal{P}$ is linear both in u and a. Other important representatives of intrusive methods, which we do not discuss in this paper, include Galerkin projection methods as developed in Beck *et al.* (2012, 2014), Cohen *et al.* (2010) and Gittelson (2013).

Concerning the second issue, an important distinction should be made between *non-adaptive* and *adaptive* methods. In non-adaptive methods, the selection of the set Λ_n for a given value of n is done in an *a priori* manner,

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

based on available information on the problem. Ideally we would like to use the set Λ_n associated to the n largest coefficients in the Taylor or Legendre expansion, but this set cannot be easily identified. Instead, we consider the set Λ_n associated to the n largest *a priori* estimates obtained in Section 3 for the V-norms of these coefficients. We detail further in Section 5.3 the algorithmic construction of the sequence $(\Lambda_n)_{n\geq 1}$ by this approach.

In adaptive methods, the selection of Λ_n is made in an *a posteriori* manner, based on the computation for downward closed values of n, for instance using knowledge of both the previous choice Λ_{n-1} and the computed approximation polynomial u_{n-1} for this choice. One reason why adaptive methods might perform significantly better than their above-described non-adaptive counterpart in the present context is because the *a priori* bound e_ν may lack sharpness and therefore gives only a limited indication as to the real set of the n largest coefficients. In particular, the guaranteed rate n^{-s} based on these *a priori* bounds may be too pessimistic, and a better rate could be obtained using an adaptive method. However, the convergence analysis of adaptive methods is usually much more delicate than that of their non-adaptive counterparts. We give examples of adaptive strategies both for interpolation in Section 5 and Taylor expansions in Section 6, convergence analysis being available only for the latter.

5.3. Non-adaptive constructions of the sets Λ_n

We recall that the *a priori* estimates obtained in Section 3 for the Taylor or Legendre coefficients have the following general form.

- For the Taylor coefficients, under the assumptions of Theorem 2.8,

$$\|t_\nu\|_V \leq Ce_\nu, \quad e_\nu := \prod_{j\in \mathrm{supp}(\nu)} \rho_j^{-\nu_j}, \tag{5.8}$$

  for any given sequence $\rho = (\rho_j)_{j\geq 1}$ of number larger than 1 that satisfies the constraint (3.48).

- For the Legendre coefficients, under the assumptions of Theorem 2.9,

$$\|w_\nu\|_V \leq Ce_\nu, \quad e_\nu := C \prod_{j\in \mathrm{supp}(\nu)} \theta(\rho_j)(1+2\nu_j)\rho_j^{-\nu_j}, \tag{5.9}$$

  for any given sequence $\rho = (\rho_j)_{j\geq 1}$ of number strictly larger than 1 that satisfies the constraint (3.48).

Also recall that for certain specific problems, we can sharpen these estimates by improving on the constraint (3.48) imposed on ρ; see Section 3.4. Once an admissible sequence $\rho = \rho(\nu)$ has been fixed for each ν, each resulting estimate e_ν is computable as a product of $\|\nu\|_0$ numbers. In the proof of Theorem 3.9 we use particular choices of admissible sequences $\rho = \rho(\nu)$, which

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

ensures the ℓ^p-summability of the resulting e_ν provided that $(\|\psi_j\|_X)_{j\geq 1}$ is ℓ^p-summable, for some $p < 1$. However, one may hope to further improve the estimate e_ν by using other sequences.

An important observation is that the above general definition of e_ν does not guarantee that the set Λ_n corresponding to the n largest e_ν is downward closed. Indeed, we do not know that the sequence $(e_\nu)_{\nu\in\mathcal{F}}$ defined in (5.8) or (5.9) is monotone non-increasing, in particular due to the fact that the sequence ρ is allowed to vary with ν. One may try to construct the sequences $\rho(\nu)$ such that the sequence $(e_\nu)_{\nu\in\mathcal{F}}$ is monotone non-increasing. However, a simpler possibility is instead to search for a surrogate s_ν, with

$$e_\nu \leq s_\nu := \prod_{j\in\mathrm{supp}(\nu)} s_j(\nu), \tag{5.10}$$

where the $s_j(\nu)$ are again explicitly given, and in addition $(s_\nu)_{\nu\in\mathcal{F}}$ is a monotone non-increasing sequence. Then, we know that at least one of the sets Λ_n corresponding to the n largest s_ν is downward closed. One example of such a surrogate in the case of Legendre coefficients is given by $s_\nu := \tilde{r}_\nu$, defined in (3.168), for which ℓ^p-summability is also established provided that $(\|\psi_j\|_X)_{j\geq 1}$ is ℓ^p-summable, for some $p < 1$.

We now discuss the complexity of identifying the downward closed set Λ_n associated to the n largest s_ν. In addition to the monotonicity of $(s_\nu)_{\nu\in\mathcal{F}}$, the following property is useful for limiting this complexity.

Definition 5.2. A monotone non-increasing positive sequence $(s_\nu)_{\nu\in\mathcal{F}}$ is said to be *anchored* if and only if

$$l \leq j \Rightarrow s_{e_j} \leq s_{e_l}, \tag{5.11}$$

where e_l and e_j are the Kronecker sequences with 1 at position l and j, respectively.

This property implies that at least one of the sets Λ_n corresponding to the n largest s_ν has the following property.

Definition 5.3. A finite downward closed set Λ is said to be *anchored* if and only if

$$e_j \in \Lambda \quad \text{and} \quad l \leq j \quad \Rightarrow \quad e_l \in \Lambda. \tag{5.12}$$

where e_l and e_j are the Kronecker sequences with 1 at position l and j, respectively.

We now show that for an anchored sequence $(s_\nu)_{\nu\in\mathcal{F}}$ the identification of the set Λ_n can be executed in at most $n^2/2$ evaluations of s_ν. For this purpose, we introduce for any downward closed set Λ its set of *neighbours* defined by

$$N(\Lambda) := \{\nu \notin \Lambda \text{ such that } \Lambda \cup \{\nu\} \text{ is downward closed}\}. \tag{5.13}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We also introduce the set of its *anchored neighbours* defined by

$$\tilde{N}(\Lambda) := \{\nu \in N(\Lambda) : \nu_j = 0 \text{ if } j > j(\Lambda) + 1\}, \tag{5.14}$$

where

$$j(\Lambda) := \max\{j : \nu_j > 0 \text{ for some } \nu \in \Lambda\}. \tag{5.15}$$

If $(s_\nu)_{\nu\in\mathcal{F}}$ is an anchored sequence, we may define the sets $\Lambda_n = \{\nu^1, \ldots, \nu^n\}$ by the following induction.

- Take $\nu^1 = 0$ to be the null multi-index.
- Given $\Lambda_k = \{\nu^1, \ldots, \nu^k\}$, pick a ν^{k+1} maximizing s_ν over $\nu \in \tilde{N}(\Lambda_k)$ and such that the new set Λ_{k+1} is anchored.

We observe that $\tilde{N}(\Lambda_k)$ is contained in the union of $\tilde{N}(\Lambda_{k-1})$ and the set consisting of the indices

$$e_{j(\Lambda_k)+1} \quad \text{and} \quad \nu^k + e_j, \quad j \le j(\Lambda_k). \tag{5.16}$$

Therefore, since the values of the s_ν have already been computed for $\nu \in \tilde{N}(\Lambda_{k-1})$, step k of the induction requires at most $j(\Lambda_k) + 1$ evaluations of s_ν. In addition, the fact that Λ_k is anchored implies that $j(\Lambda_k) \le k - 1$. Therefore, the total number of evaluations of s_ν in order to reach Λ_n is at most

$$N_n = 1 + 2 + \cdots + (n-1) \le n^2/2. \tag{5.17}$$

Finally, let us observe that the computation of a single s_ν costs $\|\nu\|_0$ multiplications, and on the other hand, for all $\nu \in \Lambda_n$,

$$2^{\|\nu\|_0} \le \prod_{j\in\mathrm{supp}(\nu)} (1 + \nu_j) \le \#\{\tilde{\nu} : \tilde{\nu} \le \nu\} \le \#(\Lambda_n) = n, \tag{5.18}$$

since Λ_n is downward closed. The total cost of identifying Λ_n is therefore at most of the order $n^2 \log(n)$ which is generally negligible compared to the computation of the approximation polynomial. Indeed, the latter involves n elements from the space V_h, and therefore has complexity at least nN_h which, in the practice of reduced modelling, is much larger than $n^2 \log(n)$ since $N_h \gg n$.

5.4. Reduced basis methods

A second class of numerical methods is not based on polynomial approximations. Instead, it directly seeks choices of functions $v_1, \ldots, v_n$ for which the approximation of the parametric PDE in the resulting n-dimensional space $V_n := \mathrm{span}\{v_1, \ldots, v_n\}$ performs almost as well as the optimal benchmarks for separable approximations. Recall that these benchmarks are measured by n-width $d_n(\mathcal{M})_V$ for the uniform error, or by the tail of the singular values (1.48) for the mean-square error.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We discuss in Section 7 the *reduced basis* method, which targets uniform error estimates, and which consists in generating V_n by a selection of n particular solution instances $u(a^i)$ for $i = 1, \dots, n$, chosen from a very large set of potential candidates. The selection process is critical for the success of this algorithm, and one main result is that a certain greedy strategy meets the benchmark of the n-width in the sense that it results in similar convergence rates.

Another representative of this second class of methods, which we do not discuss in this paper, is known as the *proper orthogonal decomposition* method and targets mean-square estimates. It builds the functions $\{v_1, \dots, v_n\}$ based on an empirical approximation of the exact covariance operator (1.44) using a sufficiently dense sampling of the random solution $u(a)$.

One main disadvantage of both reduced basis and proper orthogonal decomposition methods, compared to the first class of methods based on polynomial approximation, is that their offline stage is potentially very costly, especially in high parameter dimension. However, the potential gain is that they can get significantly closer to the optimal benchmarks for separable approximations. This is due to the fact that the best n-term polynomial approximation error may in some cases decay substantially more slowly than the n-width, as discussed in Section 4.4.

6. Sparse polynomial interpolation

In this section we discuss the construction of polynomial approximations to the solution map $y \mapsto u(y)$ by *interpolation*. We place ourselves in the same framework as in Section 3: we consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$, so that the solution map $y \mapsto u(y) := u(a(y))$ is then well defined from U to V.

Given $\Lambda \subset \mathcal{F}$ with $\#(\Lambda) = n$, we say that a discrete set

$$\Gamma \subset U, \quad \#(\Gamma) = n \tag{6.1}$$

is *unisolvent* for $\mathbb{P}_\Lambda$ if and only if, for any values $(v_\gamma)_{\gamma\in\Gamma} \in \mathbb{R}^\Gamma$, there exists a unique polynomial $\pi \in \mathbb{P}_\Lambda$ such that

$$\pi(\gamma) = v_\gamma, \quad \gamma \in \Gamma. \tag{6.2}$$

In such a case, to any real-valued function v defined over U we associate its interpolation polynomial $I_\Lambda v \in \mathbb{P}_\Lambda$, which satisfies

$$I_\Lambda v(\gamma) = v(\gamma), \quad \gamma \in \Gamma. \tag{6.3}$$

The interpolation operator I_Λ is a linear map from the space of real-valued functions defined over U onto $\mathbb{P}_\Lambda$. It may be written in the usual Lagrange

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

form

$$I_\Lambda v = \sum_{\gamma\in\Gamma} v(\gamma)\ell_{\Lambda,\gamma}, \tag{6.4}$$

where the $\ell_{\Lambda,\gamma} \in \mathbb{P}_\Lambda$ are uniquely defined by $\ell_{\Lambda,\gamma}(\tilde{\gamma}) = \delta_{\gamma,\tilde{\gamma}}$ for $\gamma, \tilde{\gamma} \in \Gamma$.

By a standard vectorization procedure, we may define a similar interpolation process that maps the space of V-valued functions defined on U onto the space V_Λ. This amounts to now using the V-valued $v(\gamma)$ in the definition of the interpolant by (3.64). With a slight abuse of notation, we again denote this operator by I_Λ. From exactly or approximately computed instances

$$u_\gamma = u(\gamma), \quad \gamma \in \Gamma, \tag{6.5}$$

of the solution map, we may thus compute $I_\Lambda u \in V_\Lambda$ such that

$$I_\Lambda u(\gamma) = u_\gamma, \quad \gamma \in \Gamma. \tag{6.6}$$

One of the main attractions of interpolation, also sometimes referred to as *collocation* in the context of parametric PDEs (Babuška *et al.* 2007, Nobile, Tempone and Webster 2008*a*, Nobile, Tempone and Webster 2008*b*) is that it is a non-intrusive process.

In addition to the existence and uniqueness of the interpolation polynomial, we point out two other properties of the interpolation process that are of interest to us.

(i) *Stability.* One typical way of quantifying the stability of the interpolation process is through its *Lebesgue constant*. If $\Gamma \subset U$ is a set of unisolvent interpolation points for $\mathbb{P}_\Lambda$ with Lagrange basis elements $\ell_{\Lambda,\gamma}$, the Lebesgue constant is defined as

$$\mathbb{L}_\Lambda := \sup \frac{\|I_\Lambda u\|_{L^\infty(U)}}{\|u\|_{L^\infty(U)}} = \max_{y\in U} \sum_{\gamma\in\Gamma} |\ell_{\Lambda,\gamma}(y)|, \tag{6.7}$$

where the first supremum is taken over all non-zero real-valued functions u which are everywhere defined and uniformly bounded over U. It is easily seen that we obtain the same quantity if we instead take the supremum over the set of V-valued functions, using the $L^\infty(U, V)$-norm in the quotient. The Lebesgue constant typically grows with the number n of interpolation points. However, it is well known that this growth strongly depends on the selection of points. For instance, on the univariate interval $[-1, 1]$, the Lebesgue constant for interpolation by polynomials of degree $n - 1$ at n points grows exponentially with n for uniformly spaced points and logarithmically for Chebyshev or Gauss–Lobatto points.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

(ii) *Progressivity.* We would like to use sequences $(\Lambda_n)_{n\geq 1}$ of index sets which have the nestedness property $\Lambda_n \subset \Lambda_{n+1}$ in order to define polynomial spaces with increasing accuracy. The sets Λ_n may be defined *a priori*, based on the analysis of best n-term polynomial approximations presented in Section 3, or adaptively generated. In both cases, it is desirable that the polynomial interpolation operators $I_{\Lambda_{n+1}}$ can be derived in a simple way from $I_{\Lambda_{n+1}}$. This requires in particular that the associated unisolvent sets of points $(\Gamma_n)_{n\geq 0}$ are themselves nested.

It was shown in Chkifa *et al.* (2014) that such progressive interpolation processes can be derived provided that the sets Λ_n are downward closed. We present this approach in Section 6.1 and discuss its stability properties in Section 6.2. We finally discuss in Section 6.3 the computational cost of such interpolation schemes, taking into account the space discretization for the computation of the instances $u(\gamma)$, for example using a finite element method.

6.1. Sparse interpolation using downward closed sets

We describe the construction of the interpolation operator for real-valued functions, since, as previously explained, it induces a similar interpolation operator for V-valued functions.

We begin by discussing progressive constructions in the case of univariate polynomial interpolation. The starting point is any sequence

$$T = (t_k)_{k\geq 0}, \tag{6.8}$$

of distinct points from $[-1,1]$. We introduce the abbreviated notation

$$I_k := I_{\{t_0,\dots,t_k\}}, \tag{6.9}$$

for the univariate interpolation operator for the k-section $\{t_0,\dots,t_k\}$ of this sequence: for any function u defined everywhere over $[-1,1]$, the polynomial $I_k u \in \mathbb{P}_k$ satisfies

$$I_k u(t_i) = u(t_i), \quad i = 0,\dots,k. \tag{6.10}$$

We can express I_k in a hierarchical form,

$$I_k u = I_0 u + \sum_{l=1}^{k} \Delta_l u, \quad \Delta_l := I_l - I_{l-1}, \tag{6.11}$$

also commonly known as the Newton form. We set $I_{-1} = 0$ so that we can also write

$$I_k u = \sum_{l=0}^{k} \Delta_l u. \tag{6.12}$$

Since $I_k u$ and $I_{k-1} u$ agree at the points $\{t_0,\dots,t_{k-1}\}$, it is readily seen

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

that, for $k > 0$,

$$\Delta_k u(t) = \alpha_k h_k(t), \tag{6.13}$$

where

$$\alpha_k = \alpha_k(u) := u(t_k) - I_{k-1}u(t_k) \tag{6.14}$$

is the error at t_k of interpolation by I_{k-1}, and

$$h_k(t) := \prod_{l=0}^{k-1} \frac{t - t_l}{t_k - t_l}. \tag{6.15}$$

We also set

$$h_0(t) := 1. \tag{6.16}$$

For all $k \geq 0$, the system $\{h_0, \ldots, h_k\}$ is a basis for $\mathbb{P}_k$, sometimes called a *hierarchical basis*.

Although the sequence T could be arbitrary, the stability of the resulting interpolation scheme, as reflected through the growth of its Lebesgue constants, depends very much on the choice of T. One interesting choice is the sequence of the so-called *Leja points*, which is initiated from an arbitrary t_0 (usually taken to be 1 or 0) and recursively defined by

$$t_k := \arg\max \left\{ \prod_{l=0}^{k-1} |t - t_l| : t \in [-1, 1] \right\}. \tag{6.17}$$

With this particular choice, we note that the hierarchical basis functions satisfy

$$\|h_k\|_{L^\infty([-1,1])} = 1, \quad k \geq 0. \tag{6.18}$$

The Leja points may be viewed as an incremental variant to the classical Fekete points

$$\{t_{0,k}, \ldots, t_{k,k}\} := \arg\max \left\{ \prod_{i \neq j} |t_i - t_j| : \{t_0, \ldots, t_k\} \in [-1, 1]^{k+1} \right\}, \tag{6.19}$$

which, in contrast to the Leja points, are *not* k-sections of a single sequence.

We now turn to the multivariate setting. Starting again with the univariate sequence T, we now define the points

$$y_\nu := (t_{\nu_j})_{j \geq 1} \in U, \quad \nu \in \mathcal{F}, \tag{6.20}$$

which are therefore extracted from the tensorized grid $T^{\mathbb{N}}$. We also define the tensorized operators

$$I_\nu := \otimes_{j \geq 1} I_{\nu_j}. \tag{6.21}$$

Recall that the application of a tensorized operator $\otimes_{j \geq 1} A_j$ to a multivariate function reduces to applying each univariate operator A_j, by freezing all

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

variables except the jth, and then applying A_j to the non-frozen variable. We may define I_ν by induction. For this, let us introduce $\mathcal{F}_k$, the set of all ν such that $\nu_j = 0$ for $j \geq k$.

- For $k = 1$, there is only $\nu = 0$, the null multi-index contained in $\mathcal{F}_1$. Then $I_0 u$ is the constant function with value $u(y_0)$, where $y_0 = (t_0, t_0, \ldots)$.
- For $k > 1$, assuming that $I_{\tilde{\nu}}$ has been defined for any $\tilde{\nu} \in \mathcal{F}_{k-1}$, and taking $\nu \in \mathcal{F}_k$, we write

$$\nu = (\nu_1, \tilde{\nu}), \quad \tilde{\nu} = (\nu_2, \nu_3, \ldots) \in \mathcal{F}_{k-1}, \tag{6.22}$$

and, for any $y \in U$,

$$y = (y_1, \tilde{y}), \quad \tilde{y} = (y_2, y_3, \ldots). \tag{6.23}$$

We then define $I_\nu := I_{\nu_1} \otimes I_{\tilde{\nu}}$, that is,

$$I_\nu u(y) = I_{\tilde{\nu}} v_{y_1}(\tilde{y}), \tag{6.24}$$

where $v_{y_1}(\tilde{y}) := I_{\nu_1} u_{\tilde{y}}(y_1)$ with $u_{\tilde{y}}$ the univariate function defined on $[-1, 1]$ by $u_{\tilde{y}}(t) = u(y)$.

Note that in the finite-dimensional case $U = [-1, 1]^d$, this induction terminates after at most d steps.

It is easily seen that I_ν is the interpolation operator on the tensor product polynomial space

$$\mathbb{P}_\nu = \otimes_{j \geq 1} \mathbb{P}_{\nu_j}, \tag{6.25}$$

for the grid of points

$$\Gamma_\nu = \otimes_{j \geq 1} \{t_0, \ldots, t_{\nu_j}\}, \tag{6.26}$$

which is unisolvent for this space. This polynomial space corresponds to a particular set Λ which has rectangular shape. That is, $\Lambda = R_\nu$, where, for any $\nu \in \mathcal{F}$, we define the *shadow* of ν as

$$R_\nu := \{\tilde{\nu} : \tilde{\nu} \leq \nu\}. \tag{6.27}$$

We thus have $\mathbb{P}_\nu = \mathbb{P}_{R_\nu}$.

We next define in a similar manner the tensorized difference operators

$$\Delta_\nu := \otimes_{j \geq 1} \Delta_{\nu_j}. \tag{6.28}$$

It follows that the range of Δ_ν is the one-dimensional subspace of $\mathbb{P}_\nu$ spanned by

$$H_\nu(y) := \prod_{\nu_j \neq 0} h_{\nu_j}(y_j), \quad \nu \in \mathcal{F}. \tag{6.29}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

To a general finite set $\Lambda \subset \mathcal{F}$, we associate the operator

$$I_\Lambda := \sum_{\nu \in \Lambda} \Delta_\nu, \tag{6.30}$$

and the grid

$$\Gamma_\Lambda := \{y_\nu : \nu \in \Lambda\}. \tag{6.31}$$

In the case where $\Lambda = R_\nu$, we find that $\Gamma_\Lambda = \Gamma_\nu$. It is thus unisolvent for $\mathbb{P}_\Lambda$. In addition, we then have

$$I_\nu = \otimes_{j \geq 1} \left(\sum_{l=0}^{\nu_j} \Delta_l \right) = \sum_{\tilde{\nu} \leq \nu} \Delta_{\tilde{\nu}} = I_\Lambda, \tag{6.32}$$

which shows that I_Λ is the interpolation operator onto $\mathbb{P}_\Lambda$ for this grid.

Let us remark that for a general set Λ, the set Γ_Λ is not unisolvent on $\mathbb{P}_\Lambda$ and I_Λ is not an interpolation operator. However, an important observation is that this is the case whenever Λ is an arbitrary downward closed set. This fact was first noticed in Kuntzman (1959) for bivariate functions, and then used in higher dimensions for particular cases of downward closed sets in Smolyak (1963).

Theorem 6.1. Let $\Lambda \subset \mathcal{F}$ be a finite downward closed set. Then the grid Γ_Λ is unisolvent for $\mathbb{P}_\Lambda$ and I_Λ is the interpolation operator onto $\mathbb{P}_\Lambda$ for this grid.

Proof. Because of the downward closed set property, $P_\nu \subset P_\Lambda$ for all $\nu \in \Lambda$. Hence the image of I_Λ is contained in $\mathbb{P}_\Lambda$. In order to prove that it is the interpolation operator for the grid Γ_Λ, we need to show that, for any function u defined over U,

$$I_\Lambda u(y_\nu) = u(y_\nu), \quad \nu \in \Lambda. \tag{6.33}$$

Since $\#(\Gamma_\Lambda) = \dim(\mathbb{P}_\Lambda)$, this also ensures the unisolvence of Γ_Λ for P_Λ.

For any $\nu \in \Lambda$, we may write

$$I_\Lambda u = I_\nu u + \sum_{\tilde{\nu} \in \Lambda, \tilde{\nu} \not\leq \nu} \Delta_{\tilde{\nu}} u. \tag{6.34}$$

Since I_ν is the interpolant on the tensor product grid Γ_ν, and this grid contains y_ν, it follows that

$$I_\nu u(y_\nu) = u(y_\nu). \tag{6.35}$$

On the other hand, if $\tilde{\nu} \in \Lambda$ is such that $\tilde{\nu} \not\leq \nu$, this means that there exists a $j \geq 0$ such that $\tilde{\nu}_j > \nu_j$. For this j we thus have $\Delta_{\tilde{\nu}} u(y) = 0$ for all $y \in U$ with jth coordinate equal to t_{ν_j} due to the application of Δ_{ν_j} in the jth variable. Therefore

$$\Delta_{\tilde{\nu}} u(y_\nu) = 0. \tag{6.36}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

It follows that $I_\Lambda u(y_\nu) = u(y_\nu)$, which concludes the proof. □

The decomposition (6.30) of I_Λ as a sum of the various Δ_ν may be viewed as a generalization of the Newton form (6.11). This decomposition also yields a simple strategy for the fast computation of $I_\Lambda u$, which we now describe.

We first observe that if Λ is a downward closed set of cardinality $n > 0$, we can find at least one $\nu \in \Lambda$ which is maximal in Λ, that is, such that

$$\tilde{\nu} \geq \nu \quad \text{and} \quad \tilde{\nu} \neq \nu \quad \Rightarrow \quad \tilde{\nu} \notin \Lambda. \tag{6.37}$$

We may then write

$$\Lambda = \tilde{\Lambda} \cup \{\nu\}, \tag{6.38}$$

where $\tilde{\Lambda}$ is a downward closed set of cardinality $n-1$. Writing

$$I_\Lambda u = I_{\tilde{\Lambda}} u + \Delta_\nu u, \tag{6.39}$$

we observe that Δ_ν is characterized by the fact that it belongs to $\mathbb{P}_\nu$ and is characterized by

$$\Delta_\nu u(y_{\tilde{\nu}}) = 0, \quad \tilde{\nu} \in R_\nu - \{\nu\}, \tag{6.40}$$

and

$$\Delta_\nu u(y_\nu) = I_\Lambda u(y_\nu) - I_{\tilde{\Lambda}} u(y_\nu) = u(y_\nu) - I_{\tilde{\Lambda}} u(y_\nu). \tag{6.41}$$

Using the tensorized hierarchical basis function H_ν, it follows that

$$\Delta_\nu u = \alpha_\nu H_\nu, \quad \alpha_\nu = \alpha_\nu(u) := u(y_\nu) - I_{\tilde{\Lambda}} u(y_\nu). \tag{6.42}$$

By iteration, we may write $\Lambda = \{\nu^1, \dots, \nu^n\}$, where the ν^i are enumerated in such a way that for each i, $\Lambda_i = \{\nu^1, \dots, \nu^i\}$ is a downward closed set. This allows us to compute I_Λ by n recursive applications of

$$I_{\Lambda_i} u = I_{\Lambda_{i-1}} u + \alpha_{\nu^i} H_{\nu^i}. \tag{6.43}$$

Note that $(H_\nu)_{\nu \in \Lambda}$ is a basis of $\mathbb{P}_\Lambda$ and that any $v \in \mathbb{P}_\Lambda$ has the unique decomposition

$$v = \sum_{\nu \in \Lambda} \alpha_\nu H_\nu, \tag{6.44}$$

where the coefficients $\alpha_\nu = \alpha_\nu(v)$ are defined by the above procedure applied to v. Therefore, although the enumeration $\{\nu^1, \dots, \nu^n\}$ is not unique, the coefficients $\alpha_\nu = \alpha_\nu(u)$ in the expression

$$I_\Lambda u = \sum_{\nu \in \Lambda} \alpha_\nu H_\nu \tag{6.45}$$

are unique. Also note that $\alpha_\nu(u)$ does not depend on the choice of Λ but only on ν and u. This computation is exactly the same in the case of V-valued functions, now with uniquely defined coefficients $\alpha_\nu \in V$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The recursive computation of the interpolation operator by (6.43) can be used in two different contexts.

- *Non-adaptive methods.* A nested sequence $(\Lambda_n)_{n\geq 0}$ of downward closed sets is prescribed in advance, and we use (6.43) to compute $I_{\Lambda_n}u$ for increasing values of n.
- *Adaptive methods.* The sequence $(\Lambda_n)_{n\geq 0}$ is not prescribed in advance, and we use the computation of $I_{\Lambda_n}u$ to define Λ_{n+1}.

We next give a typical example of an adaptive interpolation algorithm. In order to present this algorithm we begin with an analogy. Since we have

$$I_\Lambda u = \sum_{\nu\in\Lambda} \alpha_\nu H_\nu, \tag{6.46}$$

we may view the interpolant as a truncation of the formal infinite expansion of u in the hierarchical basis

$$\sum_{\nu\in\mathcal{F}} \alpha_\nu H_\nu. \tag{6.47}$$

From elementary results on polynomial interpolation, we know that this series does not converge for a general function defined everywhere over U. Even for the various models of parametric PDEs discussed in this paper, we do not know natural conditions that would ensure the unconditional convergence of this expansion to u, in contrast to the Taylor and Legendre series discussed in Section 3.

In particular, we do not know estimates for the coefficients α_ν which would allow us to establish convergence rates for the best n-term truncations.

Nevertheless, we may still take the same view as in Section 3, and use for Λ_n the set of indices corresponding to the n largest terms of (6.46) measured in some given metric $L^p(U,V,\mu)$. We take $p=\infty$ if we search for uniform approximation estimates or $p=2$ if we search for mean-square approximation estimates. This amounts to choosing the indices of the n largest $c_\nu\|\alpha_\nu\|_V$, where c_ν is given by

$$c_\nu := \|H_\nu\|_{L^p(U,\mu)}. \tag{6.48}$$

In the case where μ is the uniform measure, we also have

$$c_\nu := \prod_{j\geq 1} \|h_{\nu_j}\|_{L^p([-1,1],\frac{\mathrm{d}t}{2})}. \tag{6.49}$$

Note that if $p=\infty$ and we use the Leja sequence, then $\|H_\nu\|_{L^\infty(U)}=1$, and therefore this amounts to choosing the largest $\|\alpha_\nu\|_V$. The defect of this strategy is that the sets Λ_n are not ensured to be downward closed. In addition, we generally cannot afford an exhaustive search for the n largest contributions in (6.46).

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

In order to build a feasible adaptive algorithm, we need to limit this search. In what follows, we describe a greedy algorithm proposed in Chkifa *et al.* (2014) for the selection of the sequence $(\Lambda_n)_{n\geq 1}$, which uses the set of neighbours $N(\Lambda)$ defined by (5.13). We first give an idealized version of this algorithm, which cannot be applied as such.

Greedy interpolation algorithm. We start with the null multi-index $\Lambda_1 := \{0\}$. Assuming that Λ_{n-1} has been selected and that the $(\alpha_\nu)_{\nu\in\Lambda_{n-1}}$ have been computed, we compute the α_ν for $\nu \in N(\Lambda_{n-1})$. We then set

$$\nu^n := \arg\max\{c_\nu \|\alpha_\nu\|_V : \nu \in N(\Lambda_{n-1})\}, \tag{6.50}$$

and define $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\}$.

Note that when $p = \infty$ and T is the Leja sequence, this strategy amounts to choosing the $\nu \in N(\Lambda_{n-1})$ that maximizes the interpolation error at the new grid point, which would be added by adjoining ν, that is, setting

$$\nu^n := \arg\max\{\|u(y_\nu) - I_{\Lambda_{n-1}} u(y_\nu)\|_V : \nu \in N(\Lambda_{n-1})\}. \tag{6.51}$$

The above greedy algorithm is not computationally feasible, since we are in principle working with infinitely many variables $(y_j)_{j\geq 1}$, in which case the set of neighbours $N(\Lambda)$ to be explored has infinite cardinality. One way to circumvent this defect is to replace the infinite set $N(\Lambda_n)$ by the finite set of anchored neighbours $\tilde{N}(\Lambda_n)$ defined by (5.14).

One more serious defect of this algorithm is that it may fail to converge, even if there exist sequences $(\Lambda_n)_{n\geq 0}$ such that $I_{\Lambda_n} u$ rapidly converges to u. Indeed, if $\Delta_\nu u = 0$ for a certain ν, then no index $\tilde{\nu} \geq \nu$ will ever be selected by the algorithm. As an example, consider a two-dimensional function of the form

$$u(y) = u_1(y_1)u_2(y_2), \tag{6.52}$$

where u_1 and u_2 are non-polynomial smooth functions such that $u_2(t_0) = u_2(t_1)$. Then the sets Λ_n selected by the algorithms will consist of the indices $\nu = (k, 0)$ for $k = 0, \ldots, n-1$, since the interpolation error at the point (t_k, t_1) always vanishes. One way to avoid this problem is to change the strategy by alternating the selection of ν^n using (6.51) with a selection rule ensuring that all indices are eventually picked. For example, when n is even, we define ν^n according to (6.51), and when n is odd, for ν^n we pick the multi-index $\nu \in \tilde{N}(\Lambda_n)$, which appears at the earliest stage in the neighbours of the previous sets Λ_k. In summary, this results in the following algorithm.

Alternating greedy interpolation algorithm. We start with the null multi-index $\Lambda_1 := \{0\}$. Assuming that Λ_{n-1} has been selected and that the $(\alpha_\nu)_{\nu\in\Lambda_{n-1}}$ have been computed, we compute the α_ν for $\nu \in \tilde{N}(\Lambda_{n-1})$. If n

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

is even, we set

$$\nu^n := \arg\max\{c_\nu \|\alpha_\nu\|_V : \nu \in \tilde{N}(\Lambda_{n-1})\}, \tag{6.53}$$

and if n is odd,

$$\nu^n := \arg\min\{k(\nu) : \nu \in \tilde{N}(\Lambda_{n-1}\}, \quad k(\nu) := \min\{k : \nu \in \tilde{N}(\Lambda_k)\}. \tag{6.54}$$

We then define $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\}$.

Even with such modifications, although the adaptive algorithm seems to behave well in many practical instances, the convergence of the interpolation produced by this algorithm is still not guaranteed. It is an open problem to discover which additional assumptions on u ensure convergence and, more importantly, a convergence rate comparable to that which is proved for best n-term approximations based on Taylor and Legendre series. Note that the solution to this problem must involve the initial choice of the univariate sequence T, which, as discussed in the next subsection, greatly affects the stability and convergence properties of the interpolation process. In the next subsection we establish convergence rates for the interpolation algorithm using this stability analysis, but based on non-adaptive choices of the sequence $(\Lambda_n)_{n\geq 1}$.

Remark 6.2. A very similar greedy algorithm was proposed by Gerstner and Griebel (2003) in the slightly different context of adaptive quadrature, that is, when we want to approximate the integral of u over the domain U rather than u itself. In that case, one natural choice is to pick the new neighbour ν that maximizes the absolute value of the integral of $\Delta_\nu u$.

Let us conclude this subsection by mentioning that there exist several natural generalizations to the above-described construction of the sparse multivariate interpolation process.

The first obvious one is that we could work on more general tensor product domains of the form

$$U = \otimes_{j\geq 0} U_j, \tag{6.55}$$

where the U_j are univariate intervals or other bounded domains in $\mathbb{R}$ or $\mathbb{C}$, and define points y_ν by tensorization of sequences,

$$T_j = (t_{j,k})_{k\geq 0}, \tag{6.56}$$

of pairwise distinct points, each of them picked from U_j.

The second generalization is that we could start with univariate systems other than polynomials which still possess a hierarchical interpolation structure. We consider a general index set $\mathcal{S}$ equipped with a partial order $\leq$ and assume that there exists a root index 0 such that $0 \leq \gamma$ for all $\gamma \in \mathcal{S}$. Given a grid of pairwise distinct points $G = (t_\gamma)_{\gamma\in\mathcal{S}}$, we say that a family of functions $(h_\gamma)_{\gamma\in\Gamma}$ defined over $[-1, 1]$ is a hierarchical basis associated to

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

the grid G if and only if $h_0(t) = 1$ and

$$h_\gamma(t_\gamma) = 1 \quad \text{and} \quad h_\gamma(t_{\tilde\gamma}) = 0 \quad \text{if } \tilde\gamma \le \gamma \text{ and } \tilde\gamma \ne \gamma. \tag{6.57}$$

By tensorization, we obtain an index set $\mathcal{F} \subset G^{\mathbb{N}}$ of finitely supported sequences, equipped with a partial order $\le$ induced by its univariate counterpart. This allows us to define downward closed sets in $\mathcal{F}$ in the same way as for the particular case $G = \mathbb{N}$. For $\nu = (\nu_j)_{j\ge1} \in \mathcal{F}$, we also define the points $y_\nu \in U$ and the tensorized hierarchical functions H_ν in the same way as in (6.20) and (6.29). Then, if Λ is a downward closed set, we may inductively define an interpolation operator I_Λ onto the space

$$H_\Lambda := \operatorname{span}\{H_\nu : \nu \in \Lambda\} \tag{6.58}$$

associated to the grid Γ_Λ, using the same recursion,

$$I_\Lambda u = I_{\tilde\Lambda} u + \alpha_\nu H_\nu, \quad \alpha_\nu := \alpha_\nu(u) = u(y_\nu) - I_{\tilde\Lambda} u(y_\nu), \tag{6.59}$$

where $\nu \notin \tilde\Lambda$ and $\tilde\Lambda$ is a downward closed set such that $\Lambda = \tilde\Lambda \cup \{\nu\}$. We initialize this computation for $\Lambda = \{0\}$, where 0 is the null multi-index, by defining $I_{\{0\}}u$ as the constant function with value $u(y_0)$. Examples of relevant hierarchical systems include the classical piecewise linear, or more generally piecewise polynomial, hierarchical basis functions. With such choices the spaces H_Λ include as particular cases the well-studied piecewise polynomial *sparse grid* spaces; see Bungartz and Griebel (2004) for a survey on this topic.

6.2. Stability

We now turn to the stability analysis of the interpolation operator. We recall the Lebesgue constant defined in (6.7). One principal interest of the Lebesgue constant is that it allows us to estimate the error of interpolation in terms of the best polynomial approximation error in the L^∞-norm. Indeed, for any $u \in L^\infty(U,V)$ and any $v \in V_\Lambda$, we may write

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \le \|u - v\|_{L^\infty(U,V)} + \|I_\Lambda v - I_\Lambda u\|_{L^\infty(U,V)}, \tag{6.60}$$

which yields

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \le (1 + \mathbb{L}_\Lambda) \inf_{v\in V_\Lambda} \|u - v\|_{L^\infty(U,V)}, \tag{6.61}$$

by taking the infimum over V_Λ.

We know from the results in Section 3.8, in particular Corollary 3.26, that for relevant classes of parametric PDEs we can find nested sequences of downward closed sets $(\Lambda_n)_{n\ge1}$ with $\#(\Lambda_n) = n$, such that

$$\inf_{v\in V_{\Lambda_n}} \|u - v\|_{L^\infty(U,V)} \le Cn^{-s}, \tag{6.62}$$

where $s > 0$ is some given rate. This holds in particular with $s := \frac{1}{p} - 1$ if the

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

assumptions of Theorem 2.9 hold and if, in addition, $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$. Therefore we have the error bound

$$\|u - I_{\Lambda_n} u\|_{L^\infty(U,V)} \leq C(1 + \mathbb{L}_{\Lambda_n}) n^{-s} \tag{6.63}$$

if we use such sequences in the interpolation process. This motivates estimation of the growth of $\mathbb{L}_{\Lambda_n}$ with n.

In order to estimate $\mathbb{L}_\Lambda$, we introduce the univariate Lebesgue constants

$$\lambda_k := \sup \frac{\|I_k u\|_{L^\infty([-1,1])}}{\|u\|_{L^\infty([-1,1])}}, \tag{6.64}$$

where the supremum is taken over all non-zero real-valued functions u that are everywhere defined and uniformly bounded on $[-1,1]$. We define an analogous quantity for the difference operator Δ_k, namely

$$\delta_k := \sup \frac{\|\Delta_k u\|_{L^\infty([-1,1])}}{\|u\|_{L^\infty([-1,1])}}, \tag{6.65}$$

and observe that

$$\delta_k \leq \lambda_{k-1} + \lambda_k, \quad k \geq 0, \tag{6.66}$$

where we have set $\lambda_{-1} = 0$. We introduce for each $\nu \in \mathcal{F}$ the quantities

$$\delta_\nu := \sup \frac{\|\Delta_\nu u\|_{L^\infty(U)}}{\|u\|_{L^\infty(U)}}, \tag{6.67}$$

so that on the one hand we have

$$\delta_\nu \leq \prod_{j\geq 1} \delta_{\nu_j} \leq \prod_{j\geq 1} (\lambda_{\nu_j - 1} + \lambda_{\nu_j}), \tag{6.68}$$

and on the other hand

$$\mathbb{L}_\Lambda \leq \sum_{\nu \in \Lambda} \delta_\nu. \tag{6.69}$$

The following result from Chkifa *et al.* (2014) gives an estimate on the growth of $\mathbb{L}_\Lambda$ in terms of $\#(\Lambda)$, provided that a similar estimate holds for the univariate Lebesgue constant λ_k or for the quantity δ_k.

Theorem 6.3. If either one of the estimates

$$\lambda_k \leq (k+1)^\theta, \quad k \geq 0, \tag{6.70}$$

or

$$\delta_k \leq (k+1)^\theta, \quad k \geq 0, \tag{6.71}$$

holds for some $\theta \geq 1$, then the Lebesgue constant $\mathbb{L}_\Lambda$ satisfies

$$\mathbb{L}_\Lambda \leq (\#(\Lambda))^{\theta+1} \tag{6.72}$$

for any downward closed set Λ.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Proof. The case where (6.71) holds is elementary since the first inequality in (6.68) yields

$$\delta_\nu \le \prod_{j\ge 1}(\nu_j+1)^\theta = \left(\prod_{j\ge 1}(\nu_j+1)\right)^\theta = (\#(R_\nu))^\theta \le (\#(\Lambda))^\theta, \tag{6.73}$$

where we have used the fact that $R_\nu \subset \Lambda$ since Λ is downward closed. Using (6.69) we thus obtain (6.72).

For the case where (6.70) holds, we observe that

$$\lambda_k + \lambda_{k-1} \le (k+1)^\theta + k^\theta \le (2k+1)(k+1)^{\theta-1}. \tag{6.74}$$

The second inequality in (6.68) yields

$$\begin{aligned}\delta_\nu &\le \left(\prod_{j\ge 1}(\nu_j+1)\right)^{\theta-1}\prod_{j\ge 1}(2\nu_j+1)\\ &= (\#(R_\nu))^{\theta-1}\prod_{j\ge 1}(2\nu_j+1)\\ &\le (\#(\Lambda))^{\theta-1}\prod_{j\ge 1}(2\nu_j+1).\end{aligned}$$

In order to establish (6.72), it thus suffices to prove that $\sigma(\Lambda) \le (\#\Lambda)^2$, where

$$\sigma(\Lambda) := \sum_{\nu\in\Lambda}\prod_{j\ge 1}(2\nu_j+1). \tag{6.75}$$

For this, we use induction on $n := \#(\Lambda)$. For $n = 1$ and $\Lambda = \{0\}$ the result obviously holds. Assuming that it holds for some $n \ge 1$, we consider a downward closed set Λ of cardinality $n+1$. We may assume without loss of generality that $\nu_1 \ne 0$ for some $\nu \in \Lambda$, and denote by $K \ge 1$ the maximal value attained by the coordinate ν_1 when $\nu \in \Lambda$. For $0 \le k \le K$, we define

$$\Lambda_k := \{\hat\nu = (\nu_2, \nu_3, \dots) : (k, \hat\nu) \in \Lambda\}. \tag{6.76}$$

Each of the sets Λ_k is downward closed and, since $K \ge 1$, we have $\#(\Lambda_k) < \#(\Lambda)$ for all $k = 0, \dots, K$. The induction hypothesis implies

$$\sigma(\Lambda) = \sum_{k=0}^{K}\sum_{\nu\in\Lambda_k}\prod_{j\ge 1}(2\nu_j+1) = \sum_{k=0}^{K}(2k+1)\sigma(\Lambda_k) \le \sum_{k=0}^{K}(2k+1)(\#(\Lambda_k))^2. \tag{6.77}$$

Also, we have

$$\Lambda_K \subset \cdots \subset \Lambda_1 \subset \Lambda_0, \tag{6.78}$$

since for $k \ge 1$,

$$\nu \in \Lambda_k \Rightarrow (k,\nu) \in \Lambda \Rightarrow (k-1,\nu) \in \Lambda \Rightarrow \nu \in \Lambda_{k-1}.$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

This implies

$$k(\#(\Lambda_k))^2 \leq \#(\Lambda_k) \sum_{j=0}^{k-1} \#(\Lambda_j), \tag{6.79}$$

and therefore

$$\sigma(\Lambda) \leq \sum_{k=0}^{K} (\#(\Lambda_k))^2 + 2 \sum_{k=0}^{K} \#(\Lambda_k) \sum_{j=0}^{k-1} \#(\Lambda_j) = \left(\sum_{k=0}^{K} \#(\Lambda_k) \right)^2 = (\#(\Lambda))^2, \tag{6.80}$$

which concludes the proof. □

Remark 6.4. One noticeable feature of the above result is that the bound on $\mathbb{L}_\Lambda$ only depends on the cardinality of Λ. In particular, it is independent of the number of variables, which can be infinite, and independent of the shape of Λ.

In view of the above result, we are therefore interested in choosing univariate sequences $T = (t_k)_{k\geq 0}$ such that the Lebesgue constant λ_k or the quantity δ_k have moderate algebraic growth with k. It is well known that for particular sets of points such as the Chebyshev points

$$C_k := \left\{ \cos\left(\frac{2l+1}{2k+2} \pi \right) : l = 0, \ldots, k \right\}, \tag{6.81}$$

or the Gauss–Lobatto (or Clenshaw–Curtis) points

$$G_k := \left\{ \cos\left(\frac{l}{k} \pi \right) : l = 0, \ldots, k \right\}, \tag{6.82}$$

the Lebesgue constant has logarithmic growth $\lambda_k \sim \log(k)$, and is therefore slower than algebraic. However, these points are not adapted to our construction since the sets C_k and G_k are not nested as k grows, and are therefore not the k-sections of a single sequence.

For the Leja points defined by (6.17), numerical computation of λ_k for the first 200 values of k indicates that the linear bound

$$\lambda_k \leq (1+k) \tag{6.83}$$

seems to hold and that this bound could be sharp. However, there is currently no rigorous proof supporting this evidence or establishing another algebraic rate. Nevertheless, Leja points seem to be a good choice for the construction of our multivariate interpolation process.

Leja points have also been considered on the complex unit disc $\{|z| \leq 1\}$, taking for example $t_0 = 1$ and again using the recursion (6.17), now with $|\cdot|$ standing for the modulus. These points have the property of accumulating

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

in a regular manner on the unit circle according to the so-called Van der Corput enumeration (Calvi and Phung 2011). Their projections on the real axis are called the $\mathfrak{R}$-Leja points, and coincide with the Gauss–Lobatto points for values of k of the form $2^n + 1$ for $n \geq 0$. The growth of the Lebesgue constant λ_k has been studied by Calvi and Phung (2011, 2012), Chkifa (2013) and Chkifa (2015) for these two families of points. In the case of complex Leja points, this constant is defined as in (6.64), but taking the supremum over functions defined everywhere and bounded over the complex unit disc. It is proved in Chkifa (2013) that the linear bound (6.83) holds for the complex Leja points. For the $\mathfrak{R}$-Leja points, quadratic bounds of the type

$$\lambda_k \leq C(1+k)^2 \quad \text{and} \quad \delta_k \leq (1+k)^2, \tag{6.84}$$

with $C > 1$, are established in Chkifa (2015).

With such estimates, application of Theorem 6.3 gives us bounds of the form

$$\mathbb{L}_\Lambda \leq (\#(\Lambda))^{1+\theta}, \tag{6.85}$$

for example with $\theta = 2$ when using the $\mathfrak{R}$-Leja points. If we combine this bound with (6.63), we obtain the convergence estimate

$$\|u - I_{\Lambda_n} u\|_{L^\infty(U,V)} \leq C n^{-(s-1-\theta)}, \tag{6.86}$$

which expresses a potential deterioration of the convergence rate when using the interpolation process instead of the truncated expansions studied in Section 3.

We now present a sharper analysis, introduced in Chkifa *et al.* (2014), which reveals that this deterioration actually does not occur for the models of parametric PDEs which are of interest to us. This analysis is based on the following lemma, which gives an estimate of the interpolation error in terms of the tail of the Legendre coefficients of u multiplied by algebraic factors.

Lemma 6.5. Assume that the Legendre expansion (3.26) of u is unconditionally convergent in $L^\infty(U,V)$. If the univariate sequence $T = (t_k)_{k\geq 0}$ is such that that (6.70) or (6.71) holds for some $\theta \geq 1$, then, for any downward closed set Λ,

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \leq 2 \sum_{\nu \notin \Lambda} p_\nu(b) \|w_\nu\|_V, \tag{6.87}$$

where $b := \theta + 1$ and

$$p_\nu(b) := \prod_{j\geq 1} (1+\nu_j)^b. \tag{6.88}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Proof. The unconditional convergence of the Legendre series allows us to write

$$I_\Lambda u = I_\Lambda\left(\sum_{\nu\in\mathcal{F}} w_\nu P_\nu\right) = \sum_{\nu\in\mathcal{F}} w_\nu I_\Lambda P_\nu = \sum_{\nu\in\Lambda} w_\nu P_\nu + \sum_{\nu\notin\Lambda} w_\nu I_\Lambda P_\nu, \quad (6.89)$$

where we have used that $P_\nu \in \mathbb{P}_\Lambda$ because Λ is downward closed and hence $I_\Lambda P_\nu = P_\nu$ for every $\nu \in \Lambda$. For the second term, we observe that for each $\nu \notin \Lambda$,

$$I_\Lambda P_\nu = \sum_{\tilde\nu\in\Lambda} \Delta_{\tilde\nu} P_\nu = \sum_{\tilde\nu\in\Lambda\cap R_\nu} \Delta_{\tilde\nu} P_\nu = I_{\Lambda\cap R_\nu} P_\nu, \quad (6.90)$$

since $\Delta_{\tilde\nu} v = 0$ whenever $\tilde\nu \not\leq \nu$ and $v \in \mathbb{P}_\nu$. Therefore

$$u - I_\Lambda u = \sum_{\nu\notin\Lambda} w_\nu (I - I_{\Lambda\cap\mathcal{R}_\nu}) P_\nu, \quad (6.91)$$

where I stands for the identity operator. This implies

$$\|u - I_\Lambda u\|_{L^\infty(U,V)} \leq \sum_{\nu\notin\Lambda} (1 + \mathbb{L}_{\Lambda\cap R_\nu}) \|w_\nu\|_V \leq 2 \sum_{\nu\notin\Lambda} \mathbb{L}_{\Lambda\cap R_\nu} \|w_\nu\|_V. \quad (6.92)$$

Since (6.70) or (6.71) holds, we obtain from Theorem 6.3 that

$$\mathbb{L}_{\Lambda\cap R_\nu} \leq \#(\Lambda \cap R_\nu)^{\theta+1} \leq \#(R_\nu)^{\theta+1} = p_\nu(b), \quad (6.93)$$

which completes the proof. □

The estimate for the interpolation error in Lemma 6.5 is very similar to that of the truncated Legendre expansion, up to the presence of the factor $p_\nu(b)$. We may therefore apply to this expansion the techniques used in Section 3 in order to establish convergence rates for the interpolation error. We first establish a summability result analogous to Theorem 3.25.

Theorem 6.6. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. If the assumptions of Theorem 2.9 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then $(p_\nu(b)\|w_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p_m(\mathcal{F})$ for the same value of p, and for any $b \geq 0$.

Proof. This proof is very similar to the proof of Theorem 3.25 and so we only sketch it. We obtain a similar estimate

$$p_\nu(b)\|w_\nu\|_V \leq \tilde r_\nu := \tilde r_E(\nu)\tilde r_F(\nu), \quad (6.94)$$

where $\tilde r_E$ has exactly the same form as in (3.167) up to a change in the multiplicative constant C_0, and $\tilde r_F$ is now given by

$$\tilde r_F(\nu) := \prod_{j\in F\cap\mathrm{supp}(\nu)} c_\kappa (1+2\nu_j)^{1+b} \left(\beta + \frac{\varepsilon\nu_j}{3b_j|\nu_F|}\right)^{-\nu_j}. \quad (6.95)$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

By similar reasoning, up to a modification in the choice of J and β, one then shows that the sequence $(\tilde{r}_\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ and that it is monotone non-increasing. □

Combining this result with Lemma 6.5, we obtain the following corollary, which shows that the interpolation process converges without deterioration of the rate established for the Legendre series.

Corollary 6.7. Consider a parametric problem of the form (1.1) such that Assumption A holds for a suitable affine representation (1.15). Assume that the univariate sequence $T = (t_k)_{k\geq 0}$ is such that that (6.70) or (6.71) holds for some $\theta \geq 1$. Then, if the assumptions of Theorem 2.9 are satisfied, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, there exists a sequence of nested downward closed sets $(\Lambda_n)_{n\geq 1}$ such that $\#(\Lambda_n) = n$ and such that

$$\|u - I_{\Lambda_n} u\|_{L^\infty(U,V)} \leq Cn^{-s}, \quad n \geq 1, \quad s := \frac{1}{p} - 1. \tag{6.96}$$

6.3. Space discretization and computational cost

In practice, when applying numerical computation to parametric PDEs, as explained in Section 5, we replace the instances $u(y_\nu)$ by their approximation $u_h(y_\nu) \in V_h$ computed by a numerical solver. We denote by $I_{\Lambda,h}u$ the resulting interpolation polynomial, which belongs to the space

$$V_{\Lambda,h} := \mathbb{P}_\Lambda \otimes V_h. \tag{6.97}$$

In other words, we have

$$I_{\Lambda,h}u := I_\Lambda u_h, \tag{6.98}$$

where

$$u_h : y \mapsto u_h(y) \tag{6.99}$$

is the approximate solution map acting from U into V_h.

We begin by discussing the accuracy of this polynomial approximation. For a given set Λ_n, one first way to estimate the total interpolation error is by writing

$$\|u - I_{\Lambda_n,h}u\|_{L^\infty(U,V)} \leq \|u - I_{\Lambda_n}u\|_{L^\infty(U,V)} + \|I_{\Lambda_n}(u - u_h)\|_{L^\infty(U,V)}. \tag{6.100}$$

The first term is estimated by the results in the previous subsection, such as Corollary 6.7, and for the second term we may write

$$\|I_{\Lambda_n}(u - u_h)\|_{L^\infty(U,V)} \leq \mathbb{L}_{\Lambda_n}\varepsilon(h), \tag{6.101}$$

where $\varepsilon(h)$ is the accuracy of the numerical solver. Under the assumptions of Corollary 6.7, this results in an error estimate of the form

$$\|u - I_{\Lambda_n,h}u\|_{L^\infty(U,V)} \leq Cn^{-s} + n^{\theta+1}\varepsilon(h), \tag{6.102}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

where we have also used Theorem 6.3 for bounding $\mathbb{L}_{\Lambda_n}$.

There is a more efficient way to estimate the error in the case where the approximate solution map u_h may be viewed as the solution map of a discrete parametrized problem of the form

$$\mathcal{P}_h(u_h, a) = 0, \tag{6.103}$$

where $\mathcal{P}_h : V_h \times X \to W$ and with similar properties as the original parametric problem (1.1). Consider, for example, the case of the elliptic equation (1.5) and its Galerkin discretization on V_h defined by (5.3). It is then readily seen that the discrete solution map $a \mapsto u_h(a)$ has the same boundedness and holomorphy properties as the original map $a \mapsto u(a)$. In turn, we obtain the same estimates for the Taylor or Legendre coefficients of the solution map $y \mapsto u_h(y) := u(a(y))$. This type of problem is therefore covered by the following discrete counterpart to Corollary 6.7.

Corollary 6.8. Consider a parametric problem of the form (6.103) such that Assumption A holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. Assume that the univariate sequence $T = (t_k)_{k\geq 0}$ is such that that (6.70) or (6.71) holds for some $\theta \geq 1$. Then, if the assumptions of Theorem 2.9 are satisfied by the map $a \mapsto u_h(a)$, with the open set $\mathcal{O}$ and the bound in (2.116) independent of h, and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, there exists a sequence of nested downward closed sets $(\Lambda_n)_{n\geq 1}$ such that $\#(\Lambda_n) = n$ and such that

$$\|u_h - I_{\Lambda_n} u_h\|_{L^\infty(U,V_h)} \leq Cn^{-s}, \quad n \geq 1, \quad s := \frac{1}{p} - 1, \tag{6.104}$$

where the constant C is independent of h.

Under the assumptions of the above corollary, we may now estimate the total interpolation error by writing

$$\|u - I_{\Lambda_n,h} u\|_{L^\infty(U,V)} \leq \|u_h - I_{\Lambda_n} u_h\|_{L^\infty(U,V)} + \|u - u_h\|_{L^\infty(U,V)}, \tag{6.105}$$

which yields the bound

$$\|u - I_{\Lambda_n,h} u\|_{L^\infty(U,V)} \leq Cn^{-s} + \varepsilon(h). \tag{6.106}$$

This bound is clearly more favourable than (6.102).

We next turn to the estimate of the computational cost, starting with the offline cost. Assuming that Λ_n is given, we want to precompute the coefficients $\alpha_{\nu,h} := \alpha_\nu(u_h) \in V_h$ in the expression

$$I_{\Lambda_n,h} u = \sum_{\nu\in\Lambda_n} \alpha_{\nu,h} H_\nu. \tag{6.107}$$

We begin by computing the discretized instances $u_h(y_\nu)$ for $\nu \in \Lambda_n$, as vectors of size N_h of their coefficients in a given basis of V_h. This has a

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

cost of order nC_h, where C_h is the individual cost of one application of the discrete solver. The coefficients $\alpha_{\nu,h}$ are then computed by recursive application of the formula (6.43) based on these discretized instances. At stage i of this recursion, the coefficient $\alpha_{\nu^i,h} := \alpha_{\nu^i}(u_h)$ is computed by a linear combination of the $i-1$ previous ones, according to

$$\alpha_{\nu^i,h} = u_h(y_{\nu^i}) - I_{\Lambda_{i-1}} u_h(y_{\nu^i}) = u_h(y_{\nu^i}) - \sum_{l=1}^{i-1} \alpha_{\nu^l,h} H_{\nu^l}(y_{\nu^i}). \tag{6.108}$$

Note that in view of the definition of the H_ν, we can compute $H_\nu(y)$ for any $y \in U$ in $|\nu| = \sum_{j\geq 1} \nu_j$ multiplications. Therefore the total cost in this stage is bounded by

$$iN_h + \sum_{l=1}^{i} |\nu^l| = iN_h + \sum_{l=1}^{i} (l-1) = iN_h + i(i-1)/2. \tag{6.109}$$

Summing over $i = 1, \ldots, n$, we thus find that the total offline cost is of order

$$C_{\text{off}} \sim nC_h + n^2 N_h + n^3 \sim nC_h + n^2 N_h. \tag{6.110}$$

Here we neglect, the cost n^3 relative to the computation of the $H_{\nu^l}(y_{\nu^i})$ as well as the cost of $n^2 \log(n)$ needed for the non-adaptive selection of the sets Λ_n, as derived in Section 5.3, since N_h is typically much larger than n.

We finally evaluate the online cost. Since the online stage simply amounts to the combination of the $\alpha_{\nu^i,h}$ for computing the interpolant, we find that this cost is of order

$$C_{\text{on}} \sim nN_h, \tag{6.111}$$

where we have neglected the cost n^2 relative to the computation of the $H_\nu(y)$ for the given $y \in U$.

If ε is a desired order of accuracy, and if we have the error bound (6.106), then one way to reach this accuracy is to take both Cn^{-s} and $\varepsilon(h)$ of the order of ε. Denoting by $h(\varepsilon)$ the inverse function of $\varepsilon(h)$, that is,

$$h(\varepsilon_0) = h_0 \quad \Leftrightarrow \quad \varepsilon(h_0) = \varepsilon_0, \tag{6.112}$$

we thus find that the interpolation algorithm reaches order of accuracy ε at costs

$$C_{\text{off}}(\varepsilon) \sim \varepsilon^{-1/s} C_{h(\varepsilon)} + \varepsilon^{-2/s} N_{h(\varepsilon)} \quad \text{and} \quad C_{\text{on}}(\varepsilon) \sim \varepsilon^{-1/s} N_{h(\varepsilon)}. \tag{6.113}$$

It should be noticed that this algorithm is immune to the curse of dimensionality since the above trade-off between accuracy and complexity is obtained with infinitely many variables.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

7. Taylor approximation

The results established in Section 3 show that effective polynomial approximations (1.50) to the solution map $y \mapsto u(y)$ can be obtained by best n-term truncations of the Taylor series (3.20), for relevant classes of parametric PDEs.

In this section we discuss a strategy, proposed in Chkifa *et al.* (2013), for numerically finding a good n-term Taylor approximation to u. This numerical method rests in part on the effective computation of the Taylor coefficients t_ν. In contrast to the interpolation method discussed in Section 5, the strategy for n-term Taylor approximations is intrusive, and strongly exploits the specific structure of the parametric PDE. In fact, it only applies to the limited, yet relevant, range of parametric problems having the form of a linear operator equation

$$\mathcal{B}(a)u = f, \tag{7.1}$$

where $f \in W$ and $\mathcal{B}(a) \in \mathcal{L}(V, W)$ for a pair of Hilbert spaces (V, W), and where

$$a \mapsto \mathcal{B}(a) \tag{7.2}$$

is a continuous linear map from X to $\mathcal{L}(V, W)$. In other words, the problem map

$$\mathcal{P} : (u, a) \mapsto f - \mathcal{B}(a)u \tag{7.3}$$

is linear in both a and u, up to the constant term f. Recalling our four examples of parametric PDEs discussed in Section 2, that is, equations (1.5), (2.40), (2.58), and (2.66), the first two fall into this category while the last two do not.

Any linear parametric problem between Hilbert spaces V and W can be expressed through a variational formulation (2.38) for a pair of Hilbert spaces $(V, \tilde{V})$, where $\tilde{V}$ is the anti-dual of W. In our present setting, this formulation is as follows: Find $u(a) \in V$ such that

$$B(u(a), v; a) = L(v), \quad v \in \tilde{V}, \tag{7.4}$$

where $B(\cdot, \cdot; a)$ and L are continuous sesquilinear and antilinear forms over $V \times \tilde{V}$ and $\tilde{V}$ respectively, and where, throughout this section, we make the additional assumption that

$$a \mapsto B(\cdot, \cdot; a) \tag{7.5}$$

is a continuous *linear* map from X to $\mathfrak{B}$, the set of continuous sesquilinear forms over $V \times \tilde{V}$. We work under the following assumption.

Assumption AL. The parameter set $\mathcal{A}$ has a complete affine representer $(\psi_j)_{j \geq 1}$, and the sesquilinear form $B(\cdot, \cdot; a)$ satisfies the inf-sup conditions (2.30) for all $a \in a(U)$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Note that this assumption requires problem (7.1) to have a solution for any $a \in a(U)$ and for all $f \in W$, in contrast to Assumption A which requires it to have a solution for all $a \in a(U)$ but only for the given $f \in W$. Under this assumption, the solution map

$$y \mapsto u(y) = \mathcal{B}(a(y))^{-1} f \tag{7.6}$$

is well defined over U. Examples of problems falling into this category are the elliptic equation (1.5) and the parabolic equation (2.40) under the uniform elliptic assumption UEA(r), as discussed in Section 2.4. Since all maps in the chain

$$a \mapsto \mathcal{B}(a) \mapsto \mathcal{B}(a)^{-1} \mapsto u(a) = \mathcal{B}(a)^{-1} f \tag{7.7}$$

are infinitely Fréchet-differentiable at $a \in a(U)$, and since $y \mapsto a(y)$ is affine, the partial derivatives

$$\partial^\nu u := \left(\prod_{j \geq 1} \frac{\partial}{\partial^{\nu_j} y_j} \right) u \tag{7.8}$$

exist at every $y \in U$ and for all $\nu \in \mathcal{F}$, and therefore so do the Taylor coefficients $t_\nu = \partial^\nu u(0)/\nu! \in V$ for all $\nu \in \mathcal{F}$.

We first show in Section 7.1 that these coefficients can be computed by a simple recursive procedure that takes advantage of the linear structure of the problem. When the truncation sets Λ_n are downward closed, this procedure computes the coefficients $(t_\nu)_{\nu \in \Lambda_n}$ at the cost of solving a linear problem n times with operator $\overline{\mathcal{B}} := \mathcal{B}(\overline{a})$. Similar to the interpolation method in Section 5, the downward closed sets $(\Lambda_n)_{n \geq 1}$ for which we compute the coefficients t_ν can either be chosen in an *a priori* manner, based on the *a priori* estimates for the coefficients $\|t_\nu\|_V$, or adaptively built. Various adaptive selection strategies for finding the sets Λ_n are proposed in Section 7.2, and a convergence analysis is given for one of them in Section 7.3 in the particular case of the elliptic equation (1.5): we show that if the sequence $(\|t_\nu\|_V)_{\nu \in \mathcal{F}}$ belongs to the space $\ell^p_m(\mathcal{F})$ defined in Section 3.8 for some $p < 1$, then the adaptive strategy generates downward closed sets $(\Lambda_n)_{n \geq 1}$ such that the truncated Taylor series converges in $L^\infty(U, V)$ with the expected rate n^{-s}, where $s := \frac{1}{p} - 1$. Recall that Theorem 3.25 shows that $(\|t_\nu\|_V)_{\nu \in \mathcal{F}} \in \ell^p_m(\mathcal{F})$ whenever the assumptions of Theorem 2.8 hold and $(\|\psi_j\|_X)_{j \geq 1} \in \ell^p(\mathbb{N})$. The space discretization and computational cost of the adaptive and non-adaptive strategies are discussed in Section 7.4.

7.1. Recursive computations of the Taylor coefficients

Since $B(\cdot, \cdot, a)$ is defined for all $a \in X$, we may introduce the sesquilinear forms

$$\overline{B}(\cdot, \cdot) := B(\cdot, \cdot; \overline{a}) \quad \text{and} \quad B_j(\cdot, \cdot) := B(\cdot, \cdot; \psi_j). \tag{7.9}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Then the solution map $y \mapsto u(y) = u(a(y))$ is defined by

$$B(u(y), v; y) = L(v), \quad v \in \tilde{V}, \tag{7.10}$$

where

$$B(\cdot, \cdot; y) := \overline{B}(\cdot, \cdot) + \sum_{j \geq 1} y_j B_j(\cdot, \cdot). \tag{7.11}$$

Note that $\overline{B}$, as well as each individual B_j, is bounded, that is, $\overline{B}$ belongs to $\mathfrak{B}$. In addition, we know from Assumption AL that $\overline{B}$ satisfies the inf-sup conditions (2.30), and so does $B(\cdot, \cdot; y)$ for any $y \in U$. The following result shows that the Taylor coefficients of the solution map $y \mapsto u(y)$ satisfy simple equations allowing us to compute them in a recursive way.

Lemma 7.1. Consider a parametric problem of the form (7.1) such that Assumption AL holds for a suitable affine representer $(\psi_j)_{j \geq 1}$. Then the Taylor coefficients of the solution map $y \mapsto u(y)$ satisfy the equations

$$\overline{B}(t_\nu, v) = L_\nu(v), \quad v \in \tilde{V}, \tag{7.12}$$

where $L_\nu(v) := L(v)$ if $\nu = 0$ is the null multi-index, and

$$L_\nu(v) := - \sum_{j \in \mathrm{supp}(\nu)} B_j(t_{\nu - e_j}, v), \tag{7.13}$$

when $\nu \in \mathcal{F} - \{0\}$, where $e_j := (0, \ldots, 0, 1, 0, \ldots)$ is the Kronecker sequence with 1 at position j.

Proof. The case $\nu = 0$ is immediate since $t_0 = u(0)$ and $B(0) = \overline{B}$. For the other values of ν, we apply the operator ∂^ν to the equation

$$L(v) = B(u(y), v; y). \tag{7.14}$$

Since L does not depend on y, and due to the affine form (7.11) of $B(\cdot, \cdot, y)$, we obtain by the multivariate Leibniz rule

$$\begin{aligned} 0 &= \partial^\nu (B(u(y), v; y)) \\ &= \partial^\nu \left(\overline{B}(u(y), v) + \sum_{j \geq 1} y_j B_j(u(y), v) \right) \\ &= \overline{B}(\partial^\nu u(y), v) + \sum_{j \geq 1} y_j B_j(\partial^\nu u(y), v) + \sum_{j \in \mathrm{supp}(\nu)} \nu_j B_j(\partial^{\nu - e_j} u(y), v). \end{aligned}$$

At $y = 0$, this gives

$$\overline{B}(\partial^\nu u(0), v) = - \sum_{j \in \mathrm{supp}(\nu)} \nu_j B_j(\partial^{\nu - e_j} u(0), v). \tag{7.15}$$

Dividing by $\nu! = (\nu - e_j)! \nu_j$, this gives (7.12). □

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

For further purposes, we give the particular expression of the equations (7.12) in the case of the elliptic equation (1.5). In this case, $V = \tilde{V} = H^1_0(D)$ and the sesquilinear forms are given by

$$\overline{B}(u,v) := \int_D \overline{a}\nabla u \cdot \nabla v, \tag{7.16}$$

and

$$B_j(u,v) := \int_D \psi_j \nabla u \cdot \nabla v. \tag{7.17}$$

Therefore, for $\nu \in \mathcal{F} - \{0\}$, the coefficient t_ν is the solution of the boundary value problem

$$\int_D \overline{a}\nabla t_\nu \cdot \nabla v = - \sum_{j \in \mathrm{supp}(\nu)} \int_D \psi_j \nabla t_{\nu - e_j} \cdot \nabla v, \quad v \in V. \tag{7.18}$$

Note that, in the particular case where $\nu = e_j$, that is, when $t_\nu = \partial_{y_j} u(0)$, this relation has the form

$$\int_D \overline{a}\nabla \partial_{y_j} u(0) \cdot \nabla v = - \sum_{j \in \mathrm{supp}(\nu)} \int_D \psi_j \nabla u(0) \cdot \nabla v, \quad v \in V, \tag{7.19}$$

which corresponds to the general expression of the Fréchet derivative $\mathrm{d}u(a)$ obtained in Section 2.1, applied to $h = \psi_j$.

Lemma 7.1 shows that if Λ_n is a downward closed set, then it is possible to compute the n Taylor coefficients $(t_\nu)_{\nu \in \Lambda_n}$ by solving exactly n linear problems of the form (7.12). This is done by writing $\Lambda_n = \{\nu^1, \ldots, \nu^n\}$, where the order is such that all sections $\{\nu^1, \ldots, \nu^k\}$ are downward closed sets for $k = 1, \ldots, n$ and by recursively computing the t_ν in this order. Then, the right-hand side of the problem (7.12) for computing t_{ν^k} only depends on the t_{ν^i} for $i < k$ which have already been computed.

In practice, these linear problems can only be solved approximately via space discretization, for example using the finite element method. This induces an error in the computation of the Taylor coefficients. We deal with this issue in Section 7.4 and assume for the moment that these problems are solved exactly.

As already explained in Section 5.3, one non-adaptive strategy consist in defining Λ_n as the set of indices corresponding to the n largest *a priori* estimates e_ν for the Taylor coefficients or the n largest surrogate s_ν such that the sequence $(s_\nu)_{\nu \in \mathcal{F}}$ is monotone non-increasing. We have observed that the total cost of identifying Λ_n is then of the order $n^2 \log(n)$, which is negligible compared to the computation of the approximation polynomial. We then know from the results in Section 3.8 that if the assumptions of Theorem 2.8 hold for the solution map $a \mapsto u(a)$, and if in addition

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

$(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, we have the convergence rate

$$\sup_{y\in U}\left\|u(y) - \sum_{\nu\in\Lambda_n} t_\nu y^\nu\right\|_V \leq Cn^{-s}, \quad s := \frac{1}{p} - 1, \tag{7.20}$$

using these *a priori* selected $(\Lambda_n)_{n\geq 1}$.

In the next subsection we discuss adaptive strategies for the selection of the sets $(\Lambda_n)_{n\geq 1}$. Given the above result on *a priori* choices for the Λ_n, one might ask why we should even consider adaptive strategies. The answer is that it may be that the best n-term Taylor approximations of u actually perform much better than the proven rate $O(n^{-s})$ established using the *a priori* chosen Λ_n. This is possible, since we do not have any results that say the $O(n^{-s})$ rate is best possible under the assumption $(\|\psi_j\|)_{j\geq 1} \in \ell^p(\mathbb{N})$ and since the estimates of $\|t_\nu\|_V$ by the computable surrogate s_ν could be too pessimistic.

7.2. Adaptive algorithms

For any downward closed set Λ, given the Taylor coefficients $(t_\nu)_{\nu\in\Lambda}$, the equations (7.12) allow us to compute the Taylor coefficients t_ν for $\nu \in N(\Lambda)$, where $N(\Lambda)$ is the set of neighbours of Λ defined in (5.13). This suggests an adaptive algorithm in the same line as the greedy interpolation algorithms proposed in Section 5.2. Once again, we start with an idealized version of this algorithm which cannot be applied as such. We discuss more practical versions later.

Greedy Taylor algorithm. We start with the null multi-index $\Lambda_1 := \{0\}$. Assuming that Λ_{n-1} has been selected and that the $(t_\nu)_{\nu\in\Lambda_{n-1}}$ have been computed, we compute the t_ν for $\nu \in N(\Lambda_{n-1})$. We then set

$$\nu^n := \arg\max\{\|t_\nu\|_V : \nu \in N(\Lambda_{n-1})\}, \tag{7.21}$$

and define $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\}$.

The rationale behind this procedure is that if the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ were monotone non-increasing, it would automatically select the n largest terms of this sequence. Similar to the greedy interpolation algorithm, the above algorithm needs to be modified in the following two ways.

(i) In order to guarantee finite complexity of the search in the case of infinitely many variables, the infinite set $N(\Lambda)$ should be replaced with a finite one, such as the set of anchored neighbours $\tilde{N}(\Lambda)$ defined by (5.14).

(ii) In order to guarantee convergence, one should alternate the selection of ν^n using (6.51) with a selection rule ensuring that all indices are eventually picked.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

This results in the following algorithm, which is similar to the alternating greedy interpolation algorithm.

Alternating greedy Taylor algorithm. We start with the null multi-index $\Lambda_1 := \{0\}$. Assuming that Λ_{n-1} has been selected and that the $(t_\nu)_{\nu\in\Lambda_{n-1}}$ have been computed, we compute the t_ν for $\nu \in \tilde{N}(\Lambda_{n-1})$. If n is even, we set

$$\nu^n := \arg\max\{\|t_\nu\|_V : \nu \in \tilde{N}(\Lambda_{n-1})\}, \tag{7.22}$$

and if n is odd, we set ν^n as in (6.54). We then define $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\}$.

It is easily checked that the selection of ν^n by (6.54) ensures that the sequence $(\Lambda_n)_{n\geq 1}$ is an exhaustion of $\mathcal{F}$. Therefore, if the Taylor series (3.20) converges unconditionally to u in $L^\infty(U,V)$ or in any other norm, then the approximations produced by the greedy algorithm converge to u in the same norm. However, we do not have much information on the rate of convergence.

In view of the results obtained in Section 3, a legitimate objective is to build adaptive algorithms that can be proved to converge at a rate comparable to that for best n-term Taylor approximations, for relevant classes of parametric PDEs. Let us point out that a similar objective can be attained when considering the spatial discretization of a *single* PDE, by either adaptive wavelet methods (Cohen, Dahmen and DeVore 2000, Cohen, Dahmen and DeVore 2002, Gantumur, Harbrecht and Stevenson 2007) or by adaptive finite element methods (Morin, Nochetto and Siebert 2000, Binev, Dahmen and DeVore 2004, Stevenson 2007). More precisely, these papers show that specific refinement strategies based on *a posteriori* analysis generate adaptive wavelet sets or adaptive meshes such that the approximate solution converges with the optimal algebraic rate allowed by the exact solution, as the number of wavelets or elements grows. One key tool in these algorithms is the use of a refinement procedure which guarantees that the error decreases by a fixed amount after the refinement is performed. This procedure is called *bulk chasing*, and requires that, in general, more than one wavelet/element is added/refined at each step.

In the present context of n-term Taylor approximation, it is possible to introduce similar bulk chasing procedures, which, in general, require the selection of more than one term from the Taylor expansion at each step. Iterating these bulk chasing procedure produces a nested sequence $(\Lambda^k)_{k\geq 1}$ of downward closed sets with $\Lambda^1 = \{0\}$. Here, we use the notation Λ^k instead of Λ_k in order to stress that $\#(\Lambda^k)$ is in general larger than k. When we want to index these sets by their cardinality, we may define the sets

$$\Lambda_n := \Lambda^k \quad \text{for } n = n(k) := \#(\Lambda^k), \tag{7.23}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

which are indexed by the integers $n \in \{n(k) : k \geq 1\}$. Of course, some indices n are missed in the Λ_n notation, but these can be filled in by simply repeating the sets Λ^k. The resulting sequence $(\Lambda_n)_{n\geq 1}$ then satisfies $\#(\Lambda_n) \leq n$.

We now discuss one specific procedure of this type for which it is proved, in the particular case of the elliptic problem (1.5), that the resulting adaptive approximation converges with an algebraic convergence rate matching the rate established when keeping the largest n-terms in the Taylor expansion. For this purpose, for any finite downward closed set Λ we introduced its *margin* $M(\Lambda)$, defined by

$$M(\Lambda) := \{\nu \notin \Lambda : \exists j \in \mathrm{supp}(\nu) : \nu - e_j \in \Lambda\}. \tag{7.24}$$

Note that the set of neighbours $N(\Lambda)$ can be defined by

$$N(\Lambda) := \{\nu \notin \Lambda : \forall j \in \mathrm{supp}(\nu) : \nu - e_j \in \Lambda\}. \tag{7.25}$$

Therefore we have $N(\Lambda) \subset M(\Lambda)$, and this inclusion is generally strict. Still, for any downward closed set Λ, given the Taylor coefficients $(t_\nu)_{\nu\in\Lambda}$, the relations (7.12) allow us to compute the Taylor coefficients t_ν for $\nu \in M(\Lambda)$.

For the rest of this subsection we assume that the Taylor coefficients of the solution map are ℓ^2-summable, that is,

$$\sum_{\nu\in\mathcal{F}} \|t_\nu\|_V^2 < \infty. \tag{7.26}$$

Note that, according to Theorem 3.9, this holds if the assumptions of Theorem 2.8 are satisfied and if in addition $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$. For any set $S \subset \mathcal{F}$, we introduce the quadratic energy

$$e(S) := \sum_{\nu\in S} \|t_\nu\|_V^2. \tag{7.27}$$

We also define for any finite downward closed set Λ the quadratic error

$$\sigma(\Lambda) := \sum_{\nu\notin\Lambda} \|t_\nu\|_V^2. \tag{7.28}$$

Note that since the functions $y \mapsto y^\nu$ do not form an orthonormal system, this quantity differs from the mean-square error between u and its truncated Taylor series.

The bulk chasing procedure consists in building the new set Λ^k by adding to Λ^{k-1} a subset S^{k-1} of the margin $M^{k-1} := M(\Lambda^{k-1})$ which captures a prescribed portion of its energy, that is, such that

$$e(S^{k-1}) \geq \theta e(M^{k-1}), \tag{7.29}$$

for some fixed $0 < \theta < 1$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The objective of this procedure is to reduce the quadratic error $\sigma(\Lambda^{k-1})$ by a fixed amount. This will be achieved provided that the Taylor expansion (3.20) satisfies a so-called 'saturation property', which is analogous to that which is sometimes established in order to prove convergence of adaptive finite element methods; see Dörfler (1996), Morin *et al.* (2000) and Binev *et al.* (2004). In Section 7.3, we establish the validity of this property in the case of the parametric elliptic problem (1.5), provided that the uniform ellipticity property UEA(r) holds for some $r > 0$. For now, we start from this property viewed as a general assumption in order to analyse the convergence of adaptive algorithms based on bulk chasing.

Saturation property. There exists a fixed constant $\delta > 1$ such that, for any finite downward closed set Λ,

$$\sigma(\Lambda) \leq \delta e(M), \tag{7.30}$$

where $M = M(\Lambda)$.

If this property holds and if we use the bulk chasing procedure to construct $(\Lambda^k)_{k\geq 1}$, we may write

$$\sigma(\Lambda^k) = \sigma(\Lambda^{k-1}) - e(S^{k-1}) \leq \sigma(\Lambda^{k-1}) - \theta e(M^{k-1}) \leq \kappa\sigma(\Lambda^{k-1}), \tag{7.31}$$

where

$$\kappa = 1 - \frac{\theta}{\delta} < 1. \tag{7.32}$$

Therefore, the quadratic error decreases by a fixed amount at every step of the algorithm, and in particular, at step k we have

$$\sigma(\Lambda^k) \leq C\kappa^k, \tag{7.33}$$

where $C := \kappa^{-1} \sum_{\nu\neq 0} \|t_\nu\|_V^2$.

The set S^{k-1} should be chosen as small as possible, but we need to ensure that the new set $\Lambda^k := \Lambda^{k-1} \cup S^{k-1}$ is still downward closed. This is achieved by the following algorithm, which we first present in an idealized form that cannot be applied as such.

Bulk chasing Taylor algorithm. Having fixed $0 < \theta < 1$, we start with the null multi-index $\Lambda^1 := \{0\}$. Assume that Λ^{k-1} has been selected and that the coefficients t_ν have been computed for $\nu \in \Lambda^{k-1}$. For all $\nu \in M^{k-1}$ we compute the t_ν and the quantities

$$m_\nu := \max\{\|t_{\tilde\nu}\|_V : \tilde\nu \geq \nu \text{ and } \tilde\nu \in M^{k-1}\}. \tag{7.34}$$

We define S^{k-1} as the set of indices $\nu \in M^{k-1}$ corresponding to the l largest m_ν, for the smallest value of l such that the bulk condition (7.29) is met. We then define

$$\Lambda^k := \Lambda^{k-1} \cup S^{k-1}. \tag{7.35}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Remark 7.2. The quantity m_ν is introduced in order to guarantee that the new set Λ^k is downward closed. This can always be ensured due to the monotonicity property,

$$\nu, \tilde{\nu} \in M^{k-1} \quad \text{and} \quad \tilde{\nu} \geq \nu \quad \Rightarrow \quad m_{\tilde{\nu}} \leq m_\nu. \tag{7.36}$$

Another option would be to define S^{k-1} as the smallest subset of M^k such that the bulk condition (7.29) is met and such that $\Lambda^k := \Lambda^{k-1} \cup S^{k-1}$ is downward closed. However, it is not clear whether there is a simple algorithm for determining such a set.

The above bulk chasing algorithm is not computationally feasible due to the fact that the margin $M(\Lambda)$ of a finite downward closed set Λ has infinite cardinality in the case of countably many variable y_j. We want to modify it by restricting the computation of the t_ν and the bulk search to a finite subset of M^k. In order to accomplish this, we take the usual view in numerical computation, where we are given a target accuracy $\varepsilon > 0$ and we want the algorithm to achieve this accuracy as efficiently as possible. So, in our modified algorithm, we design the procedure so that the algorithm terminates when $\sigma(\Lambda^k) \leq C\varepsilon$, for some fixed constant C to be specified later. We begin by introducing a procedure that computes from the Taylor coefficients $(t_\nu)_{\nu\in\Lambda}$ a finite version of the margin $M = M(\Lambda)$ which captures its energy up to accuracy ε. If Λ is a finite downward closed set with margin M, then

$$\tilde{M} = \text{SPARSE}(\Lambda, (t_\nu)_{\nu\in\Lambda}, \varepsilon), \tag{7.37}$$

is a finite subset of M such that $\Lambda \cup \tilde{M}$ is, by definition, a downward closed set and such that

$$e(M \setminus \tilde{M}) \leq \varepsilon. \tag{7.38}$$

At the end of Section 7.3 we present one practical realization of such a procedure in the particular case of the elliptic equation (1.5). Our modified algorithm is the following.

Bulk chasing Taylor algorithm with ε-accuracy. Having fixed $0 < \theta < 1$, we start with the null multi-index $\Lambda^1 := \{0\}$. Assume that Λ^{k-1} has been selected and that the coefficients t_ν have been computed for $\nu \in \Lambda^{k-1}$. We define

$$\tilde{M}^{k-1} = \text{SPARSE}(\Lambda^{k-1}, (t_\nu)_{\nu\in\Lambda^{k-1}}, \varepsilon), \tag{7.39}$$

For all $\nu \in \tilde{M}^{k-1}$ we compute the t_ν and the quantities

$$m_\nu := \max\{\|t_{\tilde{\nu}}\|_V : \tilde{\nu} \geq \nu \text{ and } \tilde{\nu} \in \tilde{M}^{k-1}\}. \tag{7.40}$$

We define S^{k-1} as the set of indices $\nu \in \tilde{M}^{k-1}$ corresponding to the l largest m_ν, for the smallest value of l such that the bulk condition

$$e(S^{k-1}) \geq \theta e(\tilde{M}^{k-1}) \tag{7.41}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

is met. We then define

$$\Lambda^k := \Lambda^{k-1} \cup S^{k-1}. \tag{7.42}$$

The algorithm is stopped when $e(\tilde{M}^k) \leq 2\theta\varepsilon$.

The same computation as in (7.31), now using (7.41) and (7.38) together with the saturation property (7.30), leads to the reduction inequality

$$\sigma(\Lambda^k) \leq \sigma(\Lambda^{k-1}) - \theta e(\tilde{M}^{k-1}) \leq \kappa\sigma(\Lambda^{k-1}) + \theta\varepsilon. \tag{7.43}$$

Therefore, after sufficiently many steps k, we have

$$e(\tilde{M}^k) \leq e(M^k) \leq \sigma(\Lambda^k) \leq 2\theta\varepsilon, \tag{7.44}$$

and thus the algorithm terminates. The final step may occur before $\sigma(\Lambda^k) \leq 2\theta\varepsilon$, but we are still ensured by the saturation property (7.30) that

$$\sigma(\Lambda^k) \leq \delta e(M^k) \leq \delta(e(\tilde{M}^k) + \varepsilon) \leq C\varepsilon, \quad C := \delta(2\theta + 1), \tag{7.45}$$

and thus we have reached the stated order of accuracy for the quadratic error.

However, this does not settle the convergence analysis of the algorithm. First, we want to relate the accuracy $\sigma(\Lambda^k)$ with the number of terms $\#(\Lambda^k)$ in the Taylor approximation via a quantitative convergence estimate. Second, we also want to retrieve convergence estimates for the error between u and its truncated Taylor approximation measured in the $L^\infty(U, V)$-metric. This is the purpose of the next subsection.

7.3. Convergence analysis of adaptive algorithms

We know that if the Taylor series converges conditionally to u in $L^\infty(U, V)$ and if the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to the sequence space $\ell^p_m(\mathcal{F})$ for some $0 < p < 1$, then there exists a sequence of downward closed sets $(\Lambda_n)_{n\geq 1}$ such that $\#(\Lambda_n) = n$ and such that the uniform error bound

$$\sup_{y\in U}\left\|u(y) - \sum_{\nu\in\Lambda_n} t_\nu y^\nu\right\|_V \leq Cn^{-s}, \quad s := \frac{1}{p} - 1, \tag{7.46}$$

holds for all $n \geq 1$. This holds in particular if the assumptions of Theorem 2.8 hold and if, in addition, $(\|\psi_j\|_X)_{j\geq 1} \in \ell^p(\mathbb{N})$. In this subsection we show that this benchmark rate is met by the two previously described bulk chasing Taylor algorithms, provided that the saturation assumption hold. We begin with a result that describes the rate of decay of the quadratic error $\sigma(\Lambda_n)$.

Theorem 7.3. Consider a parametric problem of the form (7.1) such that Assumption AL holds for a suitable affine representer $(\psi_j)_{j\geq 1}$. Assume that $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to the sequence space $\ell^p_m(\mathcal{F})$ for some $0 < p < 1$,

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and that the saturation property holds. Then, for the bulk chasing Taylor algorithm, the convergence estimate

$$\sigma(\Lambda_n) \leq C\|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2 n^{-2r}, \quad r := \frac{1}{p} - \frac{1}{2}, \tag{7.47}$$

holds for all $n \in \{n(k) : k \geq 1\}$, where we have used the convention (7.23). The constant C depends on p, θ and δ. The same estimate holds for the bulk chasing Taylor algorithm with ε-accuracy for all $n \in \{n(k) : 1 \leq k \leq k(\varepsilon)\}$, where $k(\varepsilon)$ is the step where the algorithm terminates.

Proof. We begin by considering the bulk chasing Taylor algorithm. We first control the cardinality of the set S^{k-1}, which is added to Λ^{k-1}, for any $k > 1$. Recall that this set is obtained by picking indices $\nu \in M^{k-1}$ corresponding to the l largest m_ν defined by (7.34), for the smallest value of l such that the bulk condition (7.29) is met. Let $\tilde{S}^{k-1} \subset S^{k-1}$ denote the set corresponding to the $l-1$ largest m_ν for this value of l. Since the bulk condition is not met by $\tilde{S}^{k-1}$, we have

$$(1-\theta)e(M^{k-1}) \leq e(M^{k-1}) - e(\tilde{S}^{k-1}) = \sum_{\nu\in M^{k-1}\setminus\tilde{S}^{k-1}} \|t_\nu\|_V^2. \tag{7.48}$$

On the one hand, using Stechkin's Lemma 3.6, we find that

$$\sum_{\nu\in M^{k-1}\setminus\tilde{S}^{k-1}} \|t_\nu\|_V^2 \leq \sum_{\nu\in M^{k-1}\setminus\tilde{S}^{k-1}} m_\nu^2 \leq \|(m_\nu)_{\nu\in M^{k-1}}\|_{\ell^p}^2 l^{-2r} \tag{7.49}$$

and therefore, using the fact that m_ν is dominated by the monotone majorant of $\|t_\nu\|_V$, and that $l = \#(S^{k-1})$, we obtain

$$\sum_{\nu\in M^{k-1}\setminus\tilde{S}^{k-1}} \|t_\nu\|_V^2 \leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2 (\#(S^{k-1}))^{-2r}. \tag{7.50}$$

Combining this with (7.48), we find that

$$(1-\theta)e(M^{k-1}) \leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2 (\#(S^{k-1}))^{-2r}. \tag{7.51}$$

Using the saturation property, it follows that

$$\frac{1-\theta}{\delta}\sigma(\Lambda^{k-1}) \leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2 (\#(S^{k-1}))^{-2r}, \tag{7.52}$$

or equivalently

$$\#(S^{k-1}) \leq \left(\frac{\delta}{1-\theta}\right)^{1/2r} \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^{1/r} \sigma(\Lambda^{k-1})^{-1/2r}. \tag{7.53}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

For any $k > 1$, we may thus control the cardinality of Λ^k by writing

$$\#(\Lambda^k) = 1 + \sum_{l=1}^{k-1} \#(S^l) \leq 1 + \left(\frac{\delta}{1-\theta}\right)^{1/2r} \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^{1/r} \sum_{l=1}^{k-1} \sigma(\Lambda^l)^{-1/2r}. \tag{7.54}$$

On the other hand, we know from (7.31) that $\sigma(\Lambda^l) \geq \kappa^{l-k}\sigma(\Lambda^k)$ with $\kappa := 1 - \delta/\theta$, and therefore

$$\#(\Lambda^k) \leq 1 + \left(\frac{\delta}{1-\theta}\right)^{1/2r} \frac{\kappa^{1/2r}}{1-\kappa^{1/2r}} \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^{1/r} \sigma(\Lambda^k)^{-1/2r}. \tag{7.55}$$

This can be rewritten as

$$\sigma(\Lambda^k) \leq \frac{\delta}{1-\theta} \frac{\kappa}{(1-\kappa^{1/2r})^{2r}} \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2 (\#(\Lambda^k) - 1)^{-2r}. \tag{7.56}$$

Using the inequality $\#(\Lambda^k) - 1 \geq \frac{1}{2}\#(\Lambda^k)$ in the case where $k > 1$, we have thus established (7.47) with constant

$$C := 2^{2r} \frac{\delta}{1-\theta} \frac{\kappa}{(1-\kappa^{1/2r})^{2r}} > 1.$$

If $k = 1$, we simply write

$$\sigma(\Lambda^1) \leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^2}^2 \leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2, \tag{7.57}$$

which shows that (7.47) also holds in this case.

We next consider the bulk chasing Taylor algorithm with ε-accuracy and explain how (7.47) can be established by a similar argument for $n = n(k)$, with $1 \leq k \leq k(\varepsilon)$, albeit increasing the constant C. First, with exactly the same reasoning that led to (7.51), we obtain the estimate

$$(1-\theta)e(\tilde{M}^{k-1}) \leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^2 (\#(S^{k-1}))^{-2r}. \tag{7.58}$$

We next observe that since $e(\tilde{M}^{k-1}) \geq 2\theta\varepsilon$ for $k \leq k(\varepsilon)$, we have the modified saturation property

$$\sigma(\Lambda^{k-1}) \leq \delta e(M^{k-1}) \leq \delta(e(\tilde{M}^{k-1}) + \varepsilon) \leq \tilde{\delta} e(\tilde{M}^{k-1}), \tag{7.59}$$

with $\tilde{\delta} := \delta(1 + (1/2\theta))$. By the same reasoning, for any $n > 1$, we obtain the bound

$$\#(\Lambda^k) = 1 + \sum_{l=1}^{k-1} \#(S^l) \leq 1 + \left(\frac{\tilde{\delta}}{1-\theta}\right)^{1/2r} \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m}^{1/r} \sum_{l=1}^{k-1} \sigma(\Lambda^l)^{-1/2r}. \tag{7.60}$$

The saturation property implies that $\sigma(\Lambda^l) \geq \tilde{\kappa}^{l-l}\sigma(\Lambda^k)$ with $\tilde{\kappa} := 1 - (\tilde{\delta}/\theta)$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Therefore, by the same reasoning, we reach (7.47) with the larger constant

$$C := 2^{2r} \frac{\tilde{\delta}}{1-\theta} \frac{\tilde{\kappa}}{(1-\tilde{\kappa}^{1/2r})^{2r}} > 1. \qquad \square$$

Our next result shows that the benchmark rate (7.46) is met under the same assumptions as those of the above theorem.

Theorem 7.4. Consider a parametric problem of the form (7.1) such that the assumptions of Theorem 7.3 hold, and such that in addition the Taylor series converges conditionally to u in $L^\infty(U,V)$. Then, for the bulk chasing Taylor algorithm, the uniform convergence estimate

$$\sup_{y\in U} \left\| u(y) - \sum_{\nu\in\Lambda_n} t_\nu y^\nu \right\|_V \leq C \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m} n^{-s}, \quad s := \frac{1}{p} - 1, \tag{7.61}$$

holds for all $n \in \{n(k) : k \geq 1\}$. The constant C depends on p, θ and δ. The same estimate holds for the bulk chasing Taylor algorithm with ε-accuracy for all $n \in \{n(k) : 1 \leq k \leq k(\varepsilon)\}$, where $k(\varepsilon)$ is the step where the algorithm terminates.

Proof. It suffices to prove that (7.47) implies (7.61) for the same value of n, up to a change in the constant C. Since the Taylor series converges conditionally and since $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^1(\mathcal{F})$, this series also converges unconditionally. We thus have

$$\sup_{y\in U} \left\| u(y) - \sum_{\nu\in\Lambda^k} t_\nu y^\nu \right\|_V \leq \sum_{\nu\notin\Lambda^k} \|t_\nu\|_V. \tag{7.62}$$

For $n = n(k) = \#(\Lambda_n) = \#(\Lambda^k)$, we consider the set Λ_n^* of the indices corresponding to the n largest $\|t_\nu\|_V$. Using Lemma 3.6, the Cauchy–Schwarz inequality and (7.47), we write

$$\begin{aligned}
\sum_{\nu\notin\Lambda_n} \|t_\nu\|_V &\leq \sum_{\nu\notin\Lambda_n^*} \|t_\nu\|_V + \sum_{\nu\in\Lambda_n^*\setminus\Lambda_n} \|t_\nu\|_V \\
&\leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p}(n+1)^{-s} + n^{1/2} e(\Lambda_n^* \setminus \Lambda_n)^{1/2} \\
&\leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p}(n+1)^{-s} + n^{1/2} \sigma(\Lambda_n)^{1/2} \\
&\leq \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p}(n+1)^{-s} + n^{1/2} C^{1/2} \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m} n^{-r} \\
&\leq (1 + C^{1/2}) \|(\|t_\nu\|_V)_{\nu\in\mathcal{F}}\|_{\ell^p_m} n^{-s},
\end{aligned}$$

which confirms (7.61). $\square$

The above theorem assumes that the saturation property is valid. We next prove that this property always holds in the particular case of the elliptic problem (1.5). Recall that we then have $V = H^1_0(D)$ and $X = L^\infty(D)$. The

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

saturation property turns out to be a consequence of the uniform ellipticity assumption UEA(r). In order to see this, we introduce the norm

$$\|v\|_{\overline{a}} := \left(\int_D \overline{a}|\nabla v|^2\right)^2, \tag{7.63}$$

which is equivalent to the V-norm under UEA(r), since we then have

$$r\|v\|_V^2 \le \overline{a}_{\min}\|v\|_V^2 \le \|v\|_{\overline{a}}^2 \le \overline{a}_{\max}\|v\|_V^2, \tag{7.64}$$

where

$$\overline{a}_{\min} := \min_{x\in D} \overline{a}(x) \quad \text{and} \quad \overline{a}_{\max} := \max_{x\in D} \overline{a}(x) = \|\overline{a}\|_X.$$

For $\nu \in \mathcal{F}$ and $j \ge 1$, we use the notation

$$d_\nu := \|t_\nu\|_{\overline{a}}^2, \tag{7.65}$$

and

$$d_{\nu,j} := \int_D |\psi_j|\,|\nabla t_\nu|^2. \tag{7.66}$$

The proof of the saturation property uses the following lemma, which relates the above quantities.

Lemma 7.5. Consider a parametric problem of the type (1.5), with affine representer $(\psi_j)_{j\ge1}$ such that UEA(r) holds for some $r > 0$. Then we have

$$\sum_{j\ge1} d_{\nu,j} \le \gamma d_\nu, \quad \gamma := 1 - \frac{r}{\overline{a}_{\max}} < 1, \tag{7.67}$$

and

$$d_\nu \le \alpha \sum_{j\in\operatorname{supp}(\nu)} d_{\nu-e_j,j}, \quad \alpha := \frac{\overline{a}_{\max}}{r + \overline{a}_{\max}} < 1, \tag{7.68}$$

where e_j is the Kronecker sequence with 1 at position j.

Proof. The uniform ellipticity assumption implies that, for all $x \in D$,

$$\sum_{j\ge1} |\psi_j(x)| \le \overline{a}(x) - r \le \gamma\overline{a}(x), \tag{7.69}$$

which implies (7.67). On the other hand, we take $v = t_\nu$ in (7.18), which gives

$$d_\nu = -\sum_{j\in\operatorname{supp}(\nu)} \int_D \psi_j \nabla t_{\nu-e_j} \nabla t_\nu, \tag{7.70}$$

and therefore

$$d_\nu \le \frac{1}{2} \sum_{j\in\operatorname{supp}(\nu)} \int_D |\psi_j|\,|\nabla t_{\nu-e_j}|^2 + \frac{1}{2} \sum_{j\in\operatorname{supp}(\nu)} \int_D |\psi_j|\,|\nabla t_\nu|^2. \tag{7.71}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Using (7.69) in the second term of (7.71) gives

$$\left(1-\frac{\gamma}{2}\right)d_\nu \le \frac{1}{2}\sum_{j\in\mathrm{supp}(\nu)}\int_D |\psi_j|\,|\nabla t_{\nu-e_j}|^2, \tag{7.72}$$

from which we derive (7.68). □

We are now in position to establish the saturation property for the elliptic problem (1.5) under the uniform ellipticity assumption. For any downward closed set Λ and any set S, we introduce the modified quadratic error and energy

$$\overline{\sigma}(\Lambda) := \sum_{\nu\notin\Lambda} d_\nu \quad \text{and} \quad \overline{e}(S) := \sum_{\nu\in S} d_\nu. \tag{7.73}$$

Theorem 7.6. Consider a parametric problem of the type (1.5) such that UEA(r) holds for some $r>0$. Then the saturation property (7.30) holds with δ depending on r and $\overline{a}_{\max}$.

Proof. We consider an arbitrary downward closed set Λ and its margin $M := M(\Lambda)$. We first observe that

$$\overline{\sigma}(\Lambda) = \overline{e}(M) + \overline{\sigma}(\tilde{\Lambda}), \quad \tilde{\Lambda} := \Lambda \cup M. \tag{7.74}$$

Using (7.68), we may write

$$\overline{\sigma}(\tilde{\Lambda}) \le \sum_{\nu\notin\tilde{\Lambda}} d_\nu \le \alpha \sum_{\nu\notin\tilde{\Lambda}}\Big(\sum_{j\in\mathrm{supp}(\nu)} d_{\nu-e_j,j}\Big) \le A + B, \tag{7.75}$$

where

$$A := \alpha\sum_{\nu\notin\tilde{\Lambda}}\Big(\sum_{j\,\mathrm{s.t.}\,\nu-e_j\notin\tilde{\Lambda}} d_{\nu-e_j,j}\Big) = \alpha\sum_{\nu\notin\tilde{\Lambda}}\Big(\sum_{j\,\mathrm{s.t.}\,\nu+e_j\notin\tilde{\Lambda}} d_{\nu,j}\Big), \tag{7.76}$$

and

$$B := \alpha\sum_{\nu\notin\tilde{\Lambda}}\Big(\sum_{j\,\mathrm{s.t.}\,\nu-e_j\in\tilde{\Lambda}} d_{\nu-e_j,j}\Big) = \alpha\sum_{\nu\in M}\Big(\sum_{j\,\mathrm{s.t.}\,\nu+e_j\notin\tilde{\Lambda}} d_{\nu,j}\Big). \tag{7.77}$$

In this splitting, we have used the fact that if $\nu\notin\tilde{\Lambda}$ and $\nu_j\neq 0$, we have either $\nu-e_j\notin\tilde{\Lambda}$ or $\nu-e_j\in M$. Using (7.67), we control the first term A by

$$A \le \alpha\gamma\sum_{\nu\notin\tilde{\Lambda}} d_\nu = \alpha\gamma\overline{\sigma}(\tilde{\Lambda}), \tag{7.78}$$

and by the same argument we obtain

$$B \le \alpha\gamma\overline{e}(M). \tag{7.79}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Combining these estimates with (7.75), it follows that

$$(1-\alpha\gamma)\overline{\sigma}(\tilde{\Lambda}) \le \alpha\gamma\overline{e}(M), \tag{7.80}$$

and thus by (7.74)

$$\overline{\sigma}(\Lambda) \le \left(1+\frac{\alpha\gamma}{1-\alpha\gamma}\right)\overline{e}(M). \tag{7.81}$$

Finally, using the norm equivalence (7.64), we obtain the saturation property (7.30) with constant

$$\delta := \frac{\overline{a}_{\max}}{r}\left(1+\frac{\alpha\gamma}{1-\alpha\gamma}\right). \qquad \square$$

We conclude this subsection by presenting a concrete realization of the procedure SPARSE which is used in the bulk chasing Taylor algorithm with ε-accuracy, in the particular case of the elliptic problem (1.5). We again work under UEA(r). We define

$$\overline{\psi}_j := \frac{\psi_j}{\overline{a}},$$

and choose an integer $J>0$ sufficiently large that

$$\left\|\sum_{j>J}|\bar{\psi}_j|\right\|_X \le \left(\frac{\alpha\overline{e}(\Lambda)}{1-\alpha\gamma}\right)^{-1} r\varepsilon, \tag{7.82}$$

where α and γ are defined as in Lemma 7.5, and we define

$$\tilde{M} := \mathrm{SPARSE}(\Lambda,(t_\nu)_{\nu\in\Lambda},\varepsilon) := \{\nu\in M;\nu-e_j\in\Lambda\Rightarrow j\le J\}. \tag{7.83}$$

Clearly $\tilde{M}$ is finite with $\#(\tilde{M})\le J\#(\Lambda)$.

Theorem 7.7. With the above definition of $\tilde{M}$, we have

$$e(M\setminus\tilde{M}) = \sum_{\nu\in M\setminus\tilde{M}}\|t_\nu\|_V^2 \le \varepsilon. \tag{7.84}$$

Proof. We proceed in a similar way to the proof of Theorem 7.6, by first writing

$$\overline{e}(M\setminus\tilde{M}) \le \alpha\sum_{\nu\in M\setminus\tilde{M}}\left(\sum_{j\in\mathrm{supp}(\nu)} d_{\nu-e_j,j}\right) \le A+B, \tag{7.85}$$

where now

$$A := \alpha\sum_{\nu\in M\setminus\tilde{M}}\left(\sum_{j\,\mathrm{s.t.}\,\nu-e_j\in M\setminus\tilde{M}} d_{\nu-e_j,j}\right) = \alpha\sum_{\nu\in M\setminus\tilde{M}}\left(\sum_{j\,\mathrm{s.t.}\,\nu+e_j\in M\setminus\tilde{M}} d_{\nu,j}\right), \tag{7.86}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and

$$B := \alpha \sum_{\nu \in M \setminus \tilde{M}} \Bigg(\sum_{j \,\text{s.t.}\, \nu - e_j \notin M \setminus \tilde{M}} d_{\nu - e_j, j} \Bigg) = \alpha \sum_{\nu \in \Lambda \cup \tilde{M}} \Bigg(\sum_{j \,\text{s.t.}\, \nu + e_j \in M \setminus \tilde{M}} d_{\nu, j} \Bigg). \tag{7.87}$$

In this splitting, we have used the fact that if $\nu \in M \setminus \tilde{M}$ and $\nu_j \neq 0$, we have either $\nu - e_j \in M \setminus \tilde{M}$ or $\nu - e_j \in \Lambda \cup \tilde{M}$. Using (7.67), we can bound A by

$$A \leq \alpha\gamma \sum_{\nu \in M \setminus \tilde{M}} d_\nu = \alpha\gamma \overline{e}(M \setminus \tilde{M}). \tag{7.88}$$

To bound B, we first claim that for any $\nu \in \Lambda \cup \tilde{M}$ such that $\nu + e_j \in M \setminus \tilde{M}$, we must have $\nu \in \Lambda$ and $j > J$. Indeed, since $\nu + e_j \in M \setminus \tilde{M}$, the definition of $\tilde{M}$ guarantees that $\nu + e_j = \tilde{\nu} + e_k$ for some $\tilde{\nu} \in \Lambda$ and $k > J$. If $j = k$ we have our claim. If $j \neq k$ then necessarily $\tilde{\nu} - e_j \in \Lambda$ since Λ is a downward closed set, and therefore ν can be written as the sum of $\tilde{\nu} - e_j \in \Lambda$ and e_k, which means that ν is not in $\tilde{M}$. Thus, we have verified our claim. From the claim, it follows that the only j that may contribute in the summation inside B are such that $j > J$ and $\nu - e_j \in \Lambda$. Hence,

$$\begin{aligned} B &\leq \alpha \sum_{\nu \in \Lambda} \sum_{j > J} d_{\nu, j} \\ &= \alpha \sum_{\nu \in \Lambda} \int_D \Bigg(\sum_{j > J} |\psi_j| \Bigg) |\nabla t_\nu|^2 \\ &= \alpha \sum_{\nu \in \Lambda} \int_D \Bigg(\sum_{j > J} |\overline{\psi}_j| \Bigg) \overline{a} |\nabla t_\nu|^2 \\ &\leq \alpha \Bigg\| \sum_{j > J} |\overline{\psi}_j| \Bigg\|_X e(\Lambda) \leq (1 - \alpha\gamma) r\varepsilon. \end{aligned}$$

Combining the bounds for A and B with (7.85), we obtain

$$\overline{e}(M \setminus \tilde{M}) \leq \frac{B}{1 - \alpha\gamma} \leq r\varepsilon, \tag{7.89}$$

which by (7.64) implies (7.84). □

7.4. Space discretization and computational cost

In numerical computation, we need to take into account the additional space discretization of the solution map in the space $V_h \subset V$. In the case of variational problems of the form (7.4), one typical such discretization is by

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

the Petrov–Galerkin method: we define $u_h(a) \in V_h$ such that

$$B(u_h(a), v_h; a) = L(v_h), \quad v_h \in \tilde{V}_h, \tag{7.90}$$

where $\tilde{V}_h \subset \tilde{V}$ is an auxiliary finite element space such that $\dim(\tilde{V}_h) = \dim(V_h)$. For elliptic problems such as (1.5), we have $\tilde{V} = V$ and we may take $\tilde{V}_h = V_h$, which is the standard Galerkin method expressed in (5.3). We make the assumption that the discrete problem is well posed for all $a \in a(U)$, that is, Assumption AL also holds for the discrete problem.

Defining $u_h(y) = u_h(a(y))$ for a given affine representation, we thus have

$$B(u_h(y), v_h; y) = L(v_h), \quad v_h \in \tilde{V}_h, \tag{7.91}$$

and the same computation as in Lemma 7.1 shows that the Taylor coefficients $t_{\nu,h} \in V_h$ of $y \mapsto u_h(y)$ are computed by solving

$$\overline{B}(t_{\nu,h}, v_h) = L_\nu(v_h), \quad v_h \in \tilde{V}_h, \tag{7.92}$$

where $L_\nu = L$ when $\nu = 0$ is the null multi-index and

$$L_\nu(v_h) := -\sum_{j \in \mathrm{supp}(\nu)} B_j(t_{\nu - e_j,h}, v_h), \tag{7.93}$$

when $\nu \in \mathcal{F} - \{0\}$. Note that these relations amount to applying the Petrov–Galerkin approximation in the recursive computation of the Taylor coefficients t_ν.

Non-adaptive and adaptive strategies may therefore be applied in order to compute truncated Taylor expansions of the form

$$u_{n,h}(y) := \sum_{\nu \in \Lambda_n} t_{\nu,h} y^\nu, \tag{7.94}$$

with a similar convergence analysis as for the continuous problem (7.10). In particular, if the assumptions of Theorem 2.8 hold for the solution map $a \mapsto u_h(a)$, and if in addition $(\|\psi_j\|_X)_{j \geq 1} \in \ell^p(\mathbb{N})$, both non-adaptive methods based on *a priori* bounds for the $\|t_{\nu,h}\|_V$ or adaptive methods based on bulk chasing have convergence rate

$$\|u_h - u_{n,h}\|_{L^\infty(U,V)} \leq C n^{-s}, \quad s := \frac{1}{p} - 1. \tag{7.95}$$

The constant C is independent of h if, in the assumptions of Theorem 2.8, the open set $\mathcal{O}$ and the bound in (2.116) can be fixed independently of h.

Similar to the splitting (6.105) that was used for the interpolation method, we may split the resulting error into

$$\|u - u_{n,h}\|_{L^\infty(U,V)} \leq \|u_h - u_{n,h}\|_{L^\infty(U,V)} + \|u - u_h\|_{L^\infty(U,V)}. \tag{7.96}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

The second term is bounded by the error $\varepsilon(h)$ of the numerical solver. Therefore we obtain a global error bound of the form

$$\|u - u_{n,h}\|_{L^\infty(U,V)} \leq Cn^{-s} + \varepsilon(h), \tag{7.97}$$

similar to the bound (6.106) obtained for the interpolation method after space discretization.

We next turn to the estimate of the computational cost, starting with the offline cost. The computation of each individual $t_{\nu,h}$, stored as vectors of dimension N_h of their coordinates in the nodal finite element basis of V_h, requires us to solve a system. The cost of solving this system is of order C_h, where C_h is the individual cost of one application, often the discrete solver. Indeed, it amounts to solving a discrete problem, for which we invert exactly the same stiffness matrix as for the computation of the particular instance $u_h(0)$. Assembling this system requires us to compute the right-hand side, which necessitates $\|\nu\|_0$ applications of the stiffness matrices associated to the sesquilinear form B_j. Since B_j is associated to a partial differential operator, its stiffness matrix in the nodal basis is sparse, and therefore each such application has cost smaller than order N_h. We have already observed in Section 5.3 that, since Λ_n is a downward closed set, we have $2^{\|\nu\|_0} \leq n$ for each $\nu \in \Lambda_n$. The cost of computing an individual $t_{\nu,h}$ is thus at most of the order

$$C_{\text{off}}(\nu) \sim C_h + \log(n)\, N_h. \tag{7.98}$$

In the non-adaptive algorithm, we compute the n values of $t_{\nu,h}$ for $\nu \in \Lambda_n$, and therefore the total offline cost is at most of order

$$C_{\text{off}} \sim nC_h + n\log(n)\, N_h. \tag{7.99}$$

In adaptive algorithms, we need to take into account the additional computation of the $t_{\nu,h}$ for ν in the margin of Λ_n. For the bulk chasing Taylor algorithm with ε accuracy, the individual cost $C_h + \log(n)\, N_h$ is thus multiplied by $n + \#(\tilde{M}_n)$, where

$$\#(\tilde{M}_n) := \text{SPARSE}(\Lambda_n, (t_{\nu,h})_{\nu\in\Lambda_n}, \varepsilon). \tag{7.100}$$

For the SPARSE procedure that we have proposed in the case of the elliptic problem (1.5), we have $\#(\tilde{M}_n) \leq Jn$, where $J = J(\varepsilon)$ is such that

$$\left\| \sum_{j\geq J} |\psi_j| \right\|_X \lesssim \varepsilon. \tag{7.101}$$

Having assumed that $(\|\psi_j\|_X)_{j\in\mathbb{N}}$ is ℓ^p-summable, and reordering such that this sequence is non-increasing, we find by Lemma 3.6 that $J(\varepsilon) \lesssim \varepsilon^{-1/s}$, where $s := \frac{1}{p} - 1$. It follows that the total offline cost for this algorithm is

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

at most of order

$$C_{\rm off} \sim \varepsilon^{-1/s} n C_h + \varepsilon^{1/s} n \log(n)\, N_h. \tag{7.102}$$

As to the online cost, since the online stage simply consists of the combination of the $t_{\nu,h}$ for computing $u_{n,h}$, we find that this cost is of the order

$$C_{\rm on} \sim n N_h, \tag{7.103}$$

similar to sparse polynomial interpolation algorithms.

If ε is a desired order of accuracy, and if we have the error bound (6.106), then one way to reach this accuracy is to take both Cn^{-s} and $\varepsilon(h)$ of the order of ε. With $h(\varepsilon)$ the inverse function of $\varepsilon(h)$, as in (6.112), we thus find that the non-adaptive Taylor algorithm reaches the order of accuracy ε at cost at most of order

$$C_{\rm off}(\varepsilon) \sim \varepsilon^{-1/s} C_{h(\varepsilon)} + \varepsilon^{-1/s} |\log(\varepsilon)| N_{h(\varepsilon)} \quad \text{and} \quad C_{\rm on}(\varepsilon) \sim \varepsilon^{-1/s} N_{h(\varepsilon)}. \tag{7.104}$$

For the bulk chasing Taylor algorithm with ε-accuracy, we have the more pessimistic estimate

$$C_{\rm off}(\varepsilon) \sim \varepsilon^{-2/s} C_{h(\varepsilon)} + \varepsilon^{-2/s} |\log(\varepsilon)| N_{h(\varepsilon)} \quad \text{and} \quad C_{\rm on}(\varepsilon) \sim \varepsilon^{-1/s} N_{h(\varepsilon)}, \tag{7.105}$$

due to the inflation by $J(\varepsilon)$.

Similar to the interpolation algorithm discussed in Section 6, both algorithms are immune to the curse of dimensionality since these trade-offs between accuracy and complexity are obtained with infinitely many variables.

8. Reduced basis methods

We turn next to the class of numerical techniques for solving parametric PDEs known as reduced basis methods. These methods aim at finding a good subspace $V_n \subset V$, of small dimension n, to be used for approximating the elements of the solution manifold $\mathcal{M}$. We know that, for any fixed value of n, the best choice of V_n is one that achieves the infimum in the definition (1.41) of the Kolmogorov n-width with $\mathcal{K} = \mathcal{M}$. However, such a space, if it exists, is generally out of reach from a computational point of view. The reduced basis method uses a space V_n, which may be sub-optimal, spanned by n *snapshots* $u(a^1), \ldots, u(a^n)$ from the solution manifold $\mathcal{M}$. While these snapshots can be chosen in various ways, a particularly interesting strategy proceeds with a recursive greedy selection. We present this strategy in Section 8.1. In Section 8.2 we prove, for the case where V is a Hilbert space, that in a certain sense the resulting spaces V_n perform almost as well as the optimal n-width spaces in terms of convergence rates. A similar

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

analysis is given in Section 8.3 for the case of a general Banach space. The effect of space discretization on the convergence of the algorithm is discussed in Section 8.4, and the computational cost is analysed in Section 8.5.

8.1. Greedy selection algorithms

We have seen that, for relevant classes of parametric PDEs, the solution manifold $\mathcal{M}$ is a compact set in the Banach space V. While in most applications V is a Hilbert space, we describe the greedy algorithm for any compact set $\mathcal{K}$ in any Banach space V. We then analyse its performance, first in the case where V is a Hilbert space, and then later in the case of a general Banach space. We describe two versions of a greedy algorithm for generating approximation spaces for $\mathcal{K}$. The first version, called the *pure greedy algorithm*, is rather ideal, while the second version, called the *weak greedy algorithm*, is more amenable to numerical implementation.

Pure greedy algorithm. We first choose a function $g_0 \in \mathcal{K}$ such that

$$\|g_0\|_V = \max_{g\in\mathcal{K}} \|g\|_V. \tag{8.1}$$

Since $\mathcal{K}$ is compact, such a g_0 always exists but of course may not be unique. Assuming $\{g_0, \ldots, g_{n-1}\}$ have been selected, we set

$$V_n := \operatorname{span}\{g_0, \ldots, g_{n-1}\},$$

and we then take $g_n \in \mathcal{K}$ such that

$$\operatorname{dist}(g_n, V_n)_V = \max_{g\in\mathcal{K}} \operatorname{dist}(g, V_n)_V, \tag{8.2}$$

where

$$\operatorname{dist}(g, V_n)_V := \min_{h\in V_n} \|g - h\|_V. \tag{8.3}$$

We define $\sigma_0 := \sigma_0(\mathcal{K})_V = \max_{g\in\mathcal{K}} \|g\|_V$ and

$$\sigma_n := \sigma_n(\mathcal{K})_V := \sup_{g\in\mathcal{K}} \inf_{v\in V_n} \|g - v\|_V, \quad n \geq 1, \tag{8.4}$$

so that

$$\sigma_n := \operatorname{dist}(\mathcal{K}, V_n)_V = \operatorname{dist}(g_n, V_n)_V. \tag{8.5}$$

This greedy algorithm was introduced by Veroy *et al.* (2003) for the case when V is a Hilbert space, and subsequently studied extensively by Buffa *et al.* (2012) and Maday, Patera and Turinici (2002*a*, 2002*b*).

In the setting of parametric PDEs, it not possible to compute for a given $g \in \mathcal{K}$ the distance $\operatorname{dist}(g, V_n)_V$, so that one cannot exactly perform the maximization in (8.2). However, it is possible to introduce a computable

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

error indicator $d(g, V_n)_V$ which satisfies

$$cd(g, V_n)_V \le \operatorname{dist}(g, V_n)_V \le Cd(g, V_n)_V, \quad g \in \mathcal{K}, \tag{8.6}$$

for fixed constants $c, C > 0$. Performing the maximization (8.2) on $d(g, V_n)_V$ is equivalent to the application of the following weaker form of the greedy algorithm with $\gamma := c/C$, which is more appropriate.

Weak greedy algorithm. We fix a constant $0 < \gamma \le 1$. At the first step of the algorithm, one chooses a function $g_0 \in \mathcal{K}$ such that

$$\|g_0\|_V \ge \gamma \max_{g \in \mathcal{K}} \|g\|_V. \tag{8.7}$$

At the general step, if $g_0, \dots, g_{n-1}$ have been chosen, we set

$$V_n := \operatorname{span}\{g_0, \dots, g_{n-1}\},$$

and we now choose $g_n \in \mathcal{K}$ satisfying

$$\operatorname{dist}(g_n, V_n)_V \ge \gamma \max_{g \in \mathcal{K}} \operatorname{dist}(g, V_n)_V \tag{8.8}$$

to be the next element in the greedy selection. As in the pure greedy case, we introduce

$$\sigma_n := \sigma_n(\mathcal{K})_V := \operatorname{dist}(\mathcal{K}, V_n)_V, \quad n \ge 0, \tag{8.9}$$

which now measures the performance of the weak greedy algorithm. Note that if $\gamma = 1$, then the weak greedy algorithm reduces to the pure greedy algorithm introduced above. With the same definition as above for $\sigma_n := \sigma_n(\mathcal{K})_V$, we thus have

$$\operatorname{dist}(g_n, V_n)_V \ge \gamma \sigma_n. \tag{8.10}$$

For both of these algorithms, the sequence $(\sigma_n)_{n \ge 0}$ is monotone non-increasing. It is also important to note that neither the pure greedy algorithm nor the weak greedy algorithm give a unique sequence $(g_n)_{n \ge 0}$, nor is the sequence $(\sigma_n)_{n \ge 0}$ unique. In all that follows, the notation reflects any sequences that can arise in the implementation of the weak greedy selection for the fixed value of γ.

8.2. Convergence analysis of greedy algorithms in a Hilbert space

We are interested in how well the space V_n, generated by the weak greedy algorithm, approximates the elements of $\mathcal{K}$. For this purpose we would like to compare its performance measured by σ_n with the best possible performance given by the Kolmogorov width

$$d_n := d_n(\mathcal{K})_V. \tag{8.11}$$

If $(\sigma_n)_{n \ge 0}$ were bounded by $(d_n)_{n \ge 0}$ up to a fixed multiplicative constant,

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

this would mean that the greedy selection provides essentially the best possible accuracy attainable by n-dimensional subspaces. However, such a general comparison is not to be expected.

Various comparisons between σ_n and d_n have been established in the literature. A first result in this direction, in the case of the pure greedy algorithm applied to a Hilbert space V, was given by Buffa *et al.* (2012), who proved that

$$\sigma_n(\mathcal{K})_V \leq Cn2^n d_n(\mathcal{K})_V, \quad n \geq 1, \tag{8.12}$$

with C an absolute constant. The same result holds with C depending on γ for the weak greedy algorithm. While this is an interesting comparison, it is only useful if $d_n(\mathcal{K})_V$ decays to zero faster than $n^{-1}2^{-n}$, which may be a severe assumption. Unfortunately, the above result is sharp in the following sense: Binev *et al.* (2011) proved that for all $n \geq 1$ and $\varepsilon > 0$ there exists a compact set $\mathcal{K}$ such that

$$\sigma_n(\mathcal{K})_V \geq (1-\varepsilon)2^n d_n(\mathcal{K})_V. \tag{8.13}$$

This reveals that direct comparison between $\sigma_n(\mathcal{K})_V$ and $d_n(\mathcal{K})_V$ is doomed to fail.

Significant improvements on (8.12) were given by Binev *et al.* (2011), again in the Hilbert space setting, by changing the way of comparing $\sigma_n(\mathcal{K})_V$ and $d_n(\mathcal{K})_V$. Perhaps the most interesting comparison is as follows. If, for some constant $C > 0$ and some $s > 0$, the compact set $\mathcal{K}$ satisfies $d_n(\mathcal{K})_V \leq C(\max\{1,n\})^{-s}$ for all $n \geq 0$, then there is a constant $\tilde{C}$ which depends only on C and s such that

$$\sigma_n(\mathcal{K})_V \leq \tilde{C}(\max\{1,n\})^{-s}, \quad n \geq 0. \tag{8.14}$$

In other words, for the scale of polynomial decay, the greedy algorithm performs with the same decay rates as n-widths. These results were improved upon in DeVore *et al.* (2013) and extended to the case of a general Banach space V.

The analysis of the two greedy algorithms above is quite simple and executed with elementary results from linear algebra. We consider the case when V is a Hilbert space and show that the action of the weak greedy algorithm is captured by a certain lower triangular matrix. Note that in general the weak greedy algorithm does not terminate, and we obtain an infinite sequence $(g_n)_{n\geq 0}$. In order to have a consistent notation in what follows, we define $g_n := 0$, $n > m$, if the algorithm terminates at $n = m$, that is, if $\sigma_m(\mathcal{K})_V = 0$.

We let $(g_n^*)_{n\geq 0}$ denote the orthonormal system obtained from $(g_n)_{n\geq 0}$ by Gram–Schmidt orthogonalization executed in the natural order. It follows

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

that the orthogonal projector P_n from V onto V_n is given by

$$P_n g = \sum_{i=0}^{n-1} \langle g, g_i^* \rangle g_i^*,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of V, and, in particular,

$$g_n = P_{n+1} g_n = \sum_{j=0}^{n} a_{n,j} g_j^*, \quad a_{n,j} = \langle g_n, g_j^* \rangle, \quad 0 \le j \le n. \tag{8.15}$$

We consider the infinite lower triangular matrix

$$A := (a_{i,j})_{i,j=0}^{\infty}, \quad a_{i,j} := 0, \ j > i.$$

This matrix incorporates all the information about the weak greedy algorithm on $\mathcal{K}$. For example, the nth row of A gives the nth element g_n in the greedy selection. The following two properties characterize any lower triangular matrix A generated by the weak greedy algorithm with constant γ. With the notation $\sigma_n := \sigma_n(\mathcal{K})_V$, we have the following.

(P1) The diagonal elements of A satisfy $\gamma \sigma_n \le |a_{n,n}| \le \sigma_n$.

(P2) For every $m \ge n$, we have $\sum_{j=n}^{m} a_{m,j}^2 \le \sigma_n^2$.

Indeed, (P1) follows from

$$a_{n,n}^2 = \|g_n - P_n g_n\|_V^2 = \operatorname{dist}(g_n, V_n)_V^2.$$

This shows the upper bound in (P1) because each element of $\mathcal{K}$ is approximated to error σ_n. It also shows the lower bound because of the weak greedy selection property (8.8). To see (P2), we note that for $m \ge n$,

$$\sum_{j=n}^{m} a_{m,j}^2 = \|g_m - P_n g_m\|_V^2 \le \max_{g \in \mathcal{K}} \|g - P_n g\|_V^2 = \sigma_n^2.$$

Remark 8.1. If A is any infinite matrix satisfying (P1) and (P2) with $(\sigma_n)_{n \ge 0}$ a non-increasing sequence that converges to 0, then the rows of A form a compact subset of $\ell^2(\mathbb{N}_0)$ where $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. If $\mathcal{K}$ is the set consisting of these rows, then one of the possible realizations of the weak greedy algorithm on this set $\mathcal{K}$ with constant γ will choose the rows in that order and A will be the resulting matrix. In this sense, the action of the greedy algorithm on the original set $\mathcal{K}$ is completely described by the matrix A.

It follows from the above remark that there is no loss of generality in assuming that the infinite-dimensional Hilbert space V is $\ell^2(\mathbb{N}_0)$ and that $g_j^* = e_j$, where e_j is the vector with a one in the coordinate indexed by j and zero in all other coordinates, that is, $(e_j)_i = \delta_{j,i}$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

With this matrix description of the weak greedy algorithm in hand, estimates for the convergence rate of the algorithm rely on an analysis of A. Note that the diagonal elements of A give the errors σ_n, and hence we want a general way to bound the diagonal elements of matrices A with the above properties. The following lemma from DeVore *et al.* (2013) gives a general way to bound diagonal elements of a general lower triangular matrix G. It is applied later to the sections of A to obtain convergence results for the weak greedy algorithm.

Lemma 8.2. Let $G = (g_{i,j})$ be a $K \times K$ lower triangular matrix with rows $\mathbf{g}_1, \dots, \mathbf{g}_K$. If W is any m-dimensional subspace of $\mathbb{R}^K$ for some $0 < m \leq K$, and P is the orthogonal projection from $\mathbb{R}^K$ onto W, then

$$\det(G)^2 = \prod_{i=1}^{K} g_{i,i}^2 \leq \left(\frac{1}{m} \sum_{i=1}^{K} \|P\mathbf{g}_i\|_{\ell^2}^2 \right)^m \left(\frac{1}{K-m} \sum_{i=1}^{K} \|\mathbf{g}_i - P\mathbf{g}_i\|_{\ell^2}^2 \right)^{K-m}, \tag{8.16}$$

where $\|\cdot\|_{\ell^2}$ is the Euclidean norm of a vector in $\mathbb{R}^K$.

Proof. Let $\varphi_1, \dots, \varphi_m$ be any orthonormal basis for the space W and extend it to an orthonormal basis $\varphi_1, \dots, \varphi_K$ for $\mathbb{R}^K$. If we let Φ denote the $K \times K$ orthogonal matrix whose jth column is φ_j, then the matrix $C := G\Phi$ has entries $c_{i,j} = \langle \mathbf{g}_i, \varphi_j \rangle$. We let $\mathbf{c}_j$ denote the jth column of C. It follows from the arithmetic–geometric mean inequality for the numbers $\{\|\mathbf{c}_j\|_{\ell^2}^2\}_{j=1}^m$ that

$$\begin{aligned} \prod_{j=1}^{m} \|\mathbf{c}_j\|_{\ell^2}^2 &\leq \left(\frac{1}{m} \sum_{j=1}^{m} \|\mathbf{c}_j\|_{\ell^2}^2 \right)^m \\ &= \left(\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{K} \langle \mathbf{g}_i, \varphi_j \rangle^2 \right)^m \\ &= \left(\frac{1}{m} \sum_{i=1}^{K} \|P\mathbf{g}_i\|_{\ell^2}^2 \right)^m. \end{aligned} \tag{8.17}$$

Similarly,

$$\begin{aligned} \prod_{j=m+1}^{K} \|\mathbf{c}_j\|_{\ell^2}^2 &\leq \left(\frac{1}{K-m} \sum_{j=m+1}^{K} \|\mathbf{c}_j\|_{\ell^2}^2 \right)^{K-m} \\ &= \left(\frac{1}{K-m} \sum_{i=1}^{K} \|\mathbf{g}_i - P\mathbf{g}_i\|_{\ell^2}^2 \right)^{K-m}, \end{aligned} \tag{8.18}$$

where we have used the fact that φ_j is orthogonal to W when $j > m$. Now

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

we invoke Hadamard's inequality for the matrix C, which says that

$$(\det C)^2 \le \prod_{j=1}^{K} \|\mathbf{c}_j\|_{\ell^2}^2, \tag{8.19}$$

and combine it with relations (8.17) and (8.18) to obtain

$$(\det C)^2 \ \le \left(\frac{1}{m}\sum_{i=1}^{K} \|P\mathbf{g}_i\|_{\ell^2}^2\right)^m \left(\frac{1}{K-m}\sum_{i=1}^{K} \|\mathbf{g}_i - P\mathbf{g}_i\|_{\ell^2}^2\right)^{K-m}. \tag{8.20}$$

The latter inequality and the fact that $|\det C| = |\det G|$ gives (8.16). □

Let us now see how this lemma is utilized to derive convergence results for the greedy algorithm. The following theorem, taken from DeVore *et al.* (2013), relates the errors $\sigma_n(\mathcal{K})_V$ to the n-widths $d_n(\mathcal{K})_V$.

Theorem 8.3. For the weak greedy algorithm with constant γ in a Hilbert space V and for any compact set $\mathcal{K}$, the following inequalities between $\sigma_n := \sigma_n(\mathcal{K})_V$ and $d_n := d_n(\mathcal{K})_V$ hold for any $N \ge 0$, $K \ge 1$, and $1 \le m < K$:

$$\prod_{i=1}^{K} \sigma_{N+i}^2 \le \gamma^{-2K} \left(\frac{K}{m}\right)^m \left(\frac{K}{K-m}\right)^{K-m} \sigma_{N+1}^{2m} d_m^{2K-2m}. \tag{8.21}$$

Proof. In what follows, we assume that there exists a space W_m that achieves the infimum in the definition of the m-width of $\mathcal{K}$, that is, such that

$$\max_{g\in\mathcal{K}} \min_{w\in W_m} \|g - w\|_V = d_m. \tag{8.22}$$

If such a space does not exist, we may, for each $\varepsilon > 0$, find one such that

$$\max_{g\in\mathcal{K}} \min_{w\in W_m} \|g - w\|_V \le d_m + \varepsilon,$$

and modify the proof below by a limiting argument so as to reach the same conclusion.

We consider the $K \times K$ matrix $G = (g_{i,j})$, which is formed by the rows and columns of A with indices from $\{N+1, \dots, N+K\}$. Each row $\mathbf{g}_i$ is the restriction of row $N+i$ of A to the coordinates $N+1, \dots, N+K$. The space W_m determines a sequence space $\bar{W}_m \subset \ell^2$ such that $\operatorname{dist}(\mathbf{g}_i, \bar{W}_m)_{\ell^2} \le d_m$ for $i = 1, \dots, K$. Let $\tilde{W}$ be the linear space which is the restriction of W_m to the coordinates $N+1, \dots, N+K$. Obviously we have $\dim(\tilde{W}) \le m$. Let W be an m-dimensional space, $W \subset \operatorname{span}\{e_{N+1}, \dots, e_{N+K}\}$, such that $\tilde{W} \subset W$, and let P and $\tilde{P}$ be the projections in $\mathbb{R}^K$ onto W and $\tilde{W}$, respectively. Clearly,

$$\|P\mathbf{g}_i\|_{\ell^2} \le \|\mathbf{g}_i\|_{\ell^2} \le \sigma_{N+1}, \quad i = 1, \dots, K, \tag{8.23}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

where we have used property (P2) in the last inequality. Note that

$$\|\mathbf{g}_i - P\mathbf{g}_i\|_{\ell^2} \le \|\mathbf{g}_i - \tilde{P}\mathbf{g}_i\|_{\ell^2} = \operatorname{dist}(\mathbf{g}_i, \tilde{W})_{\ell^2} \le \operatorname{dist}(\mathbf{g}_i, W)_{\ell^2} \le d_m \quad (8.24)$$

for $i = 1, \dots, K$. It follows from property (P1) that

$$\gamma^K \prod_{i=1}^{K} \sigma_{N+i} \le \prod_{i=1}^{K} |a_{N+i,N+i}|. \quad (8.25)$$

We now apply Lemma 8.2 to this G and W, and use estimates (8.23), (8.24), and (8.25) to derive (8.21). □

Using this theorem, we now establish convergence results for the weak greedy algorithm, showing in particular that if $d_n(\mathcal{K})_V$ decays with an algebraic or exponential convergence rate, then a similar rate holds for $\sigma_n(\mathcal{K})_V$.

Corollary 8.4. For the weak greedy algorithm with constant γ in a Hilbert space V, we have the following.

(i) For any compact set $\mathcal{K}$, we have

$$\sigma_n(\mathcal{K})_V \le \sqrt{2}\gamma^{-1} d_0(\mathcal{K})_V^{m/n} \min_{1\le m<n} d_m^{(n-m)/n}(\mathcal{K})_V, \quad n \ge 1. \quad (8.26)$$

In particular $\sigma_{2n}(\mathcal{K})_V \le \gamma^{-1}\sqrt{2 d_0(\mathcal{K})_V d_n(\mathcal{K})_V}$ for all $n \ge 1$.

(ii) For any $s > 0$ and $C_0 > 0$,

$$\begin{aligned} &d_n(\mathcal{K})_V \le C_0(\max\{1,n\})^{-s}, \quad n \ge 0 \quad \Rightarrow \\ &\sigma_n(\mathcal{K})_V \le C_1(\max\{1,n\})^{-s}, \quad n \ge 0, \end{aligned} \quad (8.27)$$

where $C_1 := \gamma^{-2}2^{4s+1}C_0$.

(iii) For any $s > 0$ and $c_0, C_0 > 0$,

$$\begin{aligned} &d_n(\mathcal{K})_V \le C_0\, \mathrm{e}^{-c_0 n^s}, \quad n \ge 0 \quad \Rightarrow \\ &\sigma_n(\mathcal{K})_V \le \tilde{C}_1\, \mathrm{e}^{-c_1 n^s}, \quad n \ge 0, \end{aligned} \quad (8.28)$$

where $c_1 = (c_0/2)3^{-s}$ and $\tilde{C}_1 := C_0 \max\{\sqrt{2}\gamma^{-1}, \mathrm{e}^{c_1}\}$.

Proof. **(i)** We take $N = 0$, $K = n$ and any $1 \le m < n$ in Theorem 8.3. Using the monotonicity of $(\sigma_n)_{n\ge0}$ and the fact that $\sigma_1 \le \sigma_0 \le d_0$, we obtain

$$\sigma_n^{2n} \le \prod_{j=1}^{n} \sigma_j^2 \le \gamma^{-2n} \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m} d_m^{2n-2m} d_0^{2m}. \quad (8.29)$$

Since $x^{-x}(1-x)^{x-1} \le 2$ for $0 < x < 1$, we derive (8.26).

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

(ii) It follows from the monotonicity of $(\sigma_n)_{n\geq 0}$ and (8.21) for $N = K = n$ and any $1 \leq m < n$ that

$$\sigma_{2n}^{2n} \leq \prod_{j=n+1}^{2n} \sigma_j^2 \leq \gamma^{-2n} \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m} \sigma_n^{2m} d_m^{2n-2m}.$$

In the case $n = 2k$ and $m = k$, we have for any positive integer k that

$$\sigma_{4k} \leq \sqrt{2}\gamma^{-1}\sqrt{\sigma_{2k} d_k}. \tag{8.30}$$

Assuming that $d_n(\mathcal{K})_V \leq C_0(\max\{1,n\})^{-s}$ for all $n \geq 0$, we obtain by induction that for all $j \geq 0$ and $n = 2^j$,

$$\sigma_n = \sigma_{2^j} \leq C2^{-sj} \leq n^{-s}, \quad C := 2^{3s+1}\gamma^{-2}C_0. \tag{8.31}$$

Indeed, the above obviously holds for $j = 0$ or 1, since for these values we have

$$\sigma_{2^j} \leq \sigma_0 = d_0 \leq C_0 \leq C2^{-sj}. \tag{8.32}$$

Assuming its validity for some $j \geq 1$, we find that

$$\begin{aligned}\sigma_{2^{j+1}} &\leq \sqrt{2}\gamma^{-1}\sqrt{\sigma_{2^j} d_{2^{j-1}}} \\ &\leq \gamma^{-1}2^{(3s)/2}\sqrt{2CC_0}2^{-s(j+1)} \\ &= \sqrt{C}\sqrt{2^{3s+1}C_0\gamma^{-2}}2^{-s(j+1)} = C2^{-s(j+1)},\end{aligned}$$

where we have used the definition of C. For values $2^j < n < 2^{j+1}$, we obtain the general result by writing

$$\sigma_n \leq \sigma_{2^j} \leq C2^{-sj} \leq 2^s C n^{-s} = C_1 n^{-s}. \tag{8.33}$$

In the case $n = 0$, we simply write $\sigma_0 = d_0 \leq C_0 \leq C_1$.

(iii) Assuming that $d_n(\mathcal{K})_V \leq C_0\,\mathrm{e}^{-c_0 n^s}$ for all $n \geq 0$, we obtain from (i) for all $n \geq 1$,

$$\begin{aligned}\sigma_{2n+1} \leq \sigma_{2n} &\leq \sqrt{2}\gamma^{-1}\sqrt{d_n d_0} \leq \sqrt{2C_0 d_0}\gamma^{-1}\,\mathrm{e}^{-(c_0/2)n^s} \\ &\leq \sqrt{2}C_0\gamma^{-1}\,\mathrm{e}^{-(c_0/2)3^{-s}(2n+1)^s}.\end{aligned} \tag{8.34}$$

This proves

$$\sigma_n \leq \tilde{C}_1\,\mathrm{e}^{-c_1 n^s}, \quad n \geq 2. \tag{8.35}$$

For the values $n = 0$ or $n = 1$, we simply write

$$\sigma_n \leq \sigma_0 = d_0 \leq C_0 \leq \tilde{C}_1\,\mathrm{e}^{-c_1} \leq \tilde{C}_1\,\mathrm{e}^{-c_1 n^s}, \tag{8.36}$$

which concludes the proof of (iii). □

Remark 8.5. Inspection of the above proof shows that in items (ii) and (iii) of Corollary 8.4, if the same decay rates of $d_n(\mathcal{K})_V$ are only assumed

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

within a limited range $0 \le n \le N$, then the same decay rates of $\sigma_n(\mathcal{K})_V$ are achieved for the same rate $0 \le n \le N$ up to some changes in the expressions of the constants C_1, c_1 and $\tilde{C}_1$.

8.3. Convergence analysis of greedy algorithms in a Banach space

We now turn our attention to the performance of the weak greedy algorithm for a compact set $\mathcal{K}$ in a general Banach space V. We use the abbreviations $\sigma_n := \sigma_n(\mathcal{K})_V$ and $d_n := d_n(\mathcal{K})_V$. While the development is quite similar to the case of a Hilbert space, there is a slight loss in the comparison between σ_n and d_n due to the lack of Hilbert space orthogonality.

As in the Hilbert space case, we associate with the greedy procedure a lower triangular matrix $A = (a_{i,j})_{i,j=0}^{\infty}$ in the following way. For each $j = 0, 1, \dots$, we let $\lambda_j \in V^*$ be the linear functional of norm $\|\lambda_j\|_{V^*} = 1$ that satisfies

$$\text{(i) } \lambda_j(g) = 0, \quad g \in V_j, \quad \text{and} \quad \text{(ii) } \lambda_j(g_j) = \operatorname{dist}(g_j, V_j)_V. \tag{8.37}$$

The existence of such a functional is a simple consequence of the Hahn–Banach theorem. We now let A be the matrix with entries

$$a_{i,j} = \lambda_j(g_i).$$

From (ii) of (8.37), we see that A is lower triangular. Its diagonal elements $a_{j,j}$ satisfy the inequality

$$\gamma\sigma_j \le a_{j,j} = \operatorname{dist}(g_j, V_j)_V = \sigma_j, \tag{8.38}$$

because of the weak greedy selection property (8.8). Also, each entry $a_{i,j}$ satisfies

$$|a_{i,j}| = |\lambda_j(g_i)| = |\lambda_j(g_i - g)| \le \|\lambda_j\|_{V^*}\|g_i - g\|_V = \|g_i - g\|_V, \quad j < i,$$

for every $g \in V_j$, since $\lambda_j(V_j) = 0$. Therefore, we have

$$|a_{i,j}| \le \operatorname{dist}(g_i, V_j)_V \le \sigma_j, \quad j < i. \tag{8.39}$$

Theorem 8.6. For the weak greedy algorithm with constant γ in a Banach space V and for any compact set $\mathcal{K}$ contained in V, we have the following inequalities between $\sigma_n := \sigma_n(\mathcal{K})_V$ and $d_n := d_n(\mathcal{K})_V$: for any $N \ge 0$, $K \ge 1$, and $1 \le m < K$,

$$\prod_{i=1}^{K} \sigma_{N+i}^2 \le 2^K K^{K-m} \gamma^{-2K} \left(\sum_{i=1}^{K} \sigma_{N+i}^2 \right)^m d_m^{2K-2m}. \tag{8.40}$$

Proof. As in the proof of Theorem 8.3, we consider the $K \times K$ matrix G which is formed by the rows and columns of A with indices from $\{N + 1, \dots, N + K\}$. Let V_m be a Kolmogorov subspace of V for which

$$\operatorname{dist}(\mathcal{K}, V_m)_V = d_m.$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Again, we assume that such a space V_m exists. Otherwise we modify the proof given below by adding an arbitrary $\varepsilon > 0$ to d_m and then letting ε tend to zero at the end.

For each i, there is an element $h_i \in V_m$ such that

$$\|g_i - h_i\|_V = \operatorname{dist}(g_i, V_m)_V \le d_m,$$

and therefore

$$|\lambda_j(g_i) - \lambda_j(h_i)| = |\lambda_j(g_i - h_i)| \le \|\lambda_j\|_{V^*}\|g_i - h_i\|_V \le d_m. \tag{8.41}$$

We now consider the collection of vectors $(\lambda_{N+1}(h), \dots, \lambda_{N+K}(h))$ for all $h \in V_m$. They span a space $W_m \subset \mathbb{R}^K$ of dimension at most m. We assume that $\dim(W_m) = m$ (a slight notational adjustment has to be made if $\dim(W_m) < m$ without affecting the final result). It follows from (8.41) that each row $\mathbf{g}_i$ of G can be approximated by a vector from W_m in the ℓ^∞-norm to accuracy d_m, and therefore in the ℓ^2-norm to accuracy $\sqrt{K}d_m$. Let P be the orthogonal projection of $\mathbb{R}^K$ onto W. Hence, we have

$$\|\mathbf{g}_i - P\mathbf{g}_i\|_{\ell_2} \le \sqrt{K}d_m, \quad i = 1, \dots, K. \tag{8.42}$$

It also follows from (8.39) that

$$\|P\mathbf{g}_i\|_{\ell_2} \le \|\mathbf{g}_i\|_{\ell_2} \le \left(\sum_{j=1}^{i} \sigma_{N+j}^2\right)^{1/2}, \tag{8.43}$$

and therefore

$$\sum_{i=1}^{K} \|P\mathbf{g}_i\|_{\ell_2}^2 \le \sum_{i=1}^{K}\sum_{j=1}^{i} \sigma_{N+j}^2 \le K \sum_{i=1}^{K} \sigma_{N+i}^2. \tag{8.44}$$

Next, we apply Lemma 8.2 for this G and W and use estimates (8.38), (8.42) and (8.44) to derive

$$\begin{aligned}
\gamma^{2K} \prod_{i=1}^{K} \sigma_{N+i}^2 &\le \left(\frac{K}{m}\sum_{i=1}^{K} \sigma_{N+i}^2\right)^m \left(\frac{K^2}{K-m} d_m^2\right)^{K-m} \\
&= K^{K-m}\left(\frac{K}{m}\right)^m \left(\frac{K}{K-m}\right)^{K-m} \left(\sum_{i=1}^{K} \sigma_{N+i}^2\right)^m d_m^{2(K-m)} \\
&\le 2^K K^{K-m} \left(\sum_{i=1}^{K} \sigma_{N+i}^2\right)^m d_m^{2(K-m)},
\end{aligned}$$

and the proof is complete. □

In analogy with Corollary 8.4, we can use the above result to establish a convergence theorem for the weak greedy algorithm in a general Banach space. Since the proof is very similar to that of Corollary 8.4, except that

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

we use (8.40) in place of (8.21), we only state the result and refer to DeVore *et al.* (2013) for more details.

Corollary 8.7. Suppose that V is a Banach space. For the weak greedy algorithm with a constant γ, applied to a compact set $\mathcal{K} \subset V$, we have the following.

(i) For any $n \geq 1$, we have

$$\sigma_n(\mathcal{K})_V \leq \sqrt{2}\gamma^{-1} \min_{1\leq m<n} n^{(n-m)/(2n)} \left(\sum_{i=1}^{n} \sigma_i(\mathcal{K})_V^2 \right)^{m/(2n)} d_m(\mathcal{K})_V^{(n-m)/n}. \tag{8.45}$$

In particular $\sigma_{2n}(\mathcal{K})_V \leq 2\gamma^{-1}\sqrt{n d_0(\mathcal{K})_V d_n(\mathcal{K})_V}$ for all $n \geq 1$.

(ii) For any $s > 0$, $C_0 > 0$ and $\varepsilon > 0$, we have

$$\begin{aligned} &d_n(\mathcal{K})_V \leq C_0(\max\{1,n\})^{-s}, && n \geq 0 \quad \Rightarrow \\ &\sigma_n(\mathcal{K})_V \leq C_1(\max\{1,n\})^{-(s-\varepsilon-1/2)}, && n \geq 0, \end{aligned} \tag{8.46}$$

where C_1 depends on C_0, s, γ and ε.

(iii) For any $s > 0$ and $c_0, C_0 > 0$, we have

$$\begin{aligned} &d_n(\mathcal{K})_V \leq C_0\, \mathrm{e}^{-c_0 n^s}, && n \geq 0 \quad \Rightarrow \\ &\sigma_n(\mathcal{K})_V \leq \tilde{C}_1\, \mathrm{e}^{-c_1 n^s}, && n \geq 0, \end{aligned} \tag{8.47}$$

where c_1 depends on s and c_0, and $\tilde{C}_1$ depends on C_0, γ, s, and c_0.

Statement (ii) in the above corollary shows that there is a loss of $1/2$ in the algebraic rate of decay of σ_n compared to that of d_n. It is natural to ask whether this loss is unavoidable when proving results in a Banach space. We next provide an example which shows that a loss of this type is unavoidable in general. However, there is still a small gap between the above corollary and what the example below provides.

Let us begin by considering the space $V = \ell^\infty(\mathbb{N} \cup \{0\})$ equipped with its usual norm. We consider a monotone non-increasing sequence $x_0 \geq x_1 \geq x_2 \geq \cdots$ of positive real numbers which converges to zero, and we define

$$f_j := x_j e_j, \quad j = 0, 1, \ldots, \tag{8.48}$$

where e_j is the Kronecker sequence with 1 at position j. We consider the compact set

$$\mathcal{K} := \{f_0, f_1, \ldots\}. \tag{8.49}$$

From the monotonicity of the x_j, the greedy algorithm for $\mathcal{K}$ in X can choose the elements from $\mathcal{K}$ in the natural order $f_0, f_1, \ldots$. Hence,

$$\sigma_j = \sigma_j(\mathcal{K})_V = x_j, \quad j \geq 0. \tag{8.50}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We want to give an upper bound for the Kolmogorov width of $\mathcal{K}$. For this, we shall use the following result (see (7.2) of Chapter 14 in Lorentz, von Golitschek and Makovoz 1996) on Kolmogorov n-widths of the m-dimensional unit ball b_1^m of ℓ^1 in the ℓ^∞-metric, in $\mathbb{R}^m$:

$$d_n(b_1^m)_V \leq C_0\big(\log_2(m/n)\big)^{1/2} n^{-1/2}, \quad 1 \leq n \leq m/2. \tag{8.51}$$

Let us now define the sequence $\{x_j\}_{j\geq 0}$ so that in position $2^{k-1} \leq j \leq 2^k - 1$ it has the constant value 2^{-ks}, for all $k \geq 0$, where $s > 1/2$. It follows that

$$\sigma_n(\mathcal{K})_V \geq cn^{-s}, \quad n \geq 1, \tag{8.52}$$

for some $c > 0$. We now bound the n-width of $\mathcal{K}$ when $n = 2^{k+2}$ by constructing a good space V_n of dimension at most n for approximating $\mathcal{K}$. The space V_n is defined as the span of a set E of at most n vectors, which we construct as follows. First, we place into E all of the vectors, $e_0, e_1, \ldots, e_{2^k-1}$. Next, for each $j = 0, 1, \ldots, n$, we use (8.51) to choose a basis for the space of dimension 2^{n-j} whose vectors are supported on $\{2^{k+j}, 2^{k+j+1} - 1\}$, and this space approximates in V each of the f_i for $i = 2^{k+j}, \ldots, 2^{k+j+1} - 1$, to accuracy

$$C_0 2^{-(k+j+1)s}\sqrt{2j}\,2^{-(k-j)/2} \leq C_0 2^{-(k+j)s}\sqrt{j}\,2^{-(k-j)/2},$$

where we used the fact that $s > 1/2$. We place these basis vectors into E so that

$$\#(E) \leq 2^k + 2^{k+1} - 1 \leq n. \tag{8.53}$$

Note that $|x_i| \leq 2^{-2ks}$ for $i \geq 2^{2k}$. This means that for the space $V_n := \operatorname{span}(E)$, with $n = 2^{k+2}$,

$$\begin{aligned} d_n(\mathcal{K})_V &\leq \operatorname{dist}(\mathcal{K}, V_n)_V \leq \max\Big\{2^{-2ks}, \max_{1\leq j\leq n} C_0 2^{-(k+j)s} 2^{-(k-j)/2}\sqrt{j}\Big\} \\ &= \max\Big\{2^{-2ks}, C_0 2^{-k(s+1/2)} \cdot \max_{1\leq j\leq k} 2^{-j(s-1/2)}\sqrt{j}\Big\} \leq C_1 n^{-(s+1/2)}. \end{aligned}$$

From the monotonicity of $(d_n(\mathcal{K})_V)_{n\geq 0}$, we obtain that

$$d_n(\mathcal{K})_V \leq C_2 n^{-s-1/2}, \quad n \geq 1. \tag{8.54}$$

This example shows that the loss of $1/2$ which appears in (ii) of Corollary 8.7 can not be avoided in general.

8.4. Space discretization and convergence analysis

The greedy algorithms introduced in the previous subsection are at this stage only theoretical algorithms because they involve several steps that cannot be implemented numerically. To describe a numerical version of these algorithms that are applicable to solving parametric PDEs, we place ourselves in the following numerical setting. We assume that we are given

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

a target accuracy $\varepsilon > 0$ and we wish to find a space $V_n = \text{span}\{g_1, \ldots, g_n\}$, where $n = n(\varepsilon)$, such that

$$\text{dist}(\mathcal{M}, V_n)_V := \max_{v \in \mathcal{M}} \text{dist}(v, V_n)_V = \sup_{a \in \mathcal{A}} \text{dist}(u(a), V_n)_V \leq \varepsilon, \tag{8.55}$$

and of course we want n to be small. In the reduced basis method, the functions g_i are picked from the solution manifold $\mathcal{M}$, or equivalently, are of the form

$$g_i = u(a^i), \tag{8.56}$$

where $\{a^1, \ldots, a^n\}$ are picked from the parameter set $\mathcal{A}$. Our benchmark is given by the n-width of $\mathcal{M}$. Namely, we know that as soon as $d_n(\mathcal{M})_V \leq \varepsilon$ then there is a space of this dimension n which satisfies (8.55). We have seen that the theoretical greedy algorithms also give us such a space V_n with provable bounds on performance, namely with rate guarantees on the growth of n with respect to ε comparable to the n-width, as expressed by Corollaries 8.4 and 8.7. However, the greedy algorithm as it stands cannot be implemented numerically, for several reasons which we now delineate.

Issue 1: Computing the greedy selection g_k. Once the parameter a^k of the kth greedy selection is identified, the function $g_k := u(a^k)$ cannot be computed exactly. In practice, it is computed approximately by space discretization in the finite element method in the space V_h.

This means that we take

$$g_k = u_h(a^k) \in V_h, \tag{8.57}$$

and so the spaces V_n are subspaces of V_h. As explained further, this may be viewed as applying the weak greedy algorithm to the *approximate solution manifold* defined by

$$\mathcal{M}_h := \{u_h(a) : a \in \mathcal{A}\}. \tag{8.58}$$

Recall that

$$\text{dist}(\mathcal{M}_h, \mathcal{M})_V \leq \max_{a \in \mathcal{A}} \|u(a) - u_h(a)\|_V \leq \varepsilon(h), \tag{8.59}$$

where $\varepsilon(h)$ is the accuracy of the numerical solver. In order to reach the goal (8.55), we pick h such that $\varepsilon(h) \leq \varepsilon/3$.

Issue 2: Search over the manifold $\mathcal{M}_h$. The kth greedy step requires a search over the entire manifold $\mathcal{M}_h$ to choose the next basis function g_k. Since the manifold is typically an infinite set, this search has to be discretized.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

One way to handle this issue is by finding a finite set $\mathcal{M}_{h,\varepsilon} \subset \mathcal{M}_h$ such that each element in $\mathcal{M}_h$ is at distance at most $\varepsilon/3$ from $\mathcal{M}_{h,\varepsilon}$, that is,

$$\sup_{a\in\mathcal{A}} \operatorname{dist}(u(a), \mathcal{M}_{h,\varepsilon})_V \leq \varepsilon/3. \tag{8.60}$$

In practice this discretization is done on the parameter side so that each $v \in \mathcal{M}_\varepsilon$ is of the form $u(a)$, $a \in \mathcal{A}_\varepsilon$, where $\mathcal{A}_\varepsilon$ is a finite subset of $\mathcal{A}$. If we apply the weak greedy algorithm to $\mathcal{M}_{h,\varepsilon}$ until we are guaranteed that the resulting space V_n satisfies $\operatorname{dist}(\mathcal{M}_{h,\varepsilon}, V_n)_V \leq \varepsilon/3$, then the goal (8.55) is met, since

$$\operatorname{dist}(\mathcal{M}, V_n)_V \leq \operatorname{dist}(\mathcal{M}, \mathcal{M}_h)_V + \operatorname{dist}(\mathcal{M}_h, \mathcal{M}_{h,\varepsilon})_V + \operatorname{dist}(\mathcal{M}_{h,\varepsilon}, V_n)_V \leq \varepsilon. \tag{8.61}$$

Issue 3: Computation of $\operatorname{dist}(u_h(a), V_k)_V$ for $a \in \mathcal{A}$ (or $a \in \mathcal{A}_\varepsilon$). At each iteration k of the greedy algorithm, we need to compute $\operatorname{dist}(u_h(a),V_k)_V$ to sufficient accuracy such that, when selecting g_k based on these computed distances, we are certain that the weak greedy criterion (8.8) is satisfied.

Here, we want to avoid computing $u_h(a)$ itself since this is too costly and must be done many times, that is, for each $a \in \mathcal{A}_\varepsilon$. Instead, this computation is done by a surrogate $d(a, V_k)_V$, which is typically evaluated by a residual-based *a posteriori* analysis from the Galerkin approximation to $u_h(a)$ from V_k. This surrogate satisfies

$$\delta d(a, V_k)_V \leq \operatorname{dist}(u_h(a), V_k)_V \leq \beta d(a, V_k)_V. \tag{8.62}$$

A practical construction of this surrogate is discussed further in the particular case of the elliptic problem (1.5). It follows that maximizing this surrogate in place of the true error amounts to applying the weak greedy algorithm with constant $\gamma := \delta/\beta$ to the approximate solution manifold $\mathcal{M}_h$.

We can now combine the proposed solutions to each of the stated Issues 1, 2 and 3, and form the following numerical version of the weak greedy algorithm.

Numerical weak greedy algorithm. We assume we are given a numerical tolerance ε and that for each subspace V_n of V_h, we have, in hand, a surrogate $d(a, V_n)_V$ which satisfies (8.62) with uniform constants δ, β. We first construct a set $\mathcal{A}_\varepsilon$ of parameters for which the discrete set $\mathcal{M}_{h,\varepsilon}$ satisfies (8.60). We now run the pure greedy algorithm on the compact set $\mathcal{K} := \mathcal{M}_{h,\varepsilon}$, but using the surrogate $d(a, V_n)_V$ in place of $\operatorname{dist}(u(a), V_n)_V$. This means that the new element $g_n = u_h(a^n)$ is defined by

$$a^n := \arg\max\{d(a, V_{n-1})_V : a \in \mathcal{A}_\varepsilon\}. \tag{8.63}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We stop the algorithm at the first value $n = n(\varepsilon)$ for which

$$\max\{d(a, V_n)_V : a \in \mathcal{A}_\varepsilon\} \leq \frac{\varepsilon}{3\beta} \tag{8.64}$$

The output of this perturbed greedy algorithm is our reduced basis space V_n.

In view of the previous discussion, on Issues 1, 2 and 3, the output space satisfies the goal (8.55). As an immediate consequence of Corollary 8.4, we obtain one first result on its number of steps $n(\varepsilon)$, which uses assumptions on the n-width of $\mathcal{M}_h$.

Theorem 8.8. For the above algorithm, we have the following.

(i) For any $s > 0$ and $C_0 > 0$,

$$\begin{aligned} d_n(\mathcal{M}_h)_V \leq C_0(\max\{1, n\})^{-s}, \quad n \geq 0 \quad \Rightarrow \\ n(\varepsilon) \leq \left(\frac{\varepsilon}{3\beta C_1}\right)^{-1/s}, \qquad \varepsilon > 0, \end{aligned} \tag{8.65}$$

where $C_1 := \gamma^{-2} 2^{4s+1} C_0$.

(ii) For any $s > 0$ and $c_0, C_0 > 0$,

$$\begin{aligned} d_n(\mathcal{M}_h)_V \leq C_0\, \mathrm{e}^{-c_0 n^s}, \qquad\qquad n \geq 0 \quad \Rightarrow \\ n(\varepsilon) \leq \left(\frac{1}{c_1} \max\left\{\log\left(\frac{\varepsilon}{3\beta \tilde{C}_1}\right), 0\right\}\right)^{1/s}, \quad \varepsilon > 0, \end{aligned} \tag{8.66}$$

where $c_1 = (c_0/2)3^{-s}$ and $\tilde{C}_1 := C_0 \max\{\sqrt{2}\gamma^{-1}, \mathrm{e}^{c_1}\}$.

Proof. This is a direct application of items (ii) and (iii) in Corollary 8.4, using the fact that $d_n(\mathcal{M}_{h,\varepsilon})_V \leq d_n(\mathcal{M}_h)_V$. □

Let us observe that the assumptions in the above theorem are on the decay of the n-widths of $\mathcal{M}_h$, in contrast to Corollary 8.4 which uses assumptions on the decay of the n-widths of $\mathcal{M}$. As already explained in Section 6.3, the approximate solution map u_h may often be viewed as the solution map of a discrete parametrized problem of the form (6.103) with similar properties as the original parametric problem (1.1). This allows us to apply the same techniques as in Section 4 in order to evaluate $d_n(\mathcal{M}_h)_V$ and justify the validity of the assumptions in the above corollary for relevant instances of parametric PDEs.

In more general cases, we may be able to justify the decay of $d_n(\mathcal{M})_V$ but not of $d_n(\mathcal{M}_h)_V$. This occurs for example if the solver involves a different finite element space for each instance, such as in adaptive methods. Then we may still write

$$d_n(\mathcal{M}_h)_V \leq d_n(\mathcal{M})_V + \varepsilon(h). \tag{8.67}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

This means, for example, that if we start from the assumption that

$$d_n(\mathcal{M})_V \leq C_0(\max\{1,n\})^{-s},$$

we need to study the weak greedy algorithm applied to $\mathcal{M}_{h,\varepsilon}$, but under the modified assumption

$$d_n(\mathcal{M}_h)_V \leq C_0(\max\{1,n\})^{-s} + \varepsilon(h). \tag{8.68}$$

We may then separate n between the ranges $\{1,\dots,N(h)\}$ where $\varepsilon(h) \leq C_0(\max\{1,n\})^{-s}$ and the larger values of n. Then, having fixed $\varepsilon(h) = \varepsilon/3$ and using Remark 8.5, we reach a similar conclusion on the order of magnitude of $n(\varepsilon)$ as in Theorem 8.8. The same holds for exponential rates.

8.5. Computational cost

We now turn to the analysis of the computational cost required by the numerical weak greedy algorithm in order to reach the accuracy goal (8.55). For simplicity, we restrict our attention to the regime of algebraic rates described by item (ii) of Theorem 8.8. A similar analysis can be carried out for exponential rates. Here, we only consider linear elliptic problems expressed in variational form, which are particular cases of those treated in Section 7: Find $u \in V$ such that

$$B(u,v;a) = L(v), \quad v \in V, \tag{8.69}$$

where $B(\cdot,\cdot;a)$ and L are continuous sesquilinear and antilinear forms over $V \times V$ and $\tilde{V}$ respectively, and where we make the additional assumption that

$$a \mapsto B(\cdot,\cdot;a) \tag{8.70}$$

is a continuous *linear* map from X to $\mathfrak{B}$, the set of continuous sesquilinear forms over $V \times V$. We use the following symmetric elliptic version of Assumption AL.

Assumption ALE. The parameter set $\mathcal{A}$ has a complete affine representer $(\psi_j)_{j\geq 1}$ and, for all $a \in a(U)$, the sesquilinear form $B(\cdot,\cdot;a)$ satisfies the coercivity conditions (2.14) and it is symmetric when restricted to real-valued functions of V.

Under such an assumption, the approximate solution $u_h(a) \in V_h$ is defined by the Galerkin method, that is,

$$B(u_h(a),v_h;a) = L(v_h), \quad v_h \in V_h, \tag{8.71}$$

and can be computed for any given $a \in a(U)$ by the numerical solver at cost C_h.

We now turn to the online cost of the numerical weak greedy algorithm assuming that we have already computed in the offline stage the reduced

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

basis elements $g_k = u_h(a^k)$, $k = 0, \dots, n-1$, by using the possibly expensive finite element solver for u_h. Given a query $a \in \mathcal{A}$, the online stage computes $u_n(a) \in V_n$, where V_n is the reduced basis space. We recall that $V_n \subset V_h$. We find $u_n(a)$ by the Galerkin method for V_n, that is,

$$B(u_n(a), v_n; a) = L(v_n), \quad v_n \in V_n, \tag{8.72}$$

This reduces to solving an $n \times n$ linear system, where the unknowns are the coefficients $\alpha_l(a)$ in the decomposition

$$u_n(a) = \sum_{l=0}^{n-1} \alpha_l(a) g_l. \tag{8.73}$$

Note that, as opposed to stiffness matrices resulting from the discretization of PDEs in a nodal finite element basis, the resulting stiffness matrix

$$\mathbf{B}_n(a) = (B(g_k, g_l; a))_{k,l=0,\dots,n-1} \tag{8.74}$$

is generally full. Using a direct solver, such as Gauss elimination, the cost of solving this system is therefore of order

$$n(\varepsilon)^3 \sim \varepsilon^{-3/s}. \tag{8.75}$$

However, we also need to take into account the cost of assembling the system, that is, computing the above stiffness matrix, which depends on a. Since the data vector $\mathbf{F}_n := (L(g_k))_{k=0,\dots,n-1}$ of this system does not depend on a, its computation can be performed during the offline stage. In order to compute the stiffness matrix, we recall the bilinear forms $\overline{B}$, B_j and $B(\cdot, \cdot, y)$ defined in Section 7.1. If $y \in U$ and

$$a = a(y) = \overline{a} + \sum_{j \geq 1} y_j \psi_j, \tag{8.76}$$

the stiffness matrix is

$$\mathbf{B}_n(y) = \overline{\mathbf{B}}_n + \sum_{j \geq 1} y_j \mathbf{B}_{n,j}, \tag{8.77}$$

where

$$\overline{\mathbf{B}}_n := (\overline{B}(g_k, g_l))_{k,l=0,\dots,n-1} \quad \text{and} \quad \mathbf{B}_{n,j} := (B_j(g_k, g_l))_{k,l=0,\dots,n-1} \tag{8.78}$$

are $n \times n$ matrices. Each of these matrices can be computed in the offline stage, but the infinite sum over $j \geq 1$ needs to be truncated at some prescribed level J. In the case where $(\|\psi_j\|_X)_{j \geq 1}$ is ℓ^p-summable, and if the ψ_j are organized such that the $\|\psi_j\|_X$ are non-increasing with j, then we know that the $L^\infty(U, V)$ error in the approximation of the solution

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

map $y \mapsto u_h(y)$ resulting from this truncation is of order $\mathcal{O}(J^{-s})$, where $s := \frac{1}{p} - 1$, and therefore the order of accuracy ε can be preserved by taking

$$J = J(\varepsilon) \sim \varepsilon^{-1/s}. \tag{8.79}$$

We may thus incorporate such a truncation in the definition of the approximation map $y \to u_h(y)$ used to handle Issue 1. Note that the choice of J depends only on ε and is independent of h. Therefore, using this u_h, the conclusion of (i) in Theorem 8.8 is retained and the cost of assembling the system is

$$J(\varepsilon)n(\varepsilon)^2 \sim \varepsilon^{-3/s}. \tag{8.80}$$

Note that, once the coefficients $\alpha_k(a)$ are found, computing the finite element representation of the solution $u_n(a) = \sum_{k=0}^{n-1} \alpha_k(a) g_k$ has a cost of $n(\varepsilon)N_h$. In conclusion, the total online cost is of order

$$C_{\mathrm{on}}(\varepsilon) \sim \varepsilon^{-3/s} + n(\varepsilon)N_h. \tag{8.81}$$

However, note that in some applications, one may only work with the reduced basis representation $(\alpha_k(a))_{k=0,\dots,n-1}$, without the need to recompute the finite element representation. This the case, for instance, when manipulating a quantity of interest such as a linear scalar functional

$$Q(u_n(a)) = \sum_{k=0}^{n-1} \alpha_k(a) Q(g_k). \tag{8.82}$$

Having precomputed the quantities $Q(g_k)$ in the offline stage, the online evaluation of this quantity is therefore executed at cost of order $\varepsilon^{-3/s}$.

Also note that $u_n(a)$ is not the best approximation of $u(a)$ from V_n in the norm V since it is the Galerkin projection onto V_n. However, from Cea's lemma we have

$$\|u(a) - u_n(a)\|_V \le \sqrt{\frac{R}{\alpha}} \min_{v \in V_n} \|u(a) - v\|_V = \sqrt{\frac{R}{\alpha}} \mathrm{dist}(\mathcal{M}, V_n)_V, \tag{8.83}$$

where α is the constant in (2.14) and $R := \max_{a \in \mathcal{A}} \|B(\cdot, \cdot; a)\|_{\mathfrak{B}}$. This guarantees that we reach an error of the prescribed order ε between $u(a)$ and its reduced basis approximation $u_n(a)$.

We next discuss the offline cost. The first step is to find an $\varepsilon/3$ net $\mathcal{M}_{h,\varepsilon}$ for $\mathcal{M}_h$. We describe this net only through the parameter set $\mathcal{A}$, namely as

$$\mathcal{M}_{h,\varepsilon} = u_h(\mathcal{A}_\varepsilon), \tag{8.84}$$

where $\mathcal{A}_\varepsilon$ is a finite subset of $\mathcal{A}$ such that

$$\sup_{a \in \mathcal{A}} \mathrm{dist}(a, \mathcal{A}_\varepsilon)_X \le \frac{\varepsilon}{3C}, \tag{8.85}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

and C is a Lipschitz constant for the map $a \mapsto u_h(a)$. The same type of computation as done in (2.23) for the particular problem (1.5) shows that

$$C = \frac{\|L\|_W}{\alpha^2} \tag{8.86}$$

is an admissible Lipschitz constant. This implies that the resulting $\mathcal{M}_{h,\varepsilon}$ satisfies (8.60). Note that we do not need to compute the elements of $\mathcal{M}_{h,\varepsilon}$ but only the parameter values in $\mathcal{A}_\varepsilon$.

We can bound the cardinality of $\mathcal{A}_\varepsilon$ from results on covering numbers and n-widths. Let us recall that the covering number $N_\delta := N_\delta(\mathcal{A}, X)$ is the smallest number of X-balls of radius δ that cover $\mathcal{A}$. Let $B(a_i, \delta)$, $i = 1, \dots, N_\delta$, be such a covering. Note that the a_i need not be from $\mathcal{A}$, but this is easily remedied. Namely, for any such ball we have $B(a_i, \delta) \cap \mathcal{A} \neq \emptyset$, and so we choose an $a_i \in B(a_i, \delta) \cap \mathcal{A}$. Then the balls $B(a_i, 2\delta)$ are a covering of $\mathcal{A}$ with centres from $\mathcal{A}$. Therefore, taking $\eta = (2\varepsilon)/(3C)$, we can find a set $\mathcal{A}_\varepsilon \subset \mathcal{A}$ satisfying (8.85) with

$$\#(\mathcal{A}_\varepsilon) \leq N_{\eta/2}(\mathcal{A}, X) = N_{\varepsilon/(3C)}(\mathcal{A}, X). \tag{8.87}$$

The well-known Carl's inequality (Pisier 1989) gives a bound on the covering numbers $N_\delta(\mathcal{A}, X)$ in terms of the n-widths $d_n(\mathcal{A})_X$ of $\mathcal{A}$. In our case, this inequality gives

$$N_\eta(\mathcal{A}, L^\infty) \leq C_1 2^{\eta^{-1/s}}, \tag{8.88}$$

where C_1 is a constant depending on s. This gives us the bound

$$\#(\mathcal{A}_\varepsilon) \leq C_1 2^{c_1 \varepsilon^{-1/s}}, \tag{8.89}$$

for a constant c_1 that also depends on s. While it is generally not possible, at least in any reasonable way, to find a minimal set $\mathcal{A}_\varepsilon$, typically we can give a simple description of a set $\mathcal{A}_\varepsilon$ for which (8.87) is still satisfied for an appropriate constant C_1. For example, whenever we can construct a sequence of spaces X_n for which $\operatorname{dist}(\mathcal{A}, X_n)_X = \mathcal{O}(n^{-s})$, then the proof of Carl's inequality (see, *e.g.*, Lorentz, von Golitschek and Makovoz 1996) gives an explicit description of such an $\mathcal{A}_\varepsilon$. In particular, under the assumption $(\|\psi_j\|_X) \in \ell^p(\mathbb{N})$ and with the ψ_j organized in non-increasing X-norms, we can take $X_n := \operatorname{span}\{\psi_1, \dots, \psi_n\}$, for each $n \geq 1$, and the description of $\mathcal{A}_\varepsilon$ amounts to defining a specific lattice discretization U_ε of U such that $\mathcal{A}_\varepsilon = a(U_\varepsilon)$.

Let us now evaluate the cost of the kth step of the numerical weak greedy algorithm. This step includes the computation of the reduced basis element $g_k := u_h(a^k)$ once a^k has been chosen, using the possibly expensive finite element solver, which has cost of order C_h. On the other hand, we must also account for the maximization of the surrogate $d(a, V_{k-1})_V$ over the set $\mathcal{A}_e$.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

This cost is of order

$$\#(\mathcal{A}_\varepsilon)s_k, \tag{8.90}$$

where s_k is the cost of computing $d(a, V_{k-1})_V$ for one value of a.

We now give a derivation of a possible surrogate and evaluate the cost s_k for this particular surrogate. Since for the reduced basis solution $u_k(a) \in V_k$ we have

$$\sqrt{\frac{\alpha}{R}}\|u_h(a) - u_k(a)\|_V \leq \operatorname{dist}(u_h(a), V_k)_V \leq \|u_h(a) - u_k(a)\|_V, \tag{8.91}$$

this surrogate should be an equivalent quantity to $\|u_h(a) - u_k(a)\|_V$. We introduce the $N_h \times N_h$ stiffness matrix

$$\mathbf{B}_h(y) = \overline{\mathbf{B}}_h + \sum_{j=1}^{J} y_j \mathbf{B}_{h,j}, \tag{8.92}$$

for the sesquilinear form $B(\cdot, \cdot; y)$ in the finite element basis, where $\overline{\mathbf{B}}_h$ and $\mathbf{B}_{h,j}$ are the corresponding stiffness matrices for $\overline{B}$ and B_j. Therefore, the coordinate vector $U_h(y)$ of $u_h(y) = u_h(a(y))$ in the finite element basis is the solution of the $N_h \times N_h$ system

$$\mathbf{B}_h(y)U_h(y) = F_h, \tag{8.93}$$

where the right-hand side F_h does not depend on y. Here $J = J(\varepsilon)$ is the truncation level, already introduced for the evaluation of the online cost, and u_h is defined as the discrete solution of the truncated problem. We also introduce the coordinate vectors G_i of the reduced basis elements $u_h(a^i)$ in the finite element basis. Therefore, a reduced basis solution $u_k(y) = u_k(a(y))$ is represented in the finite element basis by the vector

$$U_k(y) = \sum_{i=0}^{k-1} \alpha_i(y) G_i. \tag{8.94}$$

We introduce a Hilbertian norm $\|\cdot\|_*$ on $\mathbb{R}^{N_h}$, defined in such way that

$$\|W_h\|_* := \|w_h\|_V \tag{8.95}$$

whenever W_h is the coordinate vector of $w_h \in V_h$. Note that this would coincide with the Euclidean norm if the finite basis were orthonormal in V. This is generally not the case, but nevertheless the computation of this norm is usually of complexity N_h. We may thus write

$$\|u_h(y) - u_k(y)\|_V = \|U_h(y) - U_k(y)\|_*. \tag{8.96}$$

By (8.91), it follows that

$$\frac{\alpha}{R}\|U_h(y) - U_k(y)\|_*^2 \leq \operatorname{dist}(u_h(y), V_k)_V^2 \leq \|U_h(y) - U_k(y)\|_*^2. \tag{8.97}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

Our next observation is that since, for any $y \in U_{\mathcal{A}}$,

$$\frac{\alpha}{R}\langle \overline{\mathbf{B}}_h W_h, W_h\rangle \leq \langle \mathbf{B}_h(y) W_h, W_h\rangle \leq \frac{R}{\alpha}\langle \overline{\mathbf{B}}_h W_h, W_h\rangle, \tag{8.98}$$

we have the norm equivalence

$$\frac{\alpha}{R}\|W_h\|_* \leq \|\overline{\mathbf{B}}_h^{-1}\mathbf{B}_h(y) W_h\|_* \leq \frac{R}{\alpha}\|W_h\|_*. \tag{8.99}$$

Therefore, we can define a surrogate quantity by

$$d(y, V_{k-1})_V := \|\overline{\mathbf{B}}_h^{-1}\mathbf{B}_h(y)(U_h(y) - U_k(y))\|_*, \tag{8.100}$$

and obtain the equivalence (8.62) with constants

$$\delta = \left(\frac{\alpha}{R}\right)^{3/2} \quad \text{and} \quad \beta = \frac{R}{\alpha}. \tag{8.101}$$

This surrogate is computable since we have

$$\begin{aligned} d(y, V_{k-1})_V^2 &= \left\|\overline{\mathbf{B}}_h^{-1} F_h - \overline{\mathbf{B}}_h^{-1}\mathbf{B}_h(y) U_k(y)\right\|_*^2 \\ &= \left\|\overline{\mathbf{B}}_h^{-1} F_h - \overline{\mathbf{B}}_h^{-1}\left(\overline{\mathbf{B}}_h + \sum_{j=0}^{J} y_j \mathbf{B}_{h,j}\right)\sum_{i=0}^{k-1}\alpha_i(y) G_i\right\|_*^2. \end{aligned}$$

Expanding this squared norm, we find that it is the sum of the constant term $\|\overline{\mathbf{B}}_h^{-1} F_h\|_*^2$ and of a linear combination of the real numbers $\alpha_i(y)$, $\alpha_i(y)\alpha_{i'}(y)$, $y_j\alpha_i(y)$ and $y_j y_{j'}\alpha_i(y)\alpha_{i'}(y)$ for $i, i' = 0, \ldots, k-1$ and $j, j' = 1, \ldots, J$. The coefficients of these linear combinations are given by the $\langle\cdot,\cdot\rangle_*$ inner products (associated to the $\|\cdot\|_*$-norm) between pairs of vectors chosen from

$$\overline{\mathbf{B}}_h^{-1} F_h, \quad G_i, \quad \overline{\mathbf{B}}_h^{-1}\mathbf{B}_{h,j} G_i, \quad i = 0, \ldots, k-1,\ j = 1, \ldots, J. \tag{8.102}$$

The precomputation of these vectors and of their inner product has a cost of order

$$kJC_h + k^2 J^2 N_h. \tag{8.103}$$

Then, the computation of the surrogate $d(y, V_{k-1})_V$ for each y has cost of order k^2J^2 for the linear combination, to which we must add the cost of computing the $\alpha_i(y)$, which according to the discussion of the online cost is of order k^3. Therefore

$$s_k \sim k^2 J^2 + k^3. \tag{8.104}$$

In summary, the total cost of step k of the algorithm, without including the precomputations, is

$$C_h + \#(\mathcal{A}_\varepsilon)(k^2 J^2 + k^3), \tag{8.105}$$

so the total cost up to step $n = n(\varepsilon)$ is of order

$$n(\varepsilon) C_{h(\varepsilon)} + \#(\mathcal{A}_\varepsilon) n(\varepsilon)^3 J(\varepsilon)^2 + \#(\mathcal{A}_\varepsilon) n(\varepsilon)^4. \tag{8.106}$$

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

We need to add the cost of precomputing

(i) the vectors $\overline{\mathbf{B}}_h^{-1}F_h$, G_i and $\overline{\mathbf{B}}_h^{-1}\mathbf{B}_{h,j}G_i$, for $i = 0, \dots, n-1$ and $j = 1, \dots, J$ and their $\langle \cdot, \cdot \rangle_*$ inner products,

(ii) the matrices $\overline{\mathbf{B}}_k$ and $\mathbf{B}_{k,j}$ for $k = 0, \dots, n-1$, whose entries are given by the Euclidean inner products between the vectors G_k and the vectors $\overline{\mathbf{B}}_h G_i$ and $\mathbf{B}_{h,j}G_i$.

This precomputing cost is of total order

$$n(\varepsilon)J(\varepsilon)C_{h(\varepsilon)} + n(\varepsilon)^2 J(\varepsilon)^2 N_{h(\varepsilon)}. \tag{8.107}$$

In summary, the total offline cost is of order

$$C_{\text{off}} \sim n(\varepsilon)J(\varepsilon)C_{h(\varepsilon)} + n(\varepsilon)^2 J(\varepsilon)^2 N_{h(\varepsilon)} + \#(\mathcal{A}_\varepsilon)n(\varepsilon)^3 J(\varepsilon)^2 + \#(\mathcal{A}_\varepsilon)n(\varepsilon)^4. \tag{8.108}$$

Among these terms, the largest is typically the third one which in our algebraic rate regime is of order $\varepsilon^{-5/s}2^{c_1\varepsilon^{-1/s}}$ in view of (8.87).

This offline cost is thus potentially extremely large. Note, however, that it is due to the fact that we are using a brutal discrete search over $\mathcal{A}_\varepsilon$ for the maximization of the surrogate quantity, so that there is room for improvement by using more sophisticated optimization strategies. Note also that in the case of a parametric problem with moderate number d of parameters, the quantity $J(\varepsilon)$ can simply be replaced by d.

In conclusion, we find that, compared to the polynomial methods discussed in Sections 6 and 7, the reduced basis method suffers from a very high offline cost, especially in high parametric dimension. This can be compensated by the fact that this method captures the same rate of decay as achieved by the optimal n-width spaces, so that a prescribed accuracy ε may be achieved with a number $n = n(\varepsilon)$ of reduced basis elements much smaller than the number of terms in polynomial expansions for the same accuracy, making the online cost potentially lower.

REFERENCES

R. Andreev, M. Bieri and C. Schwab (2006), 'Sparse tensor discretization of elliptic SPDEs', *SIAM J. Sci. Comput.* **31**, 4281–4304.

I. Babuška, F. Nobile and R. Tempone (2007), 'A stochastic collocation method for elliptic partial differential equations with random input data', *SIAM J. Numer. Anal.* **45**, 1005–1034.

I. Babuška, R. Tempone and G. E. Zouraris (2004), 'Galerkin finite element approximations of stochastic elliptic partial differential equations', *SIAM J. Numer. Anal.* **42**, 800–825.

J. Beck, F. Nobile, L. Tamellini and R. Tempone (2012), 'On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods', *Math. Models Methods Appl. Sci.* **22**, 1–33.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

J. Beck, F. Nobile, L. Tamellini and R. Tempone (2014), 'Convergence of quasi-optimal stochastic Galerkin methods for a class of PDEs with random coefficients', *Comput. Math. Appl.* **67**, 732–751.

M. Bergmann and L. Cordier (2003), *Proper Orthogonal Decomposition: An Overview; Post Processing of Experimental and Numerical Data*, Lecture Series 2003/2004, von Karman Institut for Fluid Dynamics.

P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova and P. Wojtaszczyk (2011), 'Convergence rates for greedy algorithms in reduced basis methods', *SIAM J. Math. Anal.* **43**, 1457–1472.

P. Binev, W. Dahmen and R. DeVore (2004), 'Adaptive finite element methods with convergence rates', *Numer. Math.* **97**, 219–268.

C. de Boor and A. Ron (1992), 'Computational aspects of polynomial interpolation in several variables', *Math. Comp.* **58**, 705–727.

S. Brenner and L. R. Scott (2008), *The Mathematical Theory of Finite Elements*, second edition, Springer.

A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme and G. Turinici (2012) '*A priori* convergence of the greedy algorithm for the parameterized reduced basis', *Math. Model. Numer. Anal.* **46**, 595–603.

H.-J. Bungartz and M. Griebel (2004), Sparse grids. In *Acta Numerica*, Vol. 13, Cambridge University Press, pp. 1–123.

J. P. Calvi and V. M. Phung (2011), 'On the Lebesgue constant of Leja sequences for the unit disk and its applications to multivariate interpolation', *J. Approx. Theory* **163**, 608–622.

J. P. Calvi and V. M. Phung (2012), 'Lagrange interpolation at real projections of Leja sequences for the unit disk', *Proc. Amer. Math. Soc.* **140**, 4271–4284.

E. R. Canfield, P. Erdős and C. Pomerance (1983), 'On a problem of Oppenheim concerning "factorisatio numerorum"', *J. Number Theory* **17**, 1–28.

A. Chkifa (2013), 'On the Lebesgue constant of Leja sequences for the complex unit disk and of their real projection', *J. Approx. Theory* **166**, 176–200.

A. Chkifa (2015), New bounds on the Lebesgue constant of Leja sequences on the unit disc and their projections $\mathfrak{R}$-Leja sequences. arXiv:1503.01731

A. Chkifa, A. Cohen, R. DeVore and C. Schwab (2013), 'Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs', *Math. Model. Numer. Anal.* **47**, 253–280.

A. Chkifa, A. Cohen, G. Migliorati, F. Nobile and R. Tempone (2015*a*), 'Discrete least squares polynomial approximation with random evaluations: Application to parametric and stochastic PDEs', *Math. Model. Numer. Anal.* doi:10.1051/m2an/2014050

A. Chkifa, A. Cohen and C. Schwab (2014), 'High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs', *Found. Comput. Math.* **14**, 601–633.

A. Chkifa, A. Cohen and C. Schwab (2015*b*), 'Breaking the curse of dimensionality in parametric PDEs', *J. Math. Pures Appliquées* **102**, 400–428.

P. G. Ciarlet (1978), *The Finite Element Method for Elliptic Problems*, Elsevier.

A. Cohen and R. DeVore (2015), Kolmogorov widths under holomorphic mappings. *IMA J. Numer. Anal.* doi:10.1093/imanum/dru066

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

A. Cohen, W. Dahmen and R. DeVore (2000), 'Adaptive wavelet methods for elliptic operator equations: Convergence rates', *Math. Comp.* **70**, 27–75.
A. Cohen, W. Dahmen and R. DeVore (2002), 'Adaptive wavelet methods for operator equations: Beyond the elliptic case', *Found. Comput. Math.* **2**, 203–245.
A. Cohen, R. DeVore and C. Schwab (2010), 'Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs', *Found. Comput. Math.* **10**, 615–646.
A. Cohen, R. DeVore and C. Schwab (2011), 'Analytic regularity and polynomial approximation of parametric and stochastic PDEs', *Anal. Appl.* **9**, 11–47.
P. Constantine, M. Eldred and E. Phipps (2012), 'Sparse pseudospectral approximation method', *Comput. Methods Appl. Mech. Engrg* **229–232**, 1–12.
R. Dautray and J.-L. Lions (1992), *Mathematical Analysis and Numerical Methods for Science and Technology*, Springer.
P. J. Davis (1963), *Interpolation and Approximation*, Blaisdell.
R. DeVore (1998), Nonlinear approximation. In *Acta Numerica*, Vol. 7, Cambridge University Press, pp. 51–150.
R. DeVore, G. Petrova and P. Wojtaszczyk (2013), 'Greedy algorithms for reduced bases in Banach spaces', *Constr. Approx.* **37**, 455–466.
J. Dieudonné (1969), *Treatise on Analysis*, Vol. I, Academic Press.
A. Doostan and G. Iaccarino (2009), 'A least-squares approximation of partial differential equations with high-dimensional random inputs', *J. Comput. Phys.* **228**, 4332–4345.
A. Doostan and H. Owadi (2011), 'A non-adapted sparse approximation of PDEs with stochastic inputs', *J. Comput. Phys.* **230**, 3015–3034.
W. Dörfler (1996), 'A convergent adaptive algorithm for Poisson's equation', *SIAM J. Numer. Anal.* **33**, 1106–1124.
P. Frauenfelder, C. Schwab and R. A. Todor (2005), 'Finite elements for elliptic problems with stochastic coefficients', *Comput. Methods Appl. Mech. Engrg* **194**, 205–228.
T. Gantumur, H. Harbrecht and R. Stevenson (2007), 'An optimal adaptive wavelet method without coarsening of the iterands', *Math. Comp.* **76**, 615–629.
T. Gerstner and M. Griebel (2003), 'Dimension-adaptive tensor-product quadrature', *Computing* **71**, 65–87.
R. Ghanem and P. Spanos (1991), *Stochastic Finite Elements: A Spectral Approach*, Springer.
R. Ghanem and P. Spanos (1997), 'Spectral techniques for stochastic finite elements', *Arch. Comput. Methods Engrg* **4**, 63–100.
G. J. Gittelson (2010), 'Stochastic Galerkin discretization of the log-normal isotropic diffusion problem', *Math. Models Methods Appl. Sci.* **20**, 237–263.
C. J. Gittelson (2013), 'An adaptive stochastic Galerkin method', *Math. Comp.* **82**, 1515–1541.
C. J. Gittelson and C. Schwab (2011), Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. In *Acta Numerica*, Vol. 20, Cambridge University Press, pp. 291–467.
I. Graham, F. Kuo, J. Nichols, R. Scheichl, C. Schwab and I. Sloan (2015), 'Quasi-Monte Carlo finite element methods for elliptic PDEs with log-normal random coefficient', *Numer. Math.* doi:10.1007/s00211-014-0689-y

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

M. Gunzburger, C. Webster and G. Zhang (2014), Stochastic finite element methods for partial differential equations with random input data. In *Acta Numerica*, Vol. 23, Cambridge University Press, pp. 521–650.

M. Hansen and C. Schwab (2013*a*), 'Analytic regularity and nonlinear approximation of a class of parametric semilinear elliptic PDEs', *Math. Nachr.* **286**, 832–860.

M. Hansen and C. Schwab (2013*b*), 'Sparse adaptive approximation of high-dimensional parametric initial value problems', *Vietnam J. Math.* **41**, 181–215.

M. Hervé (1989), *Analyticity in Infinite Dimensional Spaces*, De Gruyter.

V. H. Hoang and C. Schwab (2014), 'n-term Wiener chaos approximation rates for elliptic PDEs with lognormal Gaussian random inputs', *Math. Models Methods Appl. Sci.* **24**, 797–826.

D. Jerison and C. E. Kenig (1995), 'The inhomogeneous Dirichlet problem in Lipschitz domains', *J. Funct. Anal.* **130**, 161–219.

M. Kahlbacher and S. Volkwein (2007), 'Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems', *Discuss. Math. Differ. Incl. Control Optim.* **27**, 95–117.

G. E. Karniadakis and D. B. Xiu (2002*a*), 'The Wiener–Askey polynomial chaos for stochastic differential equations', *SIAM J. Sci. Comput.* **24**, 619–644.

G. E. Karniadakis and D. Xiu (2002*b*), 'Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos', *Comput. Methods Appl. Mech. Engrg* **191**, 4927–4948.

M. Kleiber and T. D. Hien (1992), *The Stochastic Finite Element Methods*, Wiley.

O. Knio and O. Le Maitre (2010), *Spectral Methods for Uncertainty Quantication: With Applications to Computational Fluid Dynamics*, Springer.

A. Kolmogorov (1936), 'Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse', *Ann. of Math.* **37**, 107–110.

A. Kunoth and C. Schwab (2013), 'Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs', *SIAM J. Control Optim.* **51**, 2442–2471.

J. Kuntzman (1959), *Méthodes Numériques: Interpolation, Dérivées*, Dunod.

F. Y. Kuo, C. Schwab and I. H. Sloan (2012), 'Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients', *SIAM J. Numer. Anal.* **50**, 3351–3374.

F. Y. Kuo, C. Schwab and I. H. Sloan (2015), 'Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients', *Found. Comput. Math.* doi:10.1007/s10208-014-9237-5

F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski and H. Wozniakowski (2010), 'Liberating the dimension', *J. Complexity* **26**, 422–454.

G. G. Lorentz and R. Lorentz (1986), Solvability problems of bivariate interpolation I', *Constr. Approx.* **2**, 153–169.

G. G. Lorentz, M. von Golitschek and Y. Makovoz (1996), *Constructive Approximation: Advanced Problems*, Springer.

F. Luca, A. Mukhopadhyay and K. Srinivas (2010), 'On the Oppenheim's "factorisatio numerorum" function', *Acta Arithmetica* **142**, 41–50.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

L. Machiels, Y Maday, I. B. Oliveira, A. T. Patera and D. V. Rovas (2000), 'Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems', *C. R. Acad. Sci. Paris, Sér. I* **331**, 153–158.

Y. Maday, A. T. Patera and G. Turinici (2002*a*), '*A priori* convergence theory for reduced-basis approximations of single-parametric elliptic partial differential equations', *J. Sci. Comput.* **17**, 437–446.

Y. Maday, A. T. Patera and G. Turinici (2002*b*), 'Global *a priori* convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations', *C. R. Acad. Sci. Paris, Sér. I* **335**, 289–294.

G. Migliorati, F. Nobile, R. Tempone and E. Von Schwerin (2013), 'Approximation of quantities of interest in stochastic PDEs by the random discrete L^2 projection on polynomial spaces', *SIAM J. Sci. Comput.* **35**, 1440–1460.

P. Morin, R. H. Nochetto and K. G. Siebert (2000), 'Data oscillation and convergence of adaptive FEM', *SIAM J. Numer. Anal.* **38**, 466–488.

F. Nobile, R. Tempone and C. G. Webster (2008*a*), 'A sparse grid stochastic collocation method for elliptic partial differential equations with random input data', *SIAM J. Numer. Anal.* **46**, 2309–2345.

F. Nobile, R. Tempone and C. G. Webster (2008*b*), 'An anisotropic sparse grid stochastic collocation method for elliptic partial differential equations with random input data', *SIAM J. Numer. Anal.* **46**, 2411–2442.

A. K. Noor and J. M. Peters (1980), 'Reduced basis technique for nonlinear analysis of structures', *AIAA J.* **18**, 455–462.

A. Pinkus (1985), *n-Widths in Approximation Theory*, Springer.

G. Pisier (1989), *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press.

G. Rozza, D. B. P. Huynh and A. T. Patera (2008), 'Reduced basis approximation and *a posteriori* error estimation for affinely parametrized elliptic coercive partial differential equations: Application to transport and continuum mechanics', *Arch. Comput. Methods Engrg* **15**, 229–275.

T. Runst and W. Sickel (1996), *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations*, De Gruyter Series in Nonlinear Analysis and Applications, De Gruyter.

C. Schillings and C. Schwab (2013), 'Sparse, adaptive Smolyak quadratures for Bayesian inverse problems', *Inverse Problems* **29**, 065011.

C. Schillings and C. Schwab (2014), 'Sparsity in Bayesian inversion of parametric operator equations', *Inverse Problems* **30**, 065007.

S. Sen (2008), 'Reduced-basis approximation and *a posteriori* error estimation for many-parameter heat conduction problems', *Numerical Heat Transfer* B **54**, 369–389.

C. Schwab and R. Stevenson (2009), 'Space–time adaptive wavelet methods for parabolic evolution equations', *Math. Comp.* **78**, 1293–1318.

C. Schwab and A. M. Stuart (2012), 'Sparse deterministic approximation of Bayesian inverse problems', *Inverse Problems* **28**, 045003.

C. Schwab and R. Todor (2003), 'Sparse finite elements for elliptic problems with stochastic loading', *Numer. Math.* **95**, 707–734.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press

C. Schwab and R. Todor (2007), 'Convergence rates of sparse chaos approximations of elliptic problems with stochastic coefficients', *IMA J. Numer. Anal.* **44**, 232–261.

S. Smolyak (1963), 'Quadrature and interpolation formulas for tensor products of certain classes of functions', *Doklady Akademii Nauk SSSR* **4**, 240–243.

R. Stevenson (2007), 'Optimality of a standard adaptive finite element method', *Found. Comput. Math.* **7**, 245–269.

A. M. Stuart (2010), Inverse problems: A Bayesian perspective. *Acta Numerica*, Vol. 19, Cambridge University Press, pp. 451–559.

K. Veroy, C. Prud'homme, D. V. Rovas and A. T. Patera (2003), *A posteriori* error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *Proc. 16th AIAA Computational Fluid Dynamics Conference*, paper 2003-3847.

D. Xiu (2007), 'Efficient collocational approach for parametric uncertainty analysis', *Commun. Comput. Phys.* **2**, 293–309.

D. Xiu (2010), *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press.

https://doi.org/10.1017/S0962492915000033 Published online by Cambridge University Press