HAL
open science

Model order reduction methods for parameter-dependent equations – Applications in Uncertainty Quantification.

Olivier Zahm

▶ **To cite this version:**

Olivier Zahm. Model order reduction methods for parameter-dependent equations – Applications in Uncertainty Quantification.. Mathematics [math]. Ecole Centrale de Nantes (ECN), 2015. English. NNT : . tel-01256411

HAL Id: tel-01256411

https://theses.hal.science/tel-01256411v1

Submitted on 14 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE
L'ÉCOLE CENTRALE NANTES

École doctorale Sciences Pour l'Ingénieur, Géosciences, Architecture

Présentée par

Olivier ZAHM

Pour obtenir le grade de

DOCTEUR DE L'ÉCOLE CENTRALE NANTES

Sujet de la thèse

Méthodes de réduction de modèle pour les équations paramétrées – Applications à la quantification d'incertitude.

soutenue le 20 novembre 2015, devant le jury composé de :

Président:	**Yvon MADAY**	Professeur des Universités, Université Paris 6
Rapporteurs:	**Fabio NOBILE**	Professeur associé, École Polytechnique Fédérale de Lausanne
	Christophe PRUD'HOMME	Professeur des Universités, Université de Strasbourg
Examinateurs:	**Marie BILLAUD-FRIESS**	Maître de conférence, École Centrale Nantes (co-encadrant)
	Tony LELIÈVRE	Professeur à l'École des Ponts ParisTech
	Anthony NOUY	Professeur des Universités, École Centrale Nantes (directeur)
	Clémentine PRIEUR	Professeur des Universités, Université Grenoble 1

“Der, welcher wandert diese Straße voll Beschwerden,
wird rein durch Feuer, Wasser, Luft und Erden.”

—

La Flûte enchantée, deuxième acte, vingt-huitième entrée.

À mon épouse Alexandra, pour son amour et son soutien.

Résumé

Les méthodes de réduction de modèle sont incontournables pour la résolution d'équations paramétrées de grande dimension qui apparaissent dans les problèmes de quantification d'incertitude, d'optimisation ou encore les problèmes inverses. Dans cette thèse nous nous intéressons aux méthodes d'approximation de faible rang, notamment aux méthodes de bases réduites et d'approximation de tenseur.

L'approximation obtenue par projection de Galerkin peut être de mauvaise qualité lorsque l'opérateur est mal conditionné. Pour les méthodes de projection sur des espaces réduits, nous proposons des préconditionneurs construits par interpolation d'inverse d'opérateur, calculés efficacement par des outils d'algèbre linéaire "randomisée". Des stratégies d'interpolation adaptatives sont proposées pour améliorer soit les estimateurs d'erreur, soit les projections sur les espaces réduits. Pour les méthodes d'approximation de tenseur, nous proposons une formulation en minimum de résidu avec utilisation de norme idéale. L'algorithme de résolution, qui s'interprète comme un algorithme de gradient avec préconditionneur implicite, permet d'obtenir une approximation quasi-optimale de la solution.

Enfin nous nous intéressons à l'approximation de quantités d'intérêt à valeur fonctionnelle ou vectorielle. Nous généralisons pour cela les approches de type "primale-duale" au cas non scalaire, et nous proposons de nouvelles méthodes de projection sur espaces réduits. Dans le cadre de l'approximation de tenseur, nous considérons une norme dépendant de l'erreur en quantité d'intérêt afin d'obtenir une approximation de la solution qui tient compte de l'objectif du calcul.

Mots clefs: Réduction de modèle – Quantification d'incertitude – Equations paramétrées – Bases réduites – Approximation de faible rang de tenseur – Préconditionneur – Quantité d'intérêt

Abstract

Model order reduction has become an inescapable tool for the solution of high dimensional parameter-dependent equations arising in uncertainty quantification, optimization or inverse problems. In this thesis we focus on low rank approximation methods, in particular on reduced basis methods and on tensor approximation methods.

The approximation obtained by Galerkin projections may be inaccurate when the operator is ill-conditioned. For projection based methods, we propose preconditioners built by interpolation of the operator inverse. We rely on randomized linear algebra for the efficient computation of these preconditioners. Adaptive interpolation strategies are proposed in order to improve either the error estimates or the projection onto reduced spaces. For tensor approximation methods, we propose a minimal residual formulation with ideal residual norms. The proposed algorithm, which can be interpreted as a gradient algorithm with an implicit preconditioner, allows obtaining a quasi-optimal approximation of the solution.

Finally, we address the problem of the approximation of vector-valued or functional-valued quantities of interest. For this purpose we generalize the 'primal-dual' approaches to the non-scalar case, and we propose new methods for the projection onto reduced spaces. In the context of tensor approximation we consider a norm which depends on the error on the quantity of interest. This allows obtaining approximations of the solution that take into account the objective of the numerical simulation.

Keywords: Model order reduction – Uncertainty quantification – Parameter dependent equations – Reduced Basis – Low rank tensor approximation – Preconditioner – Quantity of interest

Contents

Contents i

1 Introduction to model order reduction for parameter-dependent equations **1**

1 Context and contributions . . . 3
1.1 Parameter-dependent equations . . . 3
1.2 Functional approximation of the solution map . . . 4
1.3 Low-rank approximation of the solution map . . . 6
1.4 Contributions and organization of the manuscript . . . 7
2 Low-rank methods: a subspace point of view . . . 9
2.1 Projection on a reduced space . . . 10
2.2 Proper orthogonal decomposition . . . 11
2.3 Reduced basis method . . . 12
3 Low-rank methods based on tensor approximation . . . 15
3.1 Parameter-dependent equation as a tensor structured equation 16
3.2 Low-rank tensor formats . . . 18
3.3 Approximation in low-rank tensor format . . . 20

2 Interpolation of inverse operators for preconditioning parameter-dependent equations **29**

1 Introduction . . . 31
2 Interpolation of the inverse of a parameter-dependent matrix using Frobenius norm projection . . . 33
2.1 Projection using Frobenius norm . . . 33
2.2 Projection using a Frobenius semi-norm . . . 35
2.3 Ensuring the invertibility of the preconditioner for positive definite matrix . . . 43
2.4 Practical computation of the projection . . . 45
3 Preconditioners for projection-based model reduction . . . 46

3.1 Projection of the solution on a given reduced subspace 47
3.2 Greedy construction of the solution reduced subspace 49
4 Selection of the interpolation points 52
4.1 Greedy approximation of the inverse of a parameter-dependent matrix 52
4.2 Selection of points for improving the projection on a reduced space 55
4.3 Recycling factorizations of operator's evaluations - Application to reduced basis method 55
5 Numerical results 56
5.1 Illustration on a one parameter-dependent model 56
5.2 Multi-parameter-dependent equation 61
6 Conclusion 69

3 Projection-based model order reduction for estimating vector-valued variables of interest 73
1 Introduction 75
2 Analysis of different projection methods for the estimation of a variable of interest 76
2.1 Petrov-Galerkin projection 76
2.2 Primal-dual approach 79
2.3 Saddle point problem 82
3 Goal-oriented projections for parameter-dependent equations 86
3.1 Error estimates for vector-valued variables of interest 87
3.2 Greedy construction of the approximation spaces 89
4 Numerical results 91
4.1 Applications 91
4.2 Comparison of the projection methods 94
4.3 Greedy construction of the reduced spaces 100
5 Conclusion 104

4 Ideal minimal residual formulation for tensor approximation 107
1 Introduction 109
2 Functional framework for weakly coercive problems 112
2.1 Notations 112
2.2 Weakly coercive problems 112
3 Approximation in low-rank tensor subsets 113
3.1 Hilbert tensor spaces 113

3.2 Classical low-rank tensor subsets . . . 114
3.3 Best approximation in tensor subsets . . . 115
4 Minimal residual based approximation . . . 116
4.1 Best approximation with respect to residual norms . . . 116
4.2 Ideal choice of the residual norm . . . 117
4.3 Gradient-type algorithm . . . 118
5 Perturbation of the ideal approximation . . . 120
5.1 Approximation of the ideal approach . . . 120
5.2 Quasi-optimal approximations in $\mathcal{M}_r(X)$ . . . 121
5.3 Perturbed gradient-type algorithm . . . 121
5.4 Error indicator . . . 123
6 Computational aspects . . . 124
6.1 Best approximation in tensor subsets . . . 124
6.2 Construction of an approximation of $\Lambda^\delta(r)$ . . . 125
6.3 Summary of the algorithm . . . 129
7 Greedy algorithm . . . 129
7.1 A weak greedy algorithm . . . 130
7.2 Convergence analysis . . . 131
8 Numerical example . . . 135
8.1 Stochastic reaction-advection-diffusion problem . . . 135
8.2 Comparison of minimal residual methods . . . 136
8.3 Properties of the algorithms . . . 141
8.4 Higher dimensional case . . . 144
9 Conclusion . . . 148

5 Goal-oriented low-rank approximation for the estimation of vector-valued quantities of interest 151
1 Introduction . . . 153
2 Choice of norms . . . 154
2.1 Natural norms . . . 154
2.2 Goal-oriented norm . . . 156
3 Algorithms for goal-oriented low-rank approximations . . . 157
3.1 Iterative solver with goal-oriented truncations . . . 159
3.2 A method based on an ideal minimal residual formulation . . . 160
4 Application to uncertainty quantification . . . 162
4.1 Linear quantities of interest . . . 163
4.2 Properties of the norms . . . 164
4.3 Approximation of $u(\xi)$ by interpolation . . . 166

5 Numerical experiments 168
5.1 Iterative solver (PCG) with truncations 168
5.2 Ideal minimal residual formulation 172
6 Conclusion 177
7 Appendix: practical implementation of the approximation operators . 178

Bibliography **183**

Chapter 1

Introduction to model order reduction for parameter-dependent equations

Contents

1 Context and contributions 3

1.1 Parameter-dependent equations 3

1.2 Functional approximation of the solution map 4

1.3 Low-rank approximation of the solution map 6

1.4 Contributions and organization of the manuscript 7

2 Low-rank methods: a subspace point of view 9

2.1 Projection on a reduced space 10

2.2 Proper orthogonal decomposition 11

2.3 Reduced basis method 12

3 Low-rank methods based on tensor approximation 15

3.1 Parameter-dependent equation as a tensor structured equation 16

3.2 Low-rank tensor formats 18

3.3 Approximation in low-rank tensor format 20

1 Context and contributions

1.1 Parameter-dependent equations

Over the past decades, parameter-dependent equations have received a growing interest in different branches of science and engineering. These equations are typically used for the numerical simulation of physical phenomena governed by partial differential equations (PDEs) with different configurations of the material properties, the shape of the domain, the source terms, the boundary conditions etc. The parameter refers to these data that may vary. Various domains of application involve parameter-dependent equations. For example in *optimization* or in *control*, we search the parameter value that minimizes some cost function which is defined as a function of the solution. *Real-time simulation* requires the solution for different values of the parameter in a limited computational time. In *uncertainty quantification*, the parameter is considered as a random variable (which reflects uncertainties on the input data), and the goal is either to study the statistical properties of the solution (forward problem), or to identify the distribution law of the parameter from the knowledge of some observations of the solution (inverse problem).

Let us consider a generic parameter-dependent equation

$$A(\xi)u(\xi) = b(\xi), \tag{1.1}$$

where the parameter $\xi = (\xi_1, \ldots, \xi_d)$ is assumed to take values in a parameter set $\Xi \subset \mathbb{R}^d$ which represents the range of variations of ξ. In the present work, we mainly focus on parameter-dependent linear PDEs. The solution $u(\xi)$ belongs to a Hilbert space V (typically a Sobolev space) endowed with a norm $\|\cdot\|_V$. $A(\xi)$ is a linear operator defined from V to V', the dual space of V', and $b(\xi) \in V'$. The function $u : \xi \mapsto u(\xi)$ defined from Ξ to V is called the *solution map*. When considering the numerical solution of a PDE, a discretization scheme (such as the Finite Element Method [52]) yields a finite dimensional problem of size n, which can also be written under the form (1.1) with V either an approximation space $V = V^h$ of dimension n, or an algebraic space $V = \mathbb{R}^n$. In the latter case, $A(\xi)$ is a matrix of size n. In the rest of this introductory chapter, we consider for the sake of simplicity that V is a finite dimensional space.

When considering complex physical models, solving (1.1) for one value of the parameter requires a call to an expensive numerical solver (*e.g.* for the numerical solution of a PDE with a finite discretization, we have $n \gg 1$). This is a limit-

ing factor for the applications that require the solution $u(\xi)$ for many values of the parameter. In this context, different methods have been proposed for the approximation of the solution map. The goal of these methods, often referred as *Model Order Reduction* (MOR) methods, is to build an approximation of the solution map u that can be rapidly evaluated for any parameter value, and which is sufficiently accurate for the intended applications. This approximation is used as a surrogate for the solution map.

In the following sections, we present MOR methods based on functional approximations, and then based on low-rank approximations. Finally, the contributions of the thesis will be outlined.

1.2 Functional approximation of the solution map

Standard approximation methods construct an approximation u_N of u on a basis $\{\psi_1, \dots, \psi_N\}$ of parameter-dependent functions:

$$u_N(\xi) = \sum_{i=1}^{N} v_i \psi_i(\xi). \tag{1.2}$$

When the basis $\{\psi_1, \dots, \psi_N\}$ is fixed *a priori*, the approximation u_N linearly depends on the coefficients $v_1, \dots, v_N$: this is a *linear approximation method*. Different possibilities have been proposed for the computation of u_N. For example, one can rely on interpolation (also called *stochastic collocation* when the parameter is a random variable) [5, 10, 126], on Galerkin projection (also called *stochastic Galerkin projection*) [62, 97], or on regression [15, 19]. In the seminal work of Ghanem and Spanos [62] polynomial functions were used for the basis $\{\psi_1, \dots, \psi_N\}$. Since then, various types of bases have been considered such as piecewise polynomials [6,46,124], wavelets [93] etc. However, the main difficulty of these approaches is that the number of basis functions N dramatically increases with respect to the dimension d, *i.e.* the number of parameters. For example, when using polynomial spaces with total degree p, $N = (d+p)!/d!p!$. As a consequence the complexity of the approximation methods blows up. Since Bellman [12, 13], the expression *"curse of dimensionality"* refers to such an increase in complexity with respect to the dimension. It is then necessary to exploit particular properties of the solution map for the elaboration of efficient approximation methods. Even if the smoothness of u with respect to ξ is an essential ingredient for its numerical approximation, it is not sufficient to circumvent the curse of dimensionality (see *e.g.* [104] where it is proven that the approximation of the class of infinitely many times differentiable functions with uniformly bounded

derivatives is intractable).

In many situations, the parameters $\xi_1, \ldots, \xi_d$ do not have the same importance in the sense that the solution map may present complex variations with respect to some parameters, and simple variations with respect to the others. This *anisotropy* can be exploited for the construction of a basis that is well adapted to reproduce u, yielding an accurate approximation with a moderate number of basis functions. The idea is for example to put more effort (*e.g.* higher polynomial degree) for the description of the variations with respect to some parameters. Finding an adapted basis is the principal motivation of the *sparse approximation methods*, see [61, 99]. Given a dictionary of functions $\mathcal{D} = \{\psi_\alpha : \alpha \in \Lambda\}$, where Λ denotes a countable set, sparse approximation of u consists in finding a subset $\Lambda_r \subset \Lambda$ of cardinality r such that an element of the form

$$u_{\Lambda_r} = \sum_{\alpha \in \Lambda_r} v_\alpha \psi_\alpha(\xi) \tag{1.3}$$

approximates well the solution map. The problem of finding the best approximation of the form (1.3) is often referred as the *best r-term approximation problem*. Since the basis $\{\psi_\alpha : \alpha \in \Lambda_r\}$ is not chosen *a priori*, this approach is a *non linear approximation method*, see [48, 50]. For instance, provided u is sufficiently smooth (typically if u admits an analytical extension to a complex domain), Proposition 5.5 in [31] states that there exists an approximation u_{Λ_r} of the form (1.2) (built by polynomial interpolation) such that

$$\sup_{\xi \in \Xi} \|u(\xi) - u_{\Lambda_r}(\xi)\|_V \leq C r^{-\rho}, \tag{1.4}$$

where the constants C and ρ depend on the solution map u and on the dimension d. Furthermore, it is shown in [36] that for some particular parameter-dependent equation with infinite dimension $d \to \infty$, exploiting the anisotropy allows to "break" the curse of dimensionality in the sense that the constants C and ρ do not depend on d. However, best r-term approximation problems are known to be combinatorial optimization problems that are NP hard to solve. In practice, the sparse approximation (1.3) can be obtained by using ℓ^1 sparsity-inducing penalization techniques [19, 20, 32], or by selecting the basis functions one after the other in a greedy fashion [37, 38, 109]. We refers to [40] for a detailed introduction and analysis of the methods that exploit the anisotropy of u using sparse approximation techniques.

1.3 Low-rank approximation of the solution map

In this thesis, we focus on *low-rank* approximation methods. The principle is to build an approximation of the solution map of the form

$$u_r(\xi) = \sum_{i=1}^{r} v_i \lambda_i(\xi), \tag{1.5}$$

where the coefficients $v_1, \ldots, v_r$ and the functions $\lambda_1, \ldots, \lambda_r$ are not chosen *a priori*: they are both determined so that u_r approximates well the solution map (in a sense that remains to be defined). Note that low-rank approximations (1.5) are very similar to sparse approximations (1.3): the common feature is that the approximation basis ($\{\psi_i(\xi)\}_{i=1}^r$ or $\{\lambda_i(\xi)\}_{i=1}^r$) is not fixed *a priori*. However, instead of choosing $\lambda_1, \ldots, \lambda_r$ in a dictionary of functions $\mathcal{D}$, they will be selected in a vector space of functions S, typically a Lebesgue space, or an approximation subspace in a Lebesgue space.

We detail now some mathematical aspects of low-rank approximations. Let us introduce the Bochner space $L^p_\mu(\Xi; V) = \{v : \Xi \mapsto V : \|v\|_p < \infty\}$ where the norm $\|\cdot\|_p$ is defined by

$$\begin{aligned} \|v\|_p &= \left(\int_{\xi \in \Xi} \|v(\xi)\|_V^p \, \mathrm{d}\mu(\xi) \right)^{1/p} && \text{if } 1 \leq p < \infty, \\ \|v\|_p &= \operatorname*{ess\,sup}_{\xi \in \Xi} \|v(\xi)\|_V && \text{if } p = \infty. \end{aligned}$$

Here μ a probability measure (the law of the random variable ξ). The Bochner space $L^p_\mu(\Xi; V)$ has a tensor product structure[1] $L^p_\mu(\Xi; V) = V \otimes S$, where S denotes the Lebesgue space $L^p_\mu(\Xi)$. In the following we adopt the notation:

$$X = V \otimes S = L^p_\mu(\Xi; V). \tag{1.6}$$

The elements of X are called *tensors*. The approximation manifold associated to the elements of the form (1.5) is denoted by

$$\mathcal{C}_r(X) = \left\{ \sum_{i=1}^{r} v_i \otimes \lambda_i : v_i \in V, \lambda_i \in S \right\} \subset X.$$

Let us note that any $v \in X$ can be interpreted as a linear operator from V' to S defined by $w \mapsto \langle v, w \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the duality pair. With this point of view,

[1] $V \otimes S = \operatorname{span}\{v \otimes \lambda : v \in V, \lambda \in S\}$, where $v \otimes \lambda$ is called an elementary (or a rank-one) tensor, which can be interpreted as the application from Ξ to V defined by $\xi \mapsto (v \otimes \lambda)(\xi) = v\lambda(\xi)$.

$\mathcal{C}_r(X)$ corresponds to the set operators whose rank (*i.e.* the dimension of the range) is bounded by r. We denote by $d_r^{(p)}(u)$ the lowest possible error we can achieve, with respect to the $L^p_\mu(\Xi; V)$-norm, by approximating u by an element of $\mathcal{C}_r(X)$:

$$d_r^{(p)}(u) = \min_{u_r \in \mathcal{C}_r(X)} \|u - u_r\|_p. \tag{1.7}$$

For the applications where the relation (1.4) holds, we deduce[2] that $d_r^{(\infty)}(u) \leq Cr^{-\rho}$. Since the $L^\infty_\mu(\Xi; V)$-norm is stronger that the $L^p_\mu(\Xi; V)$-norm, we have

$$d_r^{(p)}(u) \leq d_r^{(\infty)}(u)$$

for any $1 \leq p \leq \infty$.

The value of p determines in which sense we want to approximate u. In practice, low-rank approximation methods are either based on $p = \infty$ or on $p = 2$. The choice $p = \infty$ yields an approximation u_r of u which is uniformly accurate over Ξ, meaning $\|u(\xi) - u_r(\xi)\|_V \leq \varepsilon$ for any $\xi \in \Xi$. This is the natural framework for applications such as rare event estimation or optimization. The choice $p = 2$ is natural in the stochastic context when the first moments of the solution (*e.g.* the mean, the variance etc) have to be computed. But in the literature, we observe that the choice of p is generally not driven by the application. For example, approximations built by a method based on $p = 2$ can be used to provide "numerical charts" [34], while approximations built by a method based on $p = \infty$ are used in [23] for computing moments of u. In fact, under some assumptions on the smoothness of the solution map, the approximation error measured with the $L^\infty_\mu(\Xi; V)$-norm can be controlled by the one measured with the $L^2_\mu(\Xi; V)$-norm (see [77]).

1.4 Contributions and organization of the manuscript

This thesis contains different contributions for the two main low-rank approximation methods, namely the Reduced Basis method (which is here presented as a low-rank method with a subspace point of view) and the methods based on tensor approximation techniques. These contributions address the following issues.

- The fact is that the solution map is not known *a priori*, but implicitly defined as the solution of (1.1). As a consequence, low-rank approximation methods are not able to provide optimal approximations, but only quasi-optimal

[2] In some applications we rather observe an exponential rate of convergence $d_r^{(\infty)}(u) \leq C \exp(-cr^\rho)$.

approximations in the best case scenario. This loss of accuracy can be problematic for the efficiency of the methods.

- In many applications, we only need a partial information (called a *quantity of interest*) which is a function of the solution map. The challenge is to develop *goal-oriented* approximation methods which take advantage of such situation in order to reduce the complexity.

The organization of the present manuscript is as follow.

Chapter 2. We observe in practice that a bad condition number of the operator $A(\xi)$ may yield inefficient model order reduction: the use of a preconditioner $P(\xi) \approx A(\xi)^{-1}$ is necessary to obtain accurate approximations. In Chapter 2, we propose a parameter-dependent preconditioner defined as an interpolation of $A(\xi)^{-1}$ (here we consider that equation (1.1) is an algebraic equation, *i.e.* $A(\xi)$ is a matrix). This interpolation is defined by a projection method based on the Frobenius norm. Here we use tools of the *randomized numerical linear algebra* (see [79] for an introduction) for handling large matrices. We propose strategies for the selection of the interpolation points which are dedicated either to the improvement of Galerkin projections or to the estimation of projection errors. Then we show how such preconditioner can be used for projection-based MOR, such as the reduced basis method or the proper orthogonal decomposition method.

Chapter 3. In many situations, one is not interested in the solution $u(\xi)$ itself but rather in some variable of interest defined as a function $\xi \mapsto l(u(\xi))$. In the case where $l \in \mathcal{L}(V, \mathbb{R})$ is a linear function taking scalar values, approximations of the variable of interest can be obtained using a *primal-dual* approach which consists in computing an approximation of the so-called *dual variable*. In chapter 3, we extend this approach to functional-valued or vector-valued variables of interest, *i.e.* $l \in \mathcal{L}(V, Z)$ where Z is a vector space. In particular we develop a new projection method based on a saddle point formulation for the approximation of the primal and dual variables in reduced spaces.

Chapter 4. Low-rank approximations can be obtained through the direct minimization of the residual norm associated to an equation formulated in a tensor product space. In practice, the resulting approximation can be far from being optimal with respect to the norm of interest. In Chapter 4, we introduce an ideal minimal residual formulation such that the optimality of the approximation is achieved with

respect to a desired norm, and in particular the natural norm of $L^2_\mu(\Xi; V)$ for applications to parameter-dependent equations. We propose and analyze an algorithm which provides a quasi-optimal low-rank approximation of the solution map.

Chapter 5. As in Chapter 3, we address the problem of the estimation of a functional-valued (or vector-valued) quantity of interest, which is here defined as $L(u)$ where $L \in \mathcal{L}(X, Z)$. For example when ξ is a random variable, $L(u)$ can be the conditional expectation $\mathbb{E}(u(\xi)|\xi_\tau)$ with respect to a subset of random variables ξ_τ. For this purpose, we propose an original strategy which relies on a goal-oriented norm, meaning a norm on X that takes into account the quantity of interest we want to compute. However, the best approximation problem with respect to this norm can not be solved directly. We propose an algorithm that relies on the ideal minimal residual formulation introduced in Chapter 4.

The rest of this chapter gives the basic notions of low-rank approximation methods which will be useful for the rest of the manuscript. We distinguish here the low-rank methods with a subspace point of view, such as the Reduced Basis method or the Proper Orthogonal Decomposition method, and the methods based on tensor approximation techniques.

2 Low-rank methods: a subspace point of view

Let us note that the subset of tensors with rank bounded by r can be equivalently defined by

$$\mathcal{C}_r(X) = \{V_r \otimes S : V_r \subset V, \dim(V_r) = r\},$$

so that we have

$$\min_{u_r \in \mathcal{C}_r(X)} \|u - u_r\|_p = \min_{\substack{V_r \subset V \\ \dim(V_r)=r}} \min_{u_r \in V_r \otimes S} \|u - u_r\|_p.$$

This alternative formulation of the low-rank approximation problem suggests to find a subspace $V_r \subset V$ of dimension r, often called the *reduced space*, that minimizes $\min_{u_r \in V_r \otimes S} \|u - u_r\|_p$. In practice a subspace can be constructed based on the knowledge of snapshots of the solution $\{u(\xi^1), u(\xi^2), \ldots\}$. This is the basic idea of the *Proper Orthogonal Decomposition* (POD) and of the *Reduced Basis* (RB) methods. The main difference is that the POD method aims at constructing a subspace V_r that is optimal with respect to the $L^2_\mu(\Xi; V)$-norm ($p = 2$), whereas the RB method

tries to achieve the optimality with respect to the $L^\infty_\mu(\Xi; V)$-norm ($p = \infty$). For a given reduced space V_r and for any parameter value $\xi \in \Xi$, the approximation $u_r(\xi)$ can be computed by a Galerkin-type projection (*i.e.* using equation (1.1)) of the solution $u(\xi)$ onto V_r. The term *projection-based model order reduction* is used for such strategies relying on a projection of the solution onto a reduced space. Note that we do not introduce any approximation space for S.

2.1 Projection on a reduced space

In this sub-section, we assume that we are given a reduced space $V_r \subset V$. The best approximation problem in $V_r \otimes S$ is:

$$\min_{v_r \in V_r \otimes S} \|u - v_r\|_p = \|u - \Pi_{V_r} u\|_p,$$

where Π_{V_r} denotes the V-orthogonal projector onto V_r. For the computation of $(\Pi_{V_r} u)(\xi) = \Pi_{V_r} u(\xi)$, we need to know the solution $u(\xi)$. To avoid this, we define $u_r(\xi)$ as the Galerkin projection of $u(\xi)$ onto the reduced space V_r. Here we assume that the operator $A(\xi)$ satisfies

$$\alpha(\xi) = \inf_{v \in V} \frac{\langle A(\xi)v, v\rangle}{\|v\|_V^2} > 0 \quad \text{and} \quad \beta(\xi) = \sup_{v \in V} \sup_{w \in V} \frac{\langle A(\xi)v, w\rangle}{\|v\|_V \|w\|_V} < \infty. \tag{1.8}$$

The Galerkin projection $u_r(\xi) \in V_r$ is characterized by

$$\langle A(\xi)u_r(\xi), v\rangle = \langle b(\xi), v\rangle \tag{1.9}$$

for all $v \in V_r$. Computing $u_r(\xi)$ requires the solution of a small[3] linear system of size r called the *reduced system*. Céa's lemma provides the following quasi-optimality result:

$$\|u(\xi) - u_r(\xi)\|_X \leq \kappa(\xi) \|u(\xi) - \Pi_{V_r} u(\xi)\|_V, \tag{1.10}$$

where $\kappa(\xi) = \beta(\xi)/\alpha(\xi) \geq 1$ is the condition number of $A(\xi)$. Then we have

$$\|u - u_r\|_p \leq \overline{\kappa} \min_{v_r \in V_r \otimes S} \|u - v_r\|_p$$

with $\overline{\kappa} = \sup_{\xi \in \Xi} \kappa(\xi)$. We note that a bad condition number for $A(\xi)$ can lead to an inaccurate Galerkin projection. One can find in [42] a Petrov-Galerkin projection method that aims at defining a better projection onto the reduced space by constructing a suitable test space.

[3] By small, we mean that $r \ll n$, where we recall that n denotes the dimension of the "full" problem (1.1).

In order to have a rapid evaluation of $u_r(\xi)$ for any parameter value $\xi \in \Xi$, the complexity for solving (1.9) should be independent of n (*i.e.* the complexity of the original problem). To obtain this property, a key ingredient is that both the operator $A(\xi)$ and the right hand side $b(\xi)$ admit an affine decomposition with respect to the parameter ξ, meaning that

$$A(\xi) = \sum_{k=1}^{m_A} A_k \Phi_k^A(\xi) \quad , \quad b(\xi) = \sum_{k=1}^{m_b} b_k \Phi_k^b(\xi), \tag{1.11}$$

where Φ_k^A and Φ_k^b are scalar valued functions. Such decomposition allows to precompute the reduced operators (resp. right hand sides) associated to A_k (resp. to b_k) during the so called *Offline phase*. Then for any $\xi \in \Xi$ the reduced system can be rapidly reassembled (using (1.11)) and solved during the *Online phase*, with a complexity that is independent of n. If $A(\xi)$ and $b(\xi)$ do not have an affine decomposition, one can use techniques such as the Empirical Interpolation Method [9] to build approximations of $A(\xi)$ and $b(\xi)$ of the form (1.11). Moreover, if such decompositions exist but the operators A_k or the vectors b_k are not available (for example in a non-intrusive setting), one can use the technique proposed in [29] for the computation of affine decompositions from evaluations of $A(\xi)$ and $b(\xi)$.

2.2 Proper orthogonal decomposition

We present now the principle of the POD. This method was first introduced for the analysis of turbulent flows in fluid mechanics [14]. We refer to [84] for an introduction in the context of parameter-dependent equations.

We assume that the solution map u belongs to $L^2_\mu(\Xi; V)$ ($p = 2$), and we consider its singular value decomposition[4] $u = \sum_{i=1}^{\infty} \sigma_i v_i \otimes \lambda_i$, where $\sigma_i \in \mathbb{R}$ are the singular values (sorted in decreasing order) and $v_i \in V$, $\lambda_i \in S$ are the corresponding left and right singular vectors. This decomposition is also called the Karhunen-Loève expansion when ξ is a random variable. An important feature of this decomposition is that the truncation to its first r terms gives an optimal rank-r approximation of u with respect to the $L^2_\mu(\Xi; V)$-norm:

$$\|u - \sum_{i=1}^{r} \sigma_i v_i \otimes \lambda_i\|_2 = \min_{v_r \in \mathcal{C}_r(X)} \|u - v_r\|_2. \tag{1.12}$$

As a consequence, the optimal reduced space V_r is given by the span of the dominant left singular vectors $\{v_1, \ldots, v_r\}$. The idea of the POD is to approach the $\|\cdot\|_2$-norm

[4]Note that since $\dim(V) < \infty$, the sum is finite.

of equation (1.12) using the Monte Carlo method for the estimation of the integral over Ξ:

$$\|u - v_r\|_2^2 = \int_\Xi \|u(\xi) - v_r(\xi)\|_V^2 \mathrm{d}\mu(\xi) \approx \frac{1}{K} \sum_{k=1}^{K} \|u(\xi^k) - v_r(\xi^k)\|_V^2 \tag{1.13}$$

where $\xi^1, \ldots, \xi^K$ are K independent realizations of a random variable whose probability law is μ. A reduced space V_r is then defined as the span of the first r left singular vectors of the operator

$$\lambda \in \mathbb{R}^K \mapsto \frac{1}{K} \sum_{k=1}^{K} u(\xi^k)\lambda_k \in V.$$

When $V = \mathbb{R}^n$, this operator corresponds to a matrix whose columns are the snapshots $u(\xi^1), \ldots, u(\xi^K)$. Then, efficient algorithms are available for the computation of the complete SVD [69], or for the computation of the truncated SVD [68].

Remark 2.1 (Goal-oriented POD). *Alternative constructions of the reduced space have been proposed in order to obtain accurate approximations of some variable of interest defined by $l(u(\xi))$, where $l \in \mathcal{L}(V, Z)$ is a linear function taking values in a vector space $Z = \mathbb{R}^m$. The idea proposed in [28] is to replace the norm $\|\cdot\|_V$ by a semi-norm $\|l(\cdot)\|_Z$ in the expression (1.13), so that the optimality of the reduced space is achieved with respect to the quantity of interest we want to compute. However such strategy does not take into account the projection problem on the resulting reduced space, and there is no guaranty that such a strategy improves the approximation of the quantity of interest. We believe that the computation of a* goal-oriented projection *of $u(\xi)$ is as important as the computation of a* goal-oriented reduced space *(see Chapter 3).*

2.3 Reduced basis method

The motivation of the RB method is to build a reduced space which provides a controlled approximation of the solution map with respect to the $L^\infty_\mu(\Xi; V)$-norm ($p = \infty$). The lowest error $d_r^{(\infty)}(u)$ for the approximation of u by a rank r element

satisfies

$$\begin{aligned} d_r^{(\infty)}(u) &= \min_{\substack{V_r \subset V \\ \dim(V_r)=r}} \min_{v_r \in V_r \otimes S} \sup_{\xi \in \Xi} \|u(\xi) - v_r(\xi)\|_V, \\ &= \min_{\substack{V_r \subset V \\ \dim(V_r)=r}} \sup_{\xi \in \Xi} \min_{v_r \in V_r} \|u(\xi) - v_r\|_V, \\ &= \min_{\substack{V_r \subset V \\ \dim(V_r)=r}} \sup_{w \in \mathcal{M}} \min_{v_r \in V_r} \|w - v_r\|_V, \end{aligned}$$

where $\mathcal{M} = \{u(\xi) : \xi \in \Xi\} \subset V$ denotes the *solution manifold*. Then $d_r^{(\infty)}(u)$ is the Kolmogorov r-width of $\mathcal{M}$, see [88]. There is no practical algorithm to compute the corresponding optimal subspace, even if we had access to the solution $u(\xi)$ for any $\xi \in \Xi$. The idea of the RB method is to construct V_r as the span of snapshots of the solution. Contrarily to the POD method which relies on a crude Monte Carlo sampling of the solution manifold, the RB method selects the evaluation points adaptively.

In the seminal work [123], a Greedy algorithm has been proposed for the construction of the approximation space

$$V_{r+1} = V_r + \operatorname{span}\{u(\xi^{r+1})\}.$$

Ideally, ξ^{r+1} should be chosen where the error of the current approximation $u_r(\xi)$ is maximal, that means

$$\xi^{r+1} \in \arg\max_{\xi \in \Xi} \|u(\xi) - u_r(\xi)\|_V. \tag{1.14}$$

Since $u_{r+1}(\xi)$ is defined as the Galerkin projection of $u(\xi)$ onto a subspace V_{r+1} which contains $u(\xi^{r+1})$, and thanks to relation (1.10), we have $\|u(\xi^{r+1}) - u_{r+1}(\xi^{r+1})\|_V = 0$ (in fact, we also have $\|u(\xi) - u_{r+1}(\xi)\|_V = 0$ for any $\xi \in \{\xi^1, \dots, \xi^{r+1}\}$, so that $u_{r+1}(\xi)$ is an interpolation of $u(\xi)$ on the set of points $\{\xi^1, \dots, \xi^{r+1}\}$). We understand that this approach aims at decreasing the $L^\infty_\mu(\Xi; V)$ error. But in practice, the selection strategy (1.14) is unfeasible since it requires the solution $u(\xi)$ for all $\xi \in \Xi$. To overcome such an issue, the exact error $\|u(\xi) - u_r(\xi)\|_V$ is replaced by an estimator $\Delta_r(\xi)$ that can be computed for any $\xi \in \Xi$ with low computational costs. This estimator is said to be tight if there exist two constants $c > 0$ and $C < \infty$ such that

$$c\Delta_r(\xi) \leq \|u(\xi) - u_r(\xi)\|_V \leq C\Delta_r(\xi)$$

holds for any $\xi \in \Xi$. A popular error estimator is the dual norm of the residual $\Delta_r(\xi) = \|A(\xi)u_r(\xi) - b(\xi)\|_{V'}$. In that case, we have $c = \inf_\xi \beta(\xi)^{-1}$ and $C = \sup_\xi \alpha(\xi)^{-1}$ where $\alpha(\xi)$ and $\beta(\xi)$ the constants defined in (1.8).

Remark 2.2 (Training set). *In practice, Ξ is replaced by a training set Ξ_{train} with finite cardinality, but large enough to represent well Ξ in order not to miss relevant parts of the parameter domain.*

This Greedy algorithm has been analyzed in several papers [18, 25, 49, 95]. In particular, Corollary 3.3 in [49] states that if the Kolmogorov r-width satisfies $d_r^{(\infty)}(u) \leq C_0 r^{-\rho}$ for some $C_0 > 0$ and $\rho > 0$, then the resulting approximation space V_r satisfies

$$\min_{v_r \in V_r \otimes S} \|u - v_r\|_\infty \leq \gamma^{-2} C_1 r^{-\rho} \tag{1.15}$$

with $C_1 = 2^{5\rho+1} C_0$ and $\gamma = c/C \leq 1$. Moreover, if $d_r^{(\infty)}(u) = C_0 \exp(-c_0 r^\rho)$ for some positive constants C_0, c_0 and ρ, then

$$\min_{v_r \in V_r \otimes S} \|u - v_r\|_\infty \leq \gamma^{-1} C_1 \exp(-c_1 r^\rho) \tag{1.16}$$

with $C_1 = \sqrt{2C_0}$ and $c_1 = 2^{-1-2\rho} c_0$. Qualitatively, these results tell us that the RB method is particularly interesting in the sense that it preserves (in some situations) the convergence rate of the Kolmogorov r-width. But from a quantitative point of view, a constant γ close to zero may deteriorate the quality of V_r for small r. Let us note that when the error estimator is the dual norm of the residual (*i.e.* $\Delta_r(\xi) = \|A(\xi)u_r(\xi) - b(\xi)\|_{V'}$), we have

$$\gamma^{-1} = \frac{C}{c} = \frac{\sup_\xi \beta(\xi)}{\inf_\xi \alpha(\xi)} \geq \sup_\xi \frac{\beta(\xi)}{\alpha(\xi)} =: \overline{\kappa}.$$

Here again, the condition number of $A(\xi)$ plays an important role on the quality of the approximation space V_r. If it is large, then γ^{-1} will be also large. Furthermore, note that we have

$$\gamma^{-1} = \frac{\sup_\xi(\beta(\xi)\alpha(\xi)/\alpha(\xi))}{\inf_\xi \alpha(\xi)} \leq \frac{\sup_\xi \alpha(\xi)}{\inf_\xi \alpha(\xi)} \overline{\kappa}.$$

Even with an ideal condition number $\overline{\kappa} = 1$, we have no guaranty that the constant γ^{-1} will be close to one (take for example $A(\xi) = \xi I$, where I is the identity operator and $0 < \varepsilon \leq \xi \leq 1$. Then $\kappa(\xi) = 1$ and $\gamma^{-1} = \varepsilon^{-1}$).

Remark 2.3 (Certified RB). *In order to have a certified error bound, one can use the error estimator $\Delta_r(\xi) = \alpha(\xi)^{-1}\|A(\xi)u_r(\xi) - b(\xi)\|_{V'}$ which provides an upper bound of the error: $\|u(\xi) - u_r(\xi)\|_V \leq \Delta_r(\xi)$. This corresponds to the Certified Reduced Basis method [114, 123]. Then we have $C = 1$, $c = \inf_\xi \alpha(\xi)/\beta(\xi)$ and finally $\gamma^{-1} = \overline{\kappa}$. Except for the particular case where $\alpha(\xi)$ is given for free, one can build a lower bound $\alpha_{LB}(\xi) \leq \alpha(\xi)$ using for example the Suc-*

cessive Constraint linear optimization Method (SCM), see [83]. Then for $\Delta_r(\xi) = \alpha_{LB}(\xi)^{-1}\|A(\xi)u_r(\xi) - b(\xi)\|_{V'}$*, the relation* $\|u(\xi) - u_r(\xi)\|_V \leq \Delta_r(\xi)$ *still holds and we obtain*

$$\overline{\kappa} \leq \gamma^{-1} \leq \frac{\sup_\xi \alpha(\xi)/\alpha_{LB}(\xi)}{\inf_\xi \alpha(\xi)/\alpha_{LB}(\xi)}\overline{\kappa}.$$

Remark 2.4 (Goal-oriented RB). *For some applications, one is interested in some* "region of interest" *in the parameter domain* Ξ*. For example in optimization, we need an accurate approximation of* $u(\xi)$ *around the minimizer of some cost function* $\xi \mapsto J(u(\xi))$*. Another example can be found in rare events estimation, where an accurate approximation is needed only in the border of some* failure domain *defined by* $\{\xi \in \Xi : l(u(\xi)) > 0\}$*. The idea proposed in [30] consists in using a weighted error estimate* $w_r(\xi)\Delta_r(\xi)$*, where* w_r *is a positive function that assigns more importance of the error associated to some region of the parameter domain. For rare event estimation, this function can be defined as* $w_r(\xi) = 1/|l(u_r(\xi))|$*.*

3 Low-rank methods based on tensor approximation

In this section we show how a low-rank approximation of the solution map can be obtained using tensor approximation methods. Here, the principle is to interpret u as the solution of a linear equation

$$\mathcal{A}u = \mathcal{B}, \tag{1.17}$$

which is formulated on the tensor product space $X = V \otimes S$ endowed with the $L^2_\mu(\Xi; V)$-norm ($p = 2$). Contrarily to the subspace point of view presented in Section 2, we build here an explicit representation of u_r, meaning that the functions $\lambda_1, \ldots, \lambda_r$ will be explicitly computed. In practice this requires to introduce an approximation space for S, such as a polynomial space, a piecewise polynomial space etc, or to work on a sample set (meaning a set $\Xi_{train} \subset \Xi$ of finite cardinality). But, as mentioned in section 1.2, the dimension of such spaces blows up with the number d of parameter. In order to avoid the construction of an approximation space for *multivariate* functions, we can rather consider u_r of the form

$$u_r(\xi) = \sum_{i=1}^{r} v_i \; \lambda_i^{(1)}(\xi_1) \ldots \lambda_i^{(d)}(\xi_d), \tag{1.18}$$

which is a so-called *separated representation*. This requires to introduce approximation spaces only for *univariate* functions. Here, (1.18) is an approximation in the canonical tensor format with a canonical rank bounded by r, which is known to have bad topological properties. Alternative low-rank formats have been proposed, such as the Tucker format, the Hierarchical Tucker format or the Tensor Train format.

In the rest of this section, we first show how the parameter-dependent equation (1.1) can be written as a tensor structured equation (1.17). Then we introduce standard low-rank formats, and we present different methods for the solution of (1.17) using low-rank tensor techniques.

3.1 Parameter-dependent equation as a tensor structured equation

The weak formulation of the parameter-dependent equation (1.1) consists in finding $u \in X$ such that

$$\int_\Xi \langle A(\xi)u(\xi), w(\xi)\rangle \, \mathrm{d}\mu(\xi) = \int_\Xi \langle b(\xi), w(\xi)\rangle \, \mathrm{d}\mu(\xi) \tag{1.19}$$

for any $w \in X$. We introduce the operator $\mathcal{A} : X \to X'$ and $\mathcal{B} \in X'$ defined by:

$$\langle \mathcal{A}v, w\rangle = \int_\Xi \langle A(\xi)v(\xi), w(\xi)\rangle \, \mathrm{d}\mu(\xi) \quad \text{and} \quad \langle \mathcal{B}, w\rangle = \int_\Xi \langle b(\xi), w(\xi)\rangle \, \mathrm{d}\mu(\xi),$$

for any $v, w \in X$, so that (1.19) can be equivalently written as in equation (1.17). Here, we endow X with the norm $\|\cdot\|_X = \|\cdot\|_2$. Then X is a Hilbert space, which is convenient in the present context. A sufficient condition for problem (1.17) to be well-posed is

$$\underline{\alpha} = \inf_{\xi\in\Xi} \alpha(\xi) > 0 \quad \text{and} \quad \overline{\beta} = \sup_{\xi\in\Xi} \beta(\xi) < \infty,$$

where $\alpha(\xi)$ and $\beta(\xi)$ are defined by (1.8). Indeed, this implies

$$\langle \mathcal{A}v, v\rangle \;\geq\; \int_\Xi \alpha(\xi)\|v(\xi)\|_V^2 \, \mathrm{d}\mu(\xi) \;\geq\; \underline{\alpha} \int_\Xi \|v(\xi)\|_V^2 \, \mathrm{d}\mu(\xi) = \underline{\alpha}\|v\|_X^2 \tag{1.20}$$

for any $v \in X$, and

$$\langle \mathcal{A}v, w\rangle \leq \int_\Xi \beta(\xi)\|v(\xi)\|_V \|w(\xi)\|_V \, \mathrm{d}\mu(\xi) \;\leq\; \overline{\beta}\|v\|_X\|w\|_X \tag{1.21}$$

for any $v, w \in X$. Equations (1.20) and (1.21) ensure respectively the coercivity and the continuity of $\mathcal{A}$ on X. Thanks to Lax-Milgram theorem, (1.17) is then well

posed. Furthermore, the weak formulation (1.19) (or equivalently (1.17)) is convenient to introduce an approximation space $\widetilde{X}$ in X (for example $\widetilde{X} = V \otimes \widetilde{S} \subset X$, with $\widetilde{S} \subset S$ a polynomial space). Replacing X by $\widetilde{X}$ in (1.19), the resulting solution $\widetilde{u}$ is the Galerkin approximation of u onto $\widetilde{X}$ (sometimes called the *Spectral Stochastic Galerkin* approximation). In the following, we continue working in the space $X = V \otimes S$, although it can be replaced by $\widetilde{X}$ at any time.

As mentioned earlier, we also want to introduce low-rank approximations of the form (1.18). This is made possible by the fact that, under some assumptions, the space S also admits a tensor product structure. Let us assume that Ξ admits a product structure $\Xi = \Xi_1 \times \ldots \times \Xi_d$, where $\Xi_\nu \subset \mathbb{R}$ for all $\nu \in \{1, \ldots, d\}$, and that the measure μ satisfies $\mu(\xi) = \prod_{\nu=1}^d \mu^{(\nu)}(\xi_\nu)$ for any $\xi \in \Xi$ (if μ is the probability law of $\xi = (\xi_1, \ldots, \xi_d)$, that means that the random variables $\xi_1, \ldots, \xi_d$ are mutually independent). Then the space S has the following tensor product structure[5] $S = S_1 \otimes \ldots \otimes S_d$, where $S_\nu = L^2_{\mu^{(\nu)}}(\Xi_\nu)$.

We discuss now the tensor structure of the operator $\mathcal{A}$ and the right-hand side $\mathcal{B}$. We assume here that $A(\xi)$ and $b(\xi)$ admit the following affine decompositions:

$$A(\xi) = \sum_{k=1}^{m_A} A_k \prod_{\nu=1}^{d} \Phi^A_{k,\nu}(\xi_\nu) \quad \text{and} \quad b(\xi) = \sum_{k=1}^{m_b} b_k \prod_{\nu=1}^{d} \Phi^b_{k,\nu}(\xi_\nu),$$

where $\Phi^A_{k,\nu}$ and $\Phi^b_{k,\nu}$ are scalar valued functions defined over Ξ_ν. This decomposition is similar to the previous one (1.11), but with an additional assumption for the functions Φ^A_k and Φ^b_k. Then $\mathcal{A}$ and $\mathcal{B}$ admit the following decompositions

$$\mathcal{A} = \sum_{k=1}^{m_A} A_k \otimes A_k^{(1)} \otimes \ldots \otimes A_k^{(d)} \quad \text{and} \quad \mathcal{B} = \sum_{k=1}^{m_b} b_k \otimes b_k^{(1)} \otimes \ldots \otimes b_k^{(d)},$$

where $A_k^{(\nu)} : S_\nu \to S'_\nu$ and $b_k^{(\nu)} \in S'_\nu$ are defined by

$$\langle A_k^{(\nu)} \lambda, \gamma \rangle = \int_{\Xi_\nu} \Phi^A_{k,\nu}(\xi_\nu) \lambda(\xi_\nu) \gamma(\xi_\nu) \, \mathrm{d}\mu^{(\nu)}(\xi_\nu),$$
$$\langle b_k^{(\nu)}, \gamma \rangle = \int_{\Xi_\nu} \Phi^b_{k,\nu}(\xi_\nu) \gamma(\xi_\nu) \, \mathrm{d}\mu^{(\nu)}(\xi_\nu),$$

for all $\lambda, \gamma \in S_\nu$.

[5]Here again, S_ν can be replaced by an approximation space $\widetilde{S}_\nu \subset S_\nu$. In that case, the approximation space $\widetilde{S} = \widetilde{S}_1 \otimes \ldots \otimes \widetilde{S}_d$ preserves the tensor product structure of S.

Thereafter, and for the sake of simplicity, we adopt the notation

$$X = X_1 \otimes \ldots \otimes X_d,$$

where $X_1 = V$, $X_2 = L^2_{\mu^{(1)}}(\Xi_1)$, $X_3 = L^2_{\mu^{(2)}}(\Xi_2)$ and so on. Let us note that with this new notation, we have replaced $d+1$ by d. Then, $\mathcal{A}$ and $\mathcal{B}$ are interpreted as tensors:

$$\mathcal{A} = \sum_{k=1}^{m_{\mathcal{A}}} \bigotimes_{\nu=1}^{d} \mathcal{A}_k^{(\nu)} \quad \text{and} \quad \mathcal{B} = \sum_{k=1}^{m_{\mathcal{B}}} \bigotimes_{\nu=1}^{d} \mathcal{B}_k^{(\nu)}. \tag{1.22}$$

3.2 Low-rank tensor formats

We present here different low-rank tensor formats. We refer to the monograph [76] for a detailed presentation. The most simple low-rank tensor format is the *canonical* rank-r tensor format that is defined by

$$\mathcal{C}_r(X) = \left\{ \sum_{i=1}^{r} x_i^{(1)} \otimes \ldots \otimes x_i^{(d)} : x_i^{(\nu)} \in X_\nu \right\}. \tag{1.23}$$

This allows us to define the *canonical rank* of a tensor $x \in X$ as the minimal integer $r \in \mathbb{N}$ such that $x \in \mathcal{C}_r(X)$. We use the notation $r = \mathrm{rank}(x)$. However, $\mathcal{C}_r(X)$ is not a closed subset for $d > 2$ and $r > 1$. As shown in [45, proposition 4.6], for some tensor $x \in \mathcal{C}_3(X)$ such that $x \notin \mathcal{C}_2(X)$, there exists a tensor $y \in \mathcal{C}_2(X)$ that can be arbitrarily closed to x. This is an issue for the elaboration of a robust approximation method in this subset.

Now we introduce other low-rank tensor formats that have better properties. A key ingredient is the notion of *t-rank of a tensor* $x \in X$, see [71, 78]. Here, $t \subset D = \{1, \ldots, d\}$ denotes a subset of indices, and $t^c = D \backslash t$ is the complementary of t in D. We consider $X_t = \bigotimes_{\nu \in t} X_\nu$ and $X_{t^c} = \bigotimes_{\nu \in t^c} X_\nu$, so that any $x \in X \equiv X_t \otimes X_{t^c}$ can be interpreted as a tensor of order 2. This process is called the *matricization* of the tensor with respect to t. Then, we can define the t-rank of a tensor x as the minimal integer $r \in \mathbb{N}$ such that $x \in \mathcal{C}_r(X_t \otimes X_{t^c})$ (which is the unique notion of the rank for a tensor of order 2). We then use the notation $r = \mathrm{rank}_t(x)$ (note that $\mathrm{rank}_t(x) = \mathrm{rank}_{t^c}(x)$). This allows us to introduce the subset of tensors with a Tucker rank bounded by r

$$\mathcal{T}_r(X) = \Big\{ x \in X : \mathrm{rank}_t(x) \leq r_t, t \in D \Big\},$$

where $r = (r_1, \ldots, r_d) \in \mathbb{N}^d$. As a matter of fact, this subset is closed since it is a finite intersection of closed subsets: $\mathcal{T}_r(X) = \bigcap_{t \in D} \{x \in X : \mathrm{rank}_t(x) \leq r_t\}$.

Furthermore, any element of $x_r \in \mathcal{T}_r(X)$ can be written

$$x_r = \sum_{i_1=1}^{r_1} \dots \sum_{i_d=1}^{r_d} a_{i_1,\dots,i_d}\, x_{i_1}^{(1)} \otimes \dots \otimes x_{i_d}^{(d)}, \tag{1.24}$$

where $a \in \mathbb{R}^{r_1 \times \dots \times r_d}$ is called the *core tensor*, and $x_{i_\nu}^{(\nu)} \in X_\nu$ for all $i_\nu \in \{1, \dots, r_\nu\}$ and $\nu \in \{1, \dots, d\}$. The format (1.24) is called the *Tucker format*. However this tensor format suffers from the curse of dimensionality, since the tensor core belongs to a space whose dimension increases exponentially with respect to d: $\dim(\mathbb{R}^{r_1 \times \dots \times r_d}) = \prod_{\nu=1}^d r_\nu$. To avoid this, low-rank structure also have to be imposed on the core tensor a. This can be done by considering a *dimension tree* $T \subset 2^D$ that is a hierarchical partition of set of dimension D. Examples of such trees are given on Figure 1.1 (we refer to Definition 3.1 in [71] for the definition of such dimension trees). The subset of tensors with Hierarchical Tucker rank bounded by $r \in \mathbb{N}^{\#(T)-1}$, first introduced in [78], is defined by

$$\mathcal{H}_r^T(X) = \Big\{ x \in X : \mathrm{rank}_t(x) \le r_t, t \in T \backslash D \Big\}. \tag{1.25}$$

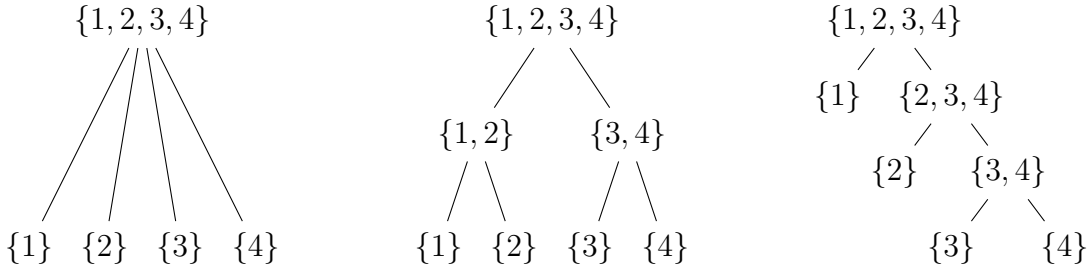
If T is the one-level tree of Figure 1.1(a), then $\mathcal{H}_r^T(X)$ is nothing but $\mathcal{T}_r(X)$. When T is the unbalanced tree of Figure 1.1(c), any tensor $x_r \in \mathcal{H}_r^T(X)$ can be written as in (1.24) with a core a having the following structure:

$$a_{i_1,\dots,i_d} = \sum_{k_1=1}^{r_1^a} \dots \sum_{k_{d-1}=1}^{r_{d-1}^a} a_{i_1,k_1}^{(1)}\, a_{k_1,i_2,k_2}^{(2)} \dots a_{k_{d-2},i_{d-1},k_{d-1}}^{(d-1)}\, a_{k_{d-1},i_d}^{(d)}, \tag{1.26}$$

where we used the notations $r_1^a = \mathrm{rank}_{\{1\}}(x)$, $r_2^a = \mathrm{rank}_{\{1,2\}}(x)$, and so on[6]. Here the tensor core a is a chained product of the tensors $a^{(\nu)} \in \mathcal{S}_\nu = \mathbb{R}^{r_{\nu-1}^a \times r_\nu \times r_\nu^a}$ (with the convention $r_0^a = r_d^a = 1$). The amount of memory for the storage of the core a is $\dim(\times_{\nu=1}^d \mathcal{S}_\nu) = \sum_{\nu=1}^d r_{\nu-1}^a r_\nu r_\nu^a$, which linearly depends on d. Let us mention that for general trees like the balanced tree of Figure 1.1(b), the tensor core possesses also a simple parametrization, see [71, 78] for more information. When considering binary trees (each node has 2 sons), the storage requirement for a also linearly depends on d.

Remark 3.1. *In fact, the trees presented on Figures 1.1(b) and 1.1(c) yield the same low-rank tensor subset $\mathcal{H}_r^T(X)$: they are both associated to subsets of tensors with bounded t-ranks for $t \in \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}\}$. Indeed, for the balanced tree 1.1(b) we have $\mathrm{rank}_{\{1,2\}}(x) = \mathrm{rank}_{\{3,4\}}(x)$ to that we can remove the condi-*

[6] here, we changed the notations by taking the complementary of the interior nodes of the tree 1.1(c): for example $r_2^a = \mathrm{rank}_{\{1,2\}}(x) = \mathrm{rank}_{\{3,\dots,d\}}(x)$.



(a) One-level tree (Tucker format).

(b) Balanced tree (Hierarchical format).

(c) Unbalanced tree (Tensor Train format).

Figure 1.1: Different dimension trees for $d = 4$.

tion on the t-rank for $t = \{3, 4\}$, the same for the unbalanced tree 1.1(c) with $\text{rank}_{\{2,3,4\}}(x) = \text{rank}_{\{1\}}(x)$. This is a particular case that can obviously not be generalized to higher dimensions $d > 4$.

We conclude this section by introducing the Tensor Train format (see [105, 106]). The subset of tensors with TT-rank bounded by $r \in \mathbb{N}^{d-1}$ is defined by

$$\mathcal{TT}_r(X) = \Big\{ x \in X : \text{rank}_t \leq r_t, t \in \big\{\{1\}, \{1, 2\}, \ldots, \{1, \ldots, d-1\}\big\} \Big\}.$$

This format is a Hierarchical Tucker format associated to a dimension tree of the form given in Figure 1.1(c) with inactive constraints on the t-rank for $t \in \{\{2\}, \ldots, \{d-1\}\}$. Any $x_r \in \mathcal{TT}_r(X)$ can be written under the form

$$x_r = \sum_{k_1=1}^{r_1} \cdots \sum_{k_{d-1}=1}^{r_{d-1}} \underbrace{x^{(1)}_{k_1}}_{\in X_1} \otimes \underbrace{x^{(2)}_{k_1,k_2}}_{\in X_2} \otimes \cdots \otimes \underbrace{x^{(d-1)}_{k_{d-2},k_{d-1}}}_{\in X_{d-1}} \otimes \underbrace{x^{(d)}_{k_{d-1}}}_{\in X_d}. \tag{1.27}$$

3.3 Approximation in low-rank tensor format

We discuss now the problem of approximating a tensor $x \in X$ in a low-rank tensor subset $\mathcal{M}_r(X) \in \{\mathcal{C}_r(X), \mathcal{T}_r(X), \mathcal{H}^T_r(X), \mathcal{TT}_r(X)\}$. In this section, we emphases on two situations: (a) for a given rank r, we want to find the best approximation of x in the set $\mathcal{M}_r(X)$, and (b) for a given $\varepsilon > 0$, we want to perform an approximation $x_r \in \mathcal{M}_r(X)$ such that $\|x - x_r\|_X \leq \varepsilon$ with an adapted rank r. Moreover, two cases have to be considered: either the tensor x is known explicitly, or the tensor x is *a priori* unknown but is the solution of the equation (1.17). In this latter case, we use the notation $x = u$ and $x_r = u_r$.

3.3.1 Methods based on singular value decomposition

We show how to address both points (a) and (b) for the approximation of a given tensor x using the Singular Value Decomposition (SVD). A key assumption is that the norm $\|\cdot\|_X$ satisfies the relation

$$\|x^{(1)} \otimes \ldots \otimes x^{(d)}\|_X = \|x^{(1)}\|_{X_1} \ldots \|x^{(d)}\|_{X_d} \tag{1.28}$$

for any $x^{(\nu)} \in X_\nu$, where $\|\cdot\|_{X_\nu}$ denotes the norm of the Hilbert space X_ν. When (1.28) is satisfied, $\|\cdot\|_X$ is called a *cross-norm*, or an *induced norm*. Once again, we begin with the case of order-two tensor, *i.e* $d = 2$. As mentioned in Section 2.2, any $x \in X_1 \otimes X_2$ admits a singular value decomposition $x = \sum_{i=1}^{\infty} \sigma_i \; x_i^{(1)} \otimes x_i^{(2)}$. Eckart Young's theorem states that the truncation to the first r terms of the SVD provides a best approximation of x in $\mathcal{C}_r(X)$:

$$\|x - \sum_{i=1}^{r} \sigma_i \; x_i^{(1)} \otimes x_i^{(2)}\|_X = \min_{y \in \mathcal{C}_r(X)} \|x - y\|_X. \tag{1.29}$$

As a consequence, the point (a) is addressed by computing the first r terms of the SVD of x. Using the notation $x_r = \sum_{i=1}^{r} \sigma_i \; x_i^{(1)} \otimes x_i^{(2)}$, we have $\|x - x_r\|_X = (\sum_{i=r+1}^{\infty} \sigma_i^2)^{1/2}$. Point (b) is addressed by truncating the SVD to the first r terms such that $(\sum_{i=r+1}^{\infty} \sigma_i^2)^{1/2} \leq \varepsilon$.

Remark 3.2 (Nested optimal subspaces). *As already mentioned in Section 2, the best low-rank approximation problem can be written as a subspace optimization problems*

$$\min_{\substack{V_r^{(1)} \subset X_1 \\ dim(V_r^{(1)})=r}} \; \min_{y \in V_r^{(1)} \otimes X_2} \|x - y\|_X \quad \text{or} \quad \min_{\substack{V_r^{(2)} \subset X_2 \\ dim(V_r^{(2)})=r}} \; \min_{y \in X_1 \otimes V_r^{(2)}} \|x - y\|_X.$$

Thanks to relation (1.29), we know that such optimal subspaces exist, and they are given by $V_r^{(\nu)}(x) = span\{x_1^{(\nu)}, \ldots, x_r^{(\nu)}\}$. These optimal subspaces are nested : $V_{r-1}^{(\nu)}(x) \subset V_r^{(\nu)}(x)$.

We now consider the case of higher order tensors ($d \geq 2$). Low-rank approximation based on singular value decomposition has been first proposed by Lathauwer in [92] for the Tucker format. The method, called Higher-Order SVD (HOSVD), consists in projecting x onto the subspace $V_{r_1}^{(1)} \otimes \ldots \otimes V_{r_d}^{(d)} \subset X$, where for all $\nu \in \{1, \ldots, d\}$, $V_{r_\nu}^{(\nu)} \subset X_\nu$ are the optimal subspaces associated to

$$\min_{\substack{V_{r_\nu}^{(\nu)} \subset X_\nu \\ \dim(V_{r_\nu}^{(\nu)})=r_\nu}} \; \min_{y \in V_{r_\nu}^{(\nu)} \otimes X^{\nu c}} \|x - y\|_X, \tag{1.30}$$

where $X^{\nu^c} = \bigotimes_{k\neq\nu} X_k$. In practice, the subspaces $V_{r_\nu}^{(\nu)} \subset X_\nu$ are obtained by computing the truncated SVD (to the first r_ν terms) of $x \in X \equiv X_\nu \otimes X_{\nu^c}$, which is seen as an order-two tensor. Lemma 2.6 in [71] states that the resulting approximation $x_r \in \mathcal{T}_r(X)$ satisfies

$$\|x - x_r\|_X \leq \sqrt{d} \min_{y\in\mathcal{T}_r(X)} \|x - y\|_X.$$

In other words, the HOSVD yields a quasi-optimal approximation in the Tucker format (point (a)). In order to address the point (b), we choose the ranks r_ν such that the approximation error (1.30) is lower than ε for all $\nu \in \{1, \ldots, d\}$. Thus we obtain an approximation $x_r \in \mathcal{T}_r(X)$ such that $\|x - x_r\|_X \leq \varepsilon\sqrt{d}$, see Property 10 in [92].

This methodology can be generalized for the low-rank approximation in the Hierarchical tensor format [71] or for the Tensor Train format [107]. Briefly, the idea is to consider the truncated SVD for all matricizations of $x \in X \equiv X_t \otimes X_{t^c}$ associated to the sets of indices t in a dimension tree. When using the balanced tree of Figure 1.1(b), this yields a quasi-optimal approximation in $\mathcal{H}_r^T(X)$ with a quasi-optimality constant $\sqrt{2d-2}$, see [71, theorem 3.11] (point (a)), and to a quasi-optimality constant $\sqrt{d-1}$ in $\mathcal{TT}_r(X)$. For the point (b), truncating each SVD with the precision ε yields a $x_r \in \mathcal{H}_r^T(X)$ such that $\|x - x_r\|_X \leq \varepsilon\sqrt{2d-2}$, and a $x_r \in \mathcal{TT}_r(X)$ such that $\|x - x_r\|_X \leq \varepsilon\sqrt{d-1}$.

Let us finally note that these methods based on SVD are particularly efficient for the low-rank approximation of a tensor. Although the truncation does not yield optimal approximations for $d > 2$ (point (a)), the quasi-optimality constant grows only moderately with respect to d. For the approximation with respect to a given precision (point (b)), the advantage is that the ranks are chosen adaptively. In particular, the anisotropy in the ranks of the tensor x (if any) is automatically detected and exploited. However, the optimal choice of the low-rank format, or the optimal choice of the tree for the Hierarchical tensor format, remain open questions.

3.3.2 Truncated iterative solvers for the solution of $\mathcal{A}u = \mathcal{B}$

We address now the problem of the low-rank approximation of a tensor $u \in X$ that is not known explicitly, but that is the solution of a tensor structured equation $\mathcal{A}u = \mathcal{B}$, with $\mathcal{A}$ and $\mathcal{B}$ of the form (1.22). It is possible here to use classical iterative solvers (Richardson, CGS, GMRES etc), coupled with the tensor approximation techniques presented in Section 3.3.1. We refer to the review Section 3.1 [72] for

related references.

Let us illustrate the methodology with a simple Richardson method. It consists in constructing the sequence $\{u^k\}_{k\geq 1}$ in X defined by the recurrence relation:

$$u^{k+1} = u^k + \omega\mathcal{P}(\mathcal{B} - \mathcal{A}u^k), \tag{1.31}$$

where $\mathcal{P} : X' \to X$ denotes a preconditioner of $\mathcal{A}$ (meaning $\mathcal{P} \approx \mathcal{A}^{-1}$), and $\omega \in \mathbb{R}$. The sequence $\{u^k\}_{k\geq 1}$ is known to converge as $\mathcal{O}(\rho^k)$ to the solution u, provided $\rho = \|\mathcal{I} - \omega\mathcal{P}\mathcal{A}\|_{X\to X}$ is strictly lower than 1 ($\mathcal{I}$ denotes the identity operator of X and $\|\cdot\|_{X\to X}$ the operator norm). An optimization of the parameter ω gives the optimal rate of convergence $\rho = (\kappa - 1)/(\kappa + 1)$, where κ denotes the condition number[7] of $\mathcal{P}\mathcal{A}$. Therefore, the Richardson iteration method requires a good preconditioner to speed up the convergence. This is also true for other iterative solvers. In the context of parameter-dependent equations, a commonly used preconditioner is $\mathcal{P} = \bigotimes_{\nu=1}^d \mathcal{P}^{(\nu)}$, where $\mathcal{P}^{(1)} = A(\bar{\xi})^{-1}$ for some parameter value $\bar{\xi} \in \Xi$, or $\mathcal{P}^{(1)} = \mathbb{E}(A(\xi))^{-1}$ when ξ is a random variable. For $\nu \geq 2$, $\mathcal{P}^{(\nu)}$ is the identity operator of X_ν. We refer to [66] for a general method for constructing low-rank preconditioners.

Iterative solvers are known to suffer from what is called the *"curse of the rank"*. To illustrate this, we assume that the iterate $u^k \in \mathcal{C}_r(X)$ is stored in the canonical tensor format. Then, the representation rank of u^{k+1} given by (1.31) is $r + m_{\mathcal{P}}(m_{\mathcal{B}} + m_{\mathcal{A}}r)$, which shows that the representation rank blows up during the iteration process. Then the idea is to "compress" each iterate u^{k+1} using for example the low-rank truncation techniques introduced in Section 3.3.1:

$$u^{k+1} = \Pi^{\varepsilon}_{\mathcal{M}_r}\Big(u^k + \omega\mathcal{P}(\mathcal{B} - \mathcal{A}u^k)\Big),$$

where $\Pi^{\varepsilon}_{\mathcal{M}_r}$ denotes an approximation operator which provides an approximation $\Pi^{\varepsilon}_{\mathcal{M}_r}(x)$ in $\mathcal{M}_r$ of a tensor $x \in X$ with a precision ε. with respect to the precision ε (point (b)). A perturbation analysis of the Richardson method shows that the sequence $\{u^k\}_{k\geq 0}$ satisfies $\limsup_{k\to\infty} \|u - u^k\|_X \leq \varepsilon/(1-\rho)$.

To conclude this section, we showed that provided efficient preconditioners are used, iterative solvers can be used for the solution of a tensor structured equation. Also, we note that such strategy only addresses the point (b).

[7] $\kappa = \|\mathcal{P}\mathcal{A}\|_{X\to X}\|(\mathcal{P}\mathcal{A})^{-1}\|_{X\to X}$.

3.3.3 Methods based on Alternating Least Squares

We present here the Alternating Least Squares (ALS) algorithm which is used for the best approximation problem in subsets of tensors with bounded ranks (point (a)). Let us note that other methods have been considered for these problems (for example a Newton algorithm is proposed in [54]), but ALS is very popular for its simplicity and efficiency in many applications.

As mentioned in Section 3.2, any low-rank tensor $x_r \in \mathcal{M}_r(X)$ admits a parametrization that takes the general form:

$$x_r = \mathcal{F}_{\mathcal{M}_r}(p_1, \dots, p_K),$$

where $p_1, \dots, p_K$ refers either to the vectors $x_{i_\nu}^{(\nu)}$, the core a or the core tensors $a^{(\nu)}$ (see equation (1.26)), and where $\mathcal{F}_{\mathcal{M}_r}$ is a multilinear map. For example, and according to relation (1.24), a possible parametrization for the Tucker format $\mathcal{T}_r(X)$ is given by

$$\mathcal{F}_{\mathcal{T}_r}\left(a, \{x_{i_1}^{(1)}\}_{i_1=1}^{r_1}, \dots, \{x_{i_d}^{(d)}\}_{i_d=1}^{r_d}\right) = \sum_{i_1=1}^{r_1} \dots \sum_{i_d=1}^{r_d} a_{i_1,\dots,i_d}\, x_{i_1}^{(1)} \otimes \dots \otimes x_{i_d}^{(d)}.$$

The map $\mathcal{F}_{\mathcal{M}_r}$ is linear with respect to the elements $p_k \in P_k$, where P_k denotes the appropriate vector space. Since $\|\cdot\|_X$ is a Hilbert norm, the minimization of the error $\|x - x_r\|_X$ with respect to any parameter p_k corresponds to a Least Squares problem. The *Alternating Least Squares* (ALS) algorithm consists in minimizing the error with respect to the parameters $p_1, \dots, p_K$ one after the other:

$$p_k^{\text{new}} = \arg\min_{p_k \in P_k} \|x - \mathcal{F}_{\mathcal{M}_r}\left(p_1^{\text{new}}, \dots, p_{k-1}^{\text{new}}, p_k, p_{k+1}^{\text{old}}, \dots, p_K^{\text{old}}\right)\|_X.$$

This operation can be repeated several times to improve the accuracy of the approximation. The convergence of the ALS has been analyzed in many papers, see for example [55, 113, 121]. In particular, provided the initial iterate is sufficiently close to the best approximation of x in $\mathcal{M}_r(X)$, ALS is proved to converge to this best approximation.

In the situation where a tensor $u \in X$ is defined as the solution of equation (1.17), the minimization over $\|u - u_r\|_X$ for $u_r \in \mathcal{M}_r(X)$ is not feasible since u is unknown. In this case, we can introduce a minimal residual problem:

$$\min_{u_r \in \mathcal{M}_r(X)} \|\mathcal{A}u_r - \mathcal{B}\|_*, \tag{1.32}$$

which can still be solved using ALS. In practice, two choices of the norm $\|\cdot\|_*$ in X' are commonly made. The first choice is to take the dual norm of X, *i.e.* $\|\cdot\|_* = \|\cdot\|_{X'}$. Thanks to equations (1.20) and (1.21), the relations $\underline{\alpha}\|v\|_X \leq \|\mathcal{A}v\|_* \leq \overline{\beta}\|v\|_X$ holds for any $v \in X$. Therefore, Céa's lemma states that the solution u_r^* of the minimization problem (1.32) satisfies

$$\|u - u_r^*\|_X \leq \gamma^{-1} \min_{u_r \in \mathcal{M}_r(X)} \|u - u_r\|_X,$$

where $\gamma^{-1} = \overline{\beta}/\underline{\alpha}$[8]. When $\mathcal{A}$ is symmetric definite positive, the other choice for the norm $\|\cdot\|_*$ is such that $\|\mathcal{A}v\|_*^2 = \langle \mathcal{A}v, v\rangle$ for any $v \in X$. For this choice of norm, we have

$$\|u - u_r^*\|_X \leq \gamma^{-1/2} \min_{u_r \in \mathcal{M}_r(X)} \|u - u_r\|_X,$$

where the quasi-optimality constant $\gamma^{-1/2}$ is improved compared to the case where $\|\cdot\|_* = \|\cdot\|_{X'}$.

Remark 3.3. *When the space X is endowed with the norm $\|\cdot\|_X$ such that $\|\cdot\|_X^2 = \|\mathcal{A}\cdot\|_*^2 = \langle \mathcal{A}\cdot, \cdot\rangle$ (this norm is often called the* energy norm *for the space X, or the norm induced by the operator), the problem (1.32) can be interpreted as a best approximation problem $\min_{u_r \in \mathcal{M}_r(X)} \|u - u_r\|_X$. If the number of terms $m_{\mathcal{A}}$ in the decomposition (1.22) is larger than 1, this norm $\|\cdot\|_X$ does not satisfies the property (1.28), so that the methods based on SVD can not be used for solving the best approximation problem or for obtaining a controlled approximation of the best approximation.*

3.3.4 Greedy algorithm and similarities with the RB method

Another possibility for the solution of $\mathcal{A}u = \mathcal{B}$ using low-rank approximations is to use a Greedy algorithm where at each iteration a rank-one correction is added to the current approximation. This algorithm, often referred as the *Proper Generalized Decomposition* (PGD, see [3, 35, 101, 102]), can be summarized as follows:

$$w^{k+1} \in \operatorname*{argmin}_{w \in \mathcal{C}_1(X)} \|\mathcal{A}\big(u^k + w\big) - \mathcal{B}\|_*, \tag{1.33}$$

$$u^{k+1} = u^k + w^{k+1}. \tag{1.34}$$

In practice, the rank-one approximation problem[9] that defines the correction w^{k+1} can be solved using the ALS algorithm. One can find in [119] an analysis of this

[8] Note that this is the same γ which is involved in the analysis of the Reduced Basis method, see Section 2.3.

[9] The minimization problem over $\mathcal{C}_1(X)$ is well posed, since the set $\mathcal{C}_r(X)$ is closed for $r = 1$.

greedy algorithm, which, under quite general assumptions, converges to u. However, we observe in practice that this greedy algorithm provides suboptimal approximations, and the number of iterations k for reaching a desired precision ε can be very large.

Remark 3.4. *Of course this greedy algorithm can also be used for the approximation of a given tensor x. The correction is then defined by*

$$w^{k+1} \in \underset{w \in \mathcal{C}_1(X)}{\operatorname{argmin}} \|x - x^k - w\|_X.$$

In the particular case where $d = 2$ and where $\|\cdot\|_X$ is an induced norm (see equation (1.28)), this greedy algorithm yields the optimal approximation x^k of x in $\mathcal{C}_k(X_1 \otimes X_2)$ (thanks to equation (1.29)). In general, greedy algorithms do not yield best approximation.

There exist many variants of greedy algorithms. For example, at each iteration one can add an *update phase* that aims at improving the accuracy of the current approximation. Such updates can be done by using an ALS for $u^{k+1} = \mathcal{F}_{\mathcal{M}_r}(p_1, \ldots, p_K)$, which is stored in some low-rank format $\mathcal{M}_r$[10].

We conclude this section by showing an analogy with the Reduced Basis method which also relies on a greedy algorithm. For that, we readopt the notation $X = V \otimes S$ without considering the possible tensor product structure of S. Let us assume that we are given an approximation $u_r = \sum_{i=1}^r v_i \otimes \lambda_i$ of u. If we want to improve this approximation, we can update the functions λ_i by solving the following minimization problem:

$$\min_{\lambda_1, \ldots, \lambda_r \in S} \|\mathcal{A}\left(\sum_{i=1}^r v_i \otimes \lambda_i\right) - \mathcal{B}\|_*.$$

If the residual norm $\|\cdot\|_*$ is the one induced by the operator ($\|\cdot\|_X^2 = \langle \mathcal{A}\cdot, \cdot\rangle$), then the stationarity condition of the latter minimization problem is:

$$\left\langle \mathcal{A}\left(\sum_{i=1}^r v_i \otimes \lambda_i\right), \left(\sum_{i=1}^r v_i \otimes \widetilde{\lambda}_i\right)\right\rangle = \left\langle \mathcal{B}, \left(\sum_{i=1}^r v_i \otimes \widetilde{\lambda}_i\right)\right\rangle$$

for any $\widetilde{\lambda}_i \in S$. Denoting $V_r = \operatorname{span}\{v_1, \ldots, v_r\}$, this is equivalent to find $u_r \in V_r \otimes S$ such that

$$\int_\Xi \langle A(\xi) u_r(\xi), \widetilde{u}_r(\xi)\rangle \mathrm{d}\mu(\xi) = \int_\Xi \langle b(\xi), \widetilde{u}_r(\xi)\rangle \mathrm{d}\mu(\xi) \tag{1.35}$$

[10] Since u^k is defined by the sum of rank-one tensors, we naturally have $u^k \in \mathcal{C}_k(X)$. However u^k can be easily "converted" in another tensor format such as the Tucker format $\mathcal{T}_{(k,\ldots,k)}(X)$.

holds for any $\widetilde{u}_r \in V_r \otimes S$. Obviously, if we define $u_r(\xi)$ as the Galerkin projection of $u(\xi)$ on the reduced space V_r (see Section 2.1), then $u_r(\xi)$ satisfies (1.35). Reciprocally, one can show that the solution of (1.35) is (almost surely) the Galerkin projection of $u(\xi)$ on V_r. This equivalence is true only for $S = L^2_\mu(\Xi)$. If S is as an approximation space in $L^2_\mu(\Xi)$, the equivalence is no longer true.

Here, we showed that the update of the functions λ_i of an approximation $u_r = \sum_{i=1}^r v_i \otimes \lambda_i$ yields to the Galerkin projection of $u(\xi)$ on $V_r = \operatorname{span}\{v_1, \ldots, v_r\}$. Now, let us consider the greedy algorithm (1.33)–(1.34) where at each iteration the functions λ_i are updated. This algorithm can be interpreted as a greedy algorithm for the construction of a reduced space V_r. The similarity is striking with the RB method: both approaches rely on a greedy construction of a reduced space. The difference is that the RB method tries to achieve the optimality with respect to the $L^\infty_\mu(\Xi; V)$-norm, whereas the other method with respect to the $L^2_\mu(\Xi; V)$-norm.

Chapter 2

Interpolation of inverse operators for preconditioning parameter-dependent equations

This chapter is based on the article [127], with additional numerical illustrations.

We propose a method for the construction of preconditioners of parameter-dependent matrices for the solution of large systems of parameter-dependent equations. The proposed method is an interpolation of the matrix inverse based on a projection of the identity matrix with respect to the Frobenius norm. Approximations of the Frobenius norm using random matrices are introduced in order to handle large matrices. The resulting statistical estimators of the Frobenius norm yield quasi-optimal projections that are controlled with high probability. Strategies for the adaptive selection of interpolation points are then proposed for different objectives in the context of projection-based model order reduction methods: the improvement of residual-based error estimators, the improvement of the projection on a given reduced approximation space, or the recycling of computations for sampling based model order reduction methods.

Contents

1 Introduction 31

2 Interpolation of the inverse of a parameter-dependent matrix using Frobenius norm projection 33

2.1 Projection using Frobenius norm 33

2.2 Projection using a Frobenius semi-norm 35

2.3 Ensuring the invertibility of the preconditioner for positive definite matrix 43

2.4 Practical computation of the projection 45

3 Preconditioners for projection-based model reduction . . . 46

3.1 Projection of the solution on a given reduced subspace 47

3.2 Greedy construction of the solution reduced subspace 49

4 Selection of the interpolation points 52

4.1 Greedy approximation of the inverse of a parameter-dependent matrix 52

4.2 Selection of points for improving the projection on a reduced space 55

4.3 Recycling factorizations of operator's evaluations - Application to reduced basis method 55

5 Numerical results 56

5.1 Illustration on a one parameter-dependent model 56

5.2 Multi-parameter-dependent equation 61

6 Conclusion 69

1 Introduction

This chapter is concerned with the solution of large systems of parameter-dependent equations of the form

$$A(\xi)u(\xi) = b(\xi), \tag{2.1}$$

where ξ takes values in some parameter set Ξ. Such problems occur in several contexts such as parametric analyses, optimization, control or uncertainty quantification, where ξ are random variables that parametrize model or data uncertainties. The efficient solution of equation (2.1) generally requires the construction of preconditioners for the operator $A(\xi)$, either for improving the performance of iterative solvers or for improving the quality of residual-based projection methods.

A basic preconditioner can be defined as the inverse (or any preconditioner) of the matrix $A(\bar{\xi})$ at some nominal parameter value $\bar{\xi} \in \Xi$ or as the inverse (or any preconditioner) of a mean value of $A(\xi)$ over Ξ (see e.g. [53,63]). When the operator only slightly varies over the parameter set Ξ, these parameter-independent preconditioners behave relatively well. However, for large variabilities, they are not able to provide a good preconditioning over the whole parameter set Ξ. A first attempt to construct a parameter-dependent preconditioner can be found in [47], where the authors compute through quadrature a polynomial expansion of the parameter-dependent factors of a LU factorization of $A(\xi)$. More recently, a linear Lagrangian interpolation of the matrix inverse has been proposed in [33]. The generalization to any standard multivariate interpolation method is straightforward. However, standard approximation or interpolation methods require the evaluation of matrix inverses (or factorizations) for many instances of ξ on a prescribed structured grid (quadrature or interpolation), that becomes prohibitive for large matrices and high dimensional parametric problems.

In this chapter, we propose an interpolation method for the inverse of matrix $A(\xi)$. The interpolation is obtained by a projection of the inverse matrix on a linear span of samples of $A(\xi)^{-1}$ and takes the form

$$P_m(\xi) = \sum_{i=1}^{m} \lambda_i(\xi) A(\xi_i)^{-1},$$

where $\xi_1, \ldots, \xi_m$ are m arbitrary interpolation points in Ξ. A natural interpolation could be obtained by minimizing the condition number of $P_m(\xi)A(\xi)$ over the $\lambda_i(\xi)$, which is a Clarke regular strongly pseudoconvex optimization problem [96]. However, the solution of this non standard optimization problem for many instances of

ξ is intractable and proposing an efficient solution method in a multi-query context remains a challenging issue. Here, the projection is defined as the minimizer of the Frobenius norm of $I - P_m(\xi)A(\xi)$, that is a quadratic optimization problem. Approximations of the Frobenius norm using random matrices are introduced in order to handle large matrices. These statistical estimations of the Frobenius norm allow to obtain quasi-optimal projections that are controlled with high probability. Since we are interested in large matrices, $A(\xi_i)^{-1}$ are here considered as implicit matrices for which only efficient matrix-vector multiplications are available. Typically, a factorization (e.g. LU) of $A(\xi_i)$ is computed and stored. Note that when the storage of factorizations of several samples of the operator is unaffordable or when efficient preconditioners are readily available, one could similarly consider projections of the inverse operator on the linear span of preconditioners of samples of the operator. However, the resulting parameter-dependent preconditioner would be no more an interpolation of preconditioners. This straightforward extension of the proposed method is not analyzed in the present chapter.

The chapter then presents several contributions in the context of projection-based model order reduction methods (e.g. Reduced Basis, Proper Orthogonal Decomposition (POD), Proper Generalized Decompositon) that rely on the projection of the solution $u(\xi)$ of (2.1) on a low-dimensional approximation space. We first show how the proposed preconditioner can be used to define a Galerkin projection-based on the preconditioned residual, which can be interpreted as a Petrov-Galerkin projection of the solution with a parameter-dependent test space. Then, we propose adaptive construction of the preconditioner, based on an adaptive selection of interpolation points, for different objectives: (i) the improvement of error estimators based on preconditioned residuals, (ii) the improvement of the quality of projections on a given low-dimensional approximation space, or (iii) the recycling of computations for sample-based model order reduction methods. Starting from a m-point interpolation, these adaptive strategies consist in choosing a new interpolation point based on different criteria. In (i), the new point is selected for minimizing the distance between the identity and the preconditioned operator. In (ii), it is selected for improving the quasi-optimality constant of Petrov-Galerkin projections which measures how far the projection is from the best approximation on the reduced approximation space. In (iii), the new interpolation point is selected as a new sample determined for the approximation of the solution and not of the operator. The interest of the latter approach is that when direct solvers are used to solve equation (2.1) at some sample points, the corresponding factorizations of the matrix can be

stored and the preconditioner can be computed with a negligible additional cost.

The chapter is organized as follows. In Section 2 we present the method for the interpolation of the inverse of a parameter-dependent matrix. In Section 3, we show how the preconditioner can be used for the definition of a Petrov-Galerkin projection of the solution of (2.1) on a given reduced approximation space, and we provide an analysis of the quasi-optimality constant of this projection. Then, different strategies for the selection of interpolation points for the preconditioner are proposed in Section 4. Finally, in Section 5, numerical experiments will illustrate the efficiency of the proposed preconditioning strategies for different projection-based model order reduction methods.

Note that the proposed preconditioner could be also used (a) for improving the quality of Galerkin projection methods where a projection of the solution $u(\xi)$ is searched on a subspace of functions of the parameters (e.g. polynomial or piecewise polynomial spaces) [46,97,101], or (b) for preconditioning iterative solvers for (2.1), in particular solvers based on low-rank truncations that require a low-rank structure of the preconditioner [65,66,86,98].

2 Interpolation of the inverse of a parameter-dependent matrix using Frobenius norm projection

In this section, we propose a construction of an interpolation of the matrix-valued function $\xi \mapsto A(\xi)^{-1} \in \mathbb{R}^{n\times n}$ for given interpolation points $\xi_1, \dots, \xi_m$ in Ξ. We let $P_i = A(\xi_i)^{-1}$, $1 \le i \le m$. For large matrices, the explicit computation of P_i is usually not affordable. Therefore, P_i is here considered as an implicit matrix and we assume that the product of P_i with a vector can be computed efficiently. In practice, factorizations of matrices $A(\xi_i)$ are stored.

2.1 Projection using Frobenius norm

We introduce the subspace $Y_m = \text{span}\{P_1, \dots, P_m\}$ of $\mathbb{R}^{n\times n}$. An approximation $P_m(\xi)$ of $A(\xi)^{-1}$ in Y_m is then defined by

$$P_m(\xi) = \underset{P\in Y_m}{\text{argmin}} \, \|I - PA(\xi)\|_F, \tag{2.2}$$

where I denotes the identity matrix of size n, and $\|\cdot\|_F$ is the Frobenius norm such that $\|B\|_F^2 = \langle B, B\rangle_F$ with $\langle B, C\rangle_F = \text{trace}(B^T C)$. Since $A(\xi_i)^{-1} \in Y_m$, we

have the interpolation property $P_m(\xi_i) = A(\xi_i)^{-1}$, $1 \leq i \leq m$. The minimization of $\|I - PA\|_F$ has been first proposed in [74] for the construction of a preconditioner P in a subspace of matrices with given sparsity pattern (SPAI method). The following proposition gives some properties of the operator $P_m(\xi)A(\xi)$ (see Lemma 2.6 and Theorem 3.2 in [70]).

Proposition 2.1. *Let $P_m(\xi)$ be defined by (2.2). We have*

$$(1 - \alpha_m(\xi))^2 \leq \|I - P_m(\xi)A(\xi)\|_F^2 \leq n(1 - \alpha_m^2(\xi)), \tag{2.3}$$

where $\alpha_m(\xi)$ is the lowest singular value of $P_m(\xi)A(\xi)$ verifying $0 \leq \alpha_m(\xi) \leq 1$, with $P_m(\xi)A(\xi) = I$ if and only if $\alpha_m(\xi) = 1$. Also, the following bound holds for the condition number of $P_m(\xi)A(\xi)$:

$$\kappa(P_m(\xi)A(\xi)) \leq \frac{\sqrt{n - (n-1)\alpha_m^2(\xi)}}{\alpha_m(\xi)}. \tag{2.4}$$

Under the condition $\|I - P_m(\xi)A(\xi)\|_F < 1$, equations (2.3) and (2.4) imply that

$$\kappa(P_m(\xi)A(\xi)) \leq \frac{\sqrt{n - (n-1)(1 - \|I - P_m(\xi)A(\xi)\|_F)^2}}{1 - \|I - P_m(\xi)A(\xi)\|_F}.$$

For all $\lambda \in \mathbb{R}^m$, we have

$$\|I - \sum_{i=1}^{m} \lambda_i P_i A(\xi)\|_F^2 = n - 2\lambda^T S(\xi) + \lambda^T M(\xi)\lambda,$$

where the matrix $M(\xi) \in \mathbb{R}^{m\times m}$ and the vector $S(\xi) \in \mathbb{R}^m$ are given by

$$M_{i,j}(\xi) = \mathrm{trace}(A^T(\xi)P_i^T P_j A(\xi)) \quad \text{and} \quad S_i(\xi) = \mathrm{trace}(P_i A(\xi)).$$

Therefore, the solution of problem (2.2) is $P_m(\xi) = \sum_{i=1}^m \lambda_i(\xi)P_i$ with $\lambda(\xi)$ the solution of $M(\xi)\lambda(\xi) = S(\xi)$. When considering a small number m of interpolation points, the computation time for solving this system of equations is negligible. However, the computation of $M(\xi)$ and $S(\xi)$ requires the evaluation of traces of matrices $A^T(\xi)P_i^T P_j A(\xi)$ and $P_i A(\xi)$ for all $1 \leq i, j \leq m$. Since the P_i are implicit matrices, the computation of such products of matrices is not affordable for large matrices. Of course, since $\mathrm{trace}(B) = \sum_{i=1}^n e_i^T B e_i$, the trace of an implicit matrix B could be obtained by computing the product of B with the canonical vectors $e_1, \ldots, e_n$, but this approach is clearly not affordable for large n.

Hereafter, we propose an approximation of the above construction using an approximation of the Frobenius norm which requires less computational efforts.

2.2 Projection using a Frobenius semi-norm

Here, we define an approximation $P_m(\xi)$ of $A(\xi)^{-1}$ in Y_m by

$$P_m(\xi) = \underset{P \in Y_m}{\operatorname{argmin}} \, \|(I - PA(\xi))\Theta\|_F, \tag{2.5}$$

where $\Theta \in \mathbb{R}^{n\times K}$, with $K \leq n$. $B \mapsto \|B\Theta\|_F$ defines a semi-norm on $\mathbb{R}^{n\times n}$. Here, we assume that the linear map $P \mapsto PA(\xi)\Theta$ is injective on Y_m so that the solution of (2.5) is unique. This requires $K \geq m$ and is satisfied when $\operatorname{rank}(\Theta) \geq m$ and Y_m is the linear span of linearly independent invertible matrices. Then, the solution $P_m(\xi) = \sum_{i=1}^m \lambda_i(\xi)P_i$ of (2.5) is such that the vector $\lambda(\xi) \in \mathbb{R}^m$ satisfies $M^\Theta(\xi)\lambda(\xi) = S^\Theta(\xi)$, with

$$M^\Theta_{i,j}(\xi) = \operatorname{trace}(\Theta^T A^T(\xi) P_i^T P_j A(\xi)\Theta) \quad \text{and} \quad S^\Theta_i(\xi) = \operatorname{trace}(\Theta^T P_i A(\xi)\Theta). \tag{2.6}$$

The procedure for the computation of $M^\Theta(\xi)$ and $S^\Theta(\xi)$ is given in Algorithm 1. Note that only mK matrix-vector products involving the implicit matrices P_i are required.

Algorithm 1 Computation of $M^\Theta(\xi)$ and $S^\Theta(\xi)$

Require: $A(\xi)$, $\{P_1, \ldots, P_m\}$ and $\Theta = (\theta_1, \ldots, \theta_K)$

Ensure: $M^\Theta(\xi)$ and $S^\Theta(\xi)$

1: Compute the vectors $w_{i,k} = P_i A(\xi)\theta_k \in \mathbb{R}^n$, for $1 \leq k \leq K$ and $1 \leq i \leq m$

2: Set $W_i = (w_{i,1}, \ldots, w_{i,K}) \in \mathbb{R}^{n\times K}$, $1 \leq i \leq m$

3: Compute $M^\Theta_{i,j}(\xi) = \operatorname{trace}(W_i^T W_j)$ for $1 \leq i, j \leq m$

4: Compute $S^\Theta_i(\xi) = \operatorname{trace}(\Theta^T W_i)$ for $1 \leq i \leq m$

Now the question is to choose a matrix Θ such that $\|(I - PA(\xi))\Theta\|_F$ provides a good approximation of $\|I - PA(\xi)\|_F$ for any $P \in Y_m$ and $\xi \in \Xi$.

2.2.1 Hadamard matrices for the estimation of the Frobenius norm of an implicit matrix

Let B an implicit n-by-n matrix (consider $B = I - PA(\xi)$, with $P \in Y_m$ and $\xi \in \Xi$). Following [11], we show how Hadamard matrices can be used for the estimation of the Frobenius norm of an implicit matrix. The goal is to find a matrix Θ such that $\|B\Theta\|_F$ is a good approximation of $\|B\|_F$. The relation $\|B\Theta\|_F^2 = \operatorname{trace}(B^T B\Theta\Theta^T)$ suggests that Θ should be such that $\Theta\Theta^T$ is as close as possible to the identity matrix. For example, we would like Θ to minimize

$$err(\Theta)^2 = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\neq i}^n (\Theta\Theta^T)^2_{i,j} = \frac{\|I - \Theta\Theta^T\|_F^2}{n(n-1)},$$

which is the mean square magnitude of the off-diagonal entries of $\Theta\Theta^T$. The bound $err(\Theta) \geq \sqrt{(n-K)/((n-1)K)}$ is known to hold for any $\Theta \in \mathbb{R}^{n\times K}$ whose rows have unit norm [125]. Hadamard matrices can be used to construct matrices Θ such that the corresponding error $err(\Theta)$ is close to the bound, see [11].

A Hadamard matrix H_s is a s-by-s matrix whose entries are ± 1, and which satisfies $H_s H_s^T = sI$ where I is the identity matrix of size s. For example,

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

is a Hadamard matrix of size $s = 2$. The Kronecker product of two Hadamard matrices is again a Hadamard matrix. Then it is possible to build a Hadamard matrix whose size s is a power of 2 using a recursive procedure: $H_{2^{k+1}} = H_2 \otimes H_{2^k}$. The (i,j)-entry of this matrix is $(-1)^{a^T b}$, where a and b are the binary vectors such that $i = \sum_{k\geq 0} 2^k a_k$ and $j = \sum_{k\geq 0} 2^k b_k$. For a sufficiently large $s = 2^k \geq \max(n, K)$, we define the *rescaled partial Hadamard matrix* $\Theta \in \mathbb{R}^{n\times K}$ as the first n rows and the first K columns of $H_s/\sqrt{K}$.

2.2.2 Statistical estimation of the Frobenius norm of an implicit matrix

For the computation of the Frobenius norm of B, we can also use a statistical estimator as first proposed in [82]. The idea is to define a random matrix $\Theta \in \mathbb{R}^{n\times K}$ with a suitable distribution law $\mathcal{D}$ such that $\|B\Theta\|_F$ provides a controlled approximation of $\|B\|_F$ with high probability.

Definition 2.2. *A distribution $\mathcal{D}$ over $\mathbb{R}^{n\times K}$ satisfies the (ε,δ)-concentration property if for any $B \in \mathbb{R}^{n\times n}$,*

$$\mathbb{P}(|\|B\Theta\|_F^2 - \|B\|_F^2| \geq \varepsilon\|B\|_F^2) \leq \delta, \tag{2.7}$$

where $\Theta \sim \mathcal{D}$.

Two distributions $\mathcal{D}$ will be considered here.

(a) The *rescaled Rademacher distribution*. Here the entries of $\Theta \in \mathbb{R}^{n\times K}$ are independent and identically distributed with $\Theta_{i,j} = \pm K^{-1/2}$ with probability 1/2. According to Theorem 13 in [4], the rescaled Rademacher distribution satisfies the (ε,δ)-concentration property for

$$K \geq 6\varepsilon^{-2}\ln(2n/\delta). \tag{2.8}$$

(b) The *subsampled Randomized Hadamard Transform distribution* (SRHT), first introduced in [1]. Here we assume that n is a power of 2. It is defined by $\Theta = K^{-1/2}(RH_nD)^T \in \mathbb{R}^{n\times K}$ where

- $D \in \mathbb{R}^{n\times n}$ is a diagonal random matrix where $D_{i,i}$ are independent Rademacher random variables (i.e. $D_{i,i} = \pm 1$ with probability 1/2),
- $H_n \in \mathbb{R}^{n\times n}$ is a Hadamard matrix of size n (see Section 2.2.1),
- $R \in \mathbb{R}^{K\times n}$ is a subset of K rows from the identity matrix of size n chosen uniformly at random and without replacement.

In other words, we randomly select K rows of H_n without replacement, and we multiply the columns by $\pm K^{-1/2}$. We can find in [22, 120] an analysis of the SRHT matrix properties. In the case where n is not a power of 2, we define the partial SRHT (P-SRHT) matrix $\Theta \in \mathbb{R}^{n\times K}$ as the first n rows of a SRHT matrix of size $s \times K$, where $s = 2^{\lceil \log_2(n)\rceil}$ is the smallest power of 2 such that $n \le s < 2n$. The following proposition shows that the (P-SRHT) distribution satisfies the (ε, δ)-concentration property.

Proposition 2.3. *The (P-SRHT) distribution satisfies the (ε, δ)-concentration property for*

$$K \ge 2(\varepsilon^2 - \varepsilon^3/3)^{-1} \ln(4/\delta)(1 + \sqrt{8\ln(4n/\delta)})^2. \tag{2.9}$$

Proof: Let $B \in \mathbb{R}^{n\times n}$. We define the square matrix $\widetilde{B}$ of size $s = 2^{\lceil \log_2(n)\rceil}$, whose first $n \times n$ diagonal block is B, and 0 elsewhere. Then we have $\|\widetilde{B}\|_F = \|B\|_F$. The rest of the proof is similar to the one of Lemma 4.10 in [22]. We consider the events $A = \{(1-\varepsilon)\|\widetilde{B}\|_F^2 \le \|\widetilde{B}\Theta\|_F^2 \le (1+\varepsilon)\|\widetilde{B}\|_F^2\}$ and $E = \{\max_i \|\widetilde{B}DH_s^T e_i\|_2^2 \le (1+\sqrt{8\ln(2s/\delta)})^2\|\widetilde{B}\|_F^2\}$, where e_i is the i-th canonical vector of $\mathbb{R}^s$. The relation $\mathbb{P}(A^c) \le \mathbb{P}(A^c|E) + \mathbb{P}(E^c)$ holds. Thanks to Lemma 4.6 in [22] (with $t = \sqrt{8\ln(2s/\delta)}$) we have $\mathbb{P}(E^c) \le \delta/2$. Now, using the scalar Chernoff bound (Theorem 2.2 in [120] with $k = 1$) we have

$$\begin{aligned}
\mathbb{P}(A^c|E) &= \mathbb{P}\big(\|\widetilde{B}\Theta\|_F^2 \le (1-\varepsilon)\|\widetilde{B}\|_F^2|E\big) + \mathbb{P}\big(\|\widetilde{B}\Theta\|_F^2 \ge (1+\varepsilon)\|\widetilde{B}\|_F^2|E\big) \\
&\le (e^{-\varepsilon}(1-\varepsilon)^{-1+\varepsilon})^{K(1+\sqrt{8\ln(2s/\delta)})^{-2}} + (e^{\varepsilon}(1+\varepsilon)^{-1-\varepsilon})^{K(1+\sqrt{8\ln(2s/\delta)})^{-2}} \\
&\le 2(e^{\varepsilon}(1+\varepsilon)^{-1-\varepsilon})^{K(1+\sqrt{8\ln(2s/\delta)})^{-2}} \le 2e^{K(-\varepsilon^2/2+\varepsilon^3/6)(1+\sqrt{8\ln(2s/\delta)})^{-2}}.
\end{aligned}$$

The condition (2.9) implies $\mathbb{P}(A^c|E) \leq \delta/2$, and then $\mathbb{P}(A^c) \leq \delta/2 + \delta/2 = \delta$, which ends the proof. ∎

Such statistical estimators are particularly interesting for that they provide approximations of the Frobenius norm of large matrices, with a number of columns K for Θ which scales as the logarithm of n, see (2.8) and (2.9). However, the concentration property (2.7) holds only for a given matrix B. The following proposition 2.4 extends these concentration results for any matrix B in a given subspace. The proof is inspired from the one of Theorem 6 in [43]. The essential ingredient is the existence of an ε-net for the unit ball of a finite dimensional space (see [21]).

Proposition 2.4. *Let $\Theta \in \mathbb{R}^{n\times K}$ be a random matrix whose distribution $\mathcal{D}$ satisfies the (ε, δ)-concentration property, with $\varepsilon \leq 1$. Then, for any L-dimensional subspace of matrices $M_L \subset \mathbb{R}^{n\times n}$ and for any $C > 1$, we have*

$$\mathbb{P}\big(|\|B\Theta\|_F^2 - \|B\|_F^2| \geq \varepsilon(C+1)/(C-1)\|B\|_F^2, \forall B \in M_L\big) \leq (9C/\varepsilon)^L\delta. \quad (2.10)$$

Proof: We consider the unit ball $\mathcal{B}_L = \{B \in M_L : \|B\|_F \leq 1\}$ of the subspace M_L. It is shown in [21] that for any $\widetilde{\varepsilon} > 0$, there exists a net $\mathcal{N}_L^{\widetilde{\varepsilon}} \subset \mathcal{B}_L$ of cardinality lower than $(3/\widetilde{\varepsilon})^L$ such that

$$\min_{B_{\widetilde{\varepsilon}} \in \mathcal{N}_L^{\widetilde{\varepsilon}}} \|B - B_{\widetilde{\varepsilon}}\|_F \leq \widetilde{\varepsilon}, \quad \forall B \in \mathcal{B}_L.$$

In other words, any element of the unit ball $\mathcal{B}_L$ can be approximated by an element of $\mathcal{N}_L^{\widetilde{\varepsilon}}$ with an error less than $\widetilde{\varepsilon}$. Using the (ε, δ)-concentration property and a union bound, we obtain

$$|\|B_{\widetilde{\varepsilon}}\Theta\|_F^2 - \|B_{\widetilde{\varepsilon}}\|_F^2| \leq \varepsilon\|B_{\widetilde{\varepsilon}}\|_F^2, \quad \forall B_{\widetilde{\varepsilon}} \in \mathcal{N}_L^{\widetilde{\varepsilon}}, \quad (2.11)$$

with a probability at least $1 - \delta(3/\widetilde{\varepsilon})^L$. We now impose the relation $\widetilde{\varepsilon} = \varepsilon/(3C)$, where $C > 1$. To prove (2.10), it remains to show that equation (2.11) implies

$$|\|B\Theta\|_F^2 - \|B\|_F^2| \leq \varepsilon(C+1)/(C-1)\|B\|_F^2, \quad \forall B \in M_L. \quad (2.12)$$

We define $B^* \in \arg\max_{B\in\mathcal{B}_L} |\|B\Theta\|_F^2 - \|B\|_F^2|$. Let $B_{\widetilde{\varepsilon}} \in \mathcal{N}_L^{\widetilde{\varepsilon}}$ be such that $\|B^* - B_{\widetilde{\varepsilon}}\|_F \leq \widetilde{\varepsilon}$, and $B_{\widetilde{\varepsilon}}^* = \arg\min_{B\in\mathrm{span}(B_{\widetilde{\varepsilon}})} \|B^* - B\|_F$. Then we have $\|B^* - B_{\widetilde{\varepsilon}}^*\|_F^2 = \|B^*\|_F^2 - \|B_{\widetilde{\varepsilon}}^*\|_F^2 \leq \widetilde{\varepsilon}^2$ and $\langle B^* - B_{\widetilde{\varepsilon}}^*, B_{\widetilde{\varepsilon}}^*\rangle = 0$, where $\langle\cdot,\cdot\rangle$ is the inner

product associated to the Frobenius norm $\|\cdot\|_F$. We have

$$\begin{aligned}\eta &:= |\|B^*\Theta\|_F^2 - \|B^*\|_F^2| = |\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta + B^*_{\widetilde{\varepsilon}}\Theta\|_F^2 - \|B^* - B^*_{\widetilde{\varepsilon}} + B^*_{\widetilde{\varepsilon}}\|_F^2| \\ &= |\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta\|_F^2 + 2\langle (B^* - B^*_{\widetilde{\varepsilon}})\Theta, B^*_{\widetilde{\varepsilon}}\Theta\rangle + \|B^*_{\widetilde{\varepsilon}}\Theta\|_F^2 - \|B^* - B^*_{\widetilde{\varepsilon}}\|_F^2 - \|B^*_{\widetilde{\varepsilon}}\|_F^2| \\ &\leq |\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta\|_F^2 - \|B^* - B^*_{\widetilde{\varepsilon}}\|_F^2| + |\|B^*_{\widetilde{\varepsilon}}\Theta\|_F^2 - \|B^*_{\widetilde{\varepsilon}}\|_F^2| + 2\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta\|_F\|B^*_{\widetilde{\varepsilon}}\Theta\|_F.\end{aligned}$$

We now have to bound the three terms in the previous expression. Firstly, since $(B^* - B^*_{\widetilde{\varepsilon}})/\|B^* - B^*_{\widetilde{\varepsilon}}\|_F \in \mathcal{B}_L$, the relation $|\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta\|_F^2 - \|B^* - B^*_{\widetilde{\varepsilon}}\|_F^2| \leq \|B^* - B^*_{\widetilde{\varepsilon}}\|_F^2\eta \leq \widetilde{\varepsilon}^2\eta$ holds. Secondly, (2.11) gives $|\|B^*_{\widetilde{\varepsilon}}\Theta\|_F^2 - \|B^*_{\widetilde{\varepsilon}}\|_F^2| \leq \varepsilon\|B^*_{\widetilde{\varepsilon}}\|_F^2 \leq \varepsilon$. Thirdly, by definition of η, we can write $\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta\|_F^2 \leq (1+\eta)\|B^* - B^*_{\widetilde{\varepsilon}}\|_F^2 \leq \widetilde{\varepsilon}^2(1+\eta)$ and $\|B^*_{\widetilde{\varepsilon}}\Theta\|_F^2 \leq (1+\varepsilon)\|B^*_{\widetilde{\varepsilon}}\|_F^2 \leq 1+\varepsilon$, so that we obtain $2\|(B^* - B^*_{\widetilde{\varepsilon}})\Theta\|_F\|B^*_{\widetilde{\varepsilon}}\Theta\|_F \leq 2\widetilde{\varepsilon}\sqrt{1+\varepsilon}\sqrt{1+\eta}$. Finally, from (2.11), we obtain

$$\eta \leq \widetilde{\varepsilon}^2\eta + \varepsilon + 2\widetilde{\varepsilon}\sqrt{1+\varepsilon}\sqrt{1+\eta} \tag{2.13}$$

Since $\varepsilon \leq 1$, we have $\widetilde{\varepsilon} = \varepsilon/(3C) < 1/3$. Then $\widetilde{\varepsilon}^2 \leq \widetilde{\varepsilon}$ and $\sqrt{1+\varepsilon} \leq 3/2$, so that (2.13) implies

$$\eta \leq \widetilde{\varepsilon}\eta + \varepsilon + 3\widetilde{\varepsilon}\sqrt{1+\eta} \leq \widetilde{\varepsilon}\eta + \varepsilon + 3\widetilde{\varepsilon}(1+\eta/2) \leq 3\widetilde{\varepsilon}\eta + \varepsilon + 3\widetilde{\varepsilon},$$

and then $\eta \leq (\varepsilon + 3\widetilde{\varepsilon})/(1 - 3\widetilde{\varepsilon}) \leq \varepsilon(C+1)/(C-1)$. By definition of η, we can write $|\|B\Theta\|_F^2 - \|B\|_F^2| \leq \varepsilon(C+1)/(C-1)$ for any $B \in \mathcal{B}_L$, that implies (2.12). ■

Proposition 2.5. *Let $\xi \in \Xi$, and let $P_m(\xi) \in Y_m$ be defined by (2.5) where $\Theta \in \mathbb{R}^{n\times K}$ is a realization of a rescaled Rademacher matrix with*

$$K \geq 6\varepsilon^{-2}\ln(2n(9C/\varepsilon)^{m+1}/\delta), \tag{2.14}$$

or a realization of a P-SRHT matrix with

$$K \geq 2(\varepsilon^2 - \varepsilon^3/3)^{-1}\ln(4(9C/\varepsilon)^{m+1}/\delta)(1 + \sqrt{8\ln(4n(9C/\varepsilon)^{m+1}/\delta)})^2 \tag{2.15}$$

for some $\delta > 0$, $\varepsilon \leq 1$ and $C > 1$. Assuming $\varepsilon' = \varepsilon(C+1)/(C-1) < 1$,

$$\|I - P_m(\xi)A(\xi)\|_F \leq \sqrt{\frac{1+\varepsilon'}{1-\varepsilon'}}\min_{P\in Y_m}\|I - PA(\xi)\|_F \tag{2.16}$$

holds with a probability higher than $1-\delta$.

Proof: Let us introduce the subspace $M_{m+1} = Y_m A(\xi) + \mathrm{span}(I)$ of dimension less than $m+1$, such that $\{I - PA(\xi) : P \in Y_m\} \subset M_{m+1}$. Then, we note that with the conditions (2.14) or (2.15), the distribution law $\mathcal{D}$ of the random matrix Θ satisfies the $(\varepsilon, \delta(\varepsilon/(9C))^{m+1})$-concentration property. Thanks to Proposition 2.4, the probability that

$$|\|(I - PA(\xi))\Theta\|_F^2 - \|I - PA(\xi)\|_F^2| \leq \varepsilon' \|I - PA(\xi)\|_F^2$$

holds for any $P \in Y_m$ is higher than $1-\delta$. Then, by definition of $P_m(\xi)$ (2.5), we have with a probability at least $1-\delta$ that for any $P \in Y_m$, it holds

$$\begin{aligned}\|I - P_m(\xi)A(\xi)\|_F &\leq \frac{1}{\sqrt{1-\varepsilon'}}\|(I - P_m(\xi)A(\xi))\Theta\|_F, \\ &\leq \frac{1}{\sqrt{1-\varepsilon'}}\|(I - PA(\xi))\Theta\|_F \leq \frac{\sqrt{1+\varepsilon'}}{\sqrt{1-\varepsilon'}}\|I - PA(\xi)\|_F.\end{aligned}$$

Then, taking the minimum over $P \in Y_m$, we obtain (2.16). ∎

Similarly to Proposition 2.1, we obtain the following properties for $P_m(\xi)A(\xi)$, with $P_m(\xi)$ the solution of (2.5).

Proposition 2.6. *Under the assumptions of Proposition 2.5, the inequalities*

$$(1 - \alpha_m(\xi))^2(1-\varepsilon')^{-1} \leq \|(I - P_m(\xi)A(\xi))\Theta\|_F^2 \leq n\left(1 - (1-\varepsilon')\alpha_m^2(\xi)\right) \quad (2.17)$$

and

$$\kappa(P_m(\xi)A(\xi)) \leq \alpha_m(\xi)^{-1}\sqrt{n(1-\varepsilon')^{-1} - (n-1)\alpha_m^2(\xi)} \quad (2.18)$$

hold with probability $1-\delta$, where $\alpha_m(\xi)$ is the lowest singular value of $P_m(\xi)A(\xi)$.

Proof: The optimality condition for $P_m(\xi)$ yields $\|(I - P_m(\xi)A(\xi))\Theta\|_F^2 = \|\Theta\|_F^2 - \|P_m(\xi)A(\xi)\Theta\|_F^2$. Since $P_m(\xi)A(\xi) \in M_{m+1}$ (where M_{m+1} is the subspace introduced in the proof of Proposition (2.5)), we have

$$\|P_m(\xi)A(\xi)\Theta\|_F^2 \geq (1-\varepsilon')\|P_m(\xi)A(\xi)\|_F^2 \quad (2.19)$$

with a probability higher than $1-\delta$. Using $\|\Theta\|_F^2 = n$ (which is satisfies for any realization of the rescaled Rademacher or the P-SRHT distribution), we obtain $\|(I - P_m(\xi)A(\xi))\Theta\|_F^2 \leq n - (1-\varepsilon')\|P_m(\xi)A(\xi)\|_F^2$ with a probability higher than

$1-\delta$. Then, $\|P_m(\xi)A(\xi)\|_F^2 \geq n\alpha_m(\xi)^2$ yields the right inequality of (2.17). Following the proof of Lemma 2.6 in [70], we have $(1-\alpha_m(\xi)^2) \leq \|I - P_m(\xi)A(\xi)\|_F^2$. Together with (2.19), it yields the left inequality of (2.17). Furthermore, with probability $1-\delta$, we have $n-(1-\varepsilon')\|P_m(\xi)A(\xi)\|_F^2 \geq 0$. Since the square of the Frobenius norm of matrix $P_m(\xi)A(\xi)$ is the sum of squares of its singular values, we deduce

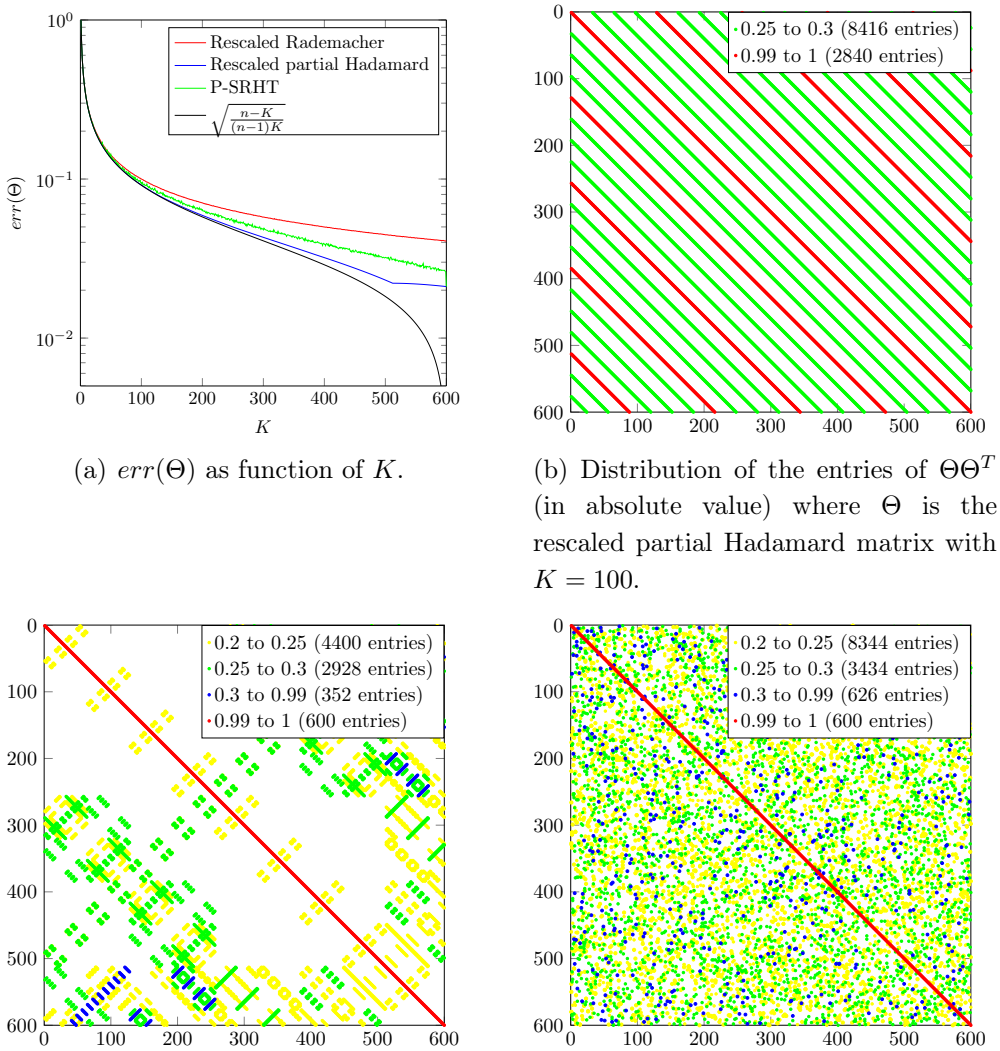
$$(n-1)\alpha_m(\xi)^2 + \beta_m(\xi)^2 \leq \|P_m(\xi)A(\xi)\|_F^2 \leq n(1-\varepsilon')^{-1}$$

with a probability higher than $1-\delta$, where $\beta_m(\xi)$ is the largest singular value of $P_m(\xi)A(\xi)$. Then (2.18) follows from the definition of $\kappa(P_m(\xi)A(\xi)) = \beta_m(\xi)/\alpha_m(\xi)$. ■

2.2.3 Comparison and comments

We have presented different possibilities for the definition of Θ. The rescaled partial Hadamard matrices introduced in section 2.2.1 have the advantage that the error $err(\Theta)$ is close to the theoretical bound $\sqrt{(n-K)/((n-1)K)}$, see Figure 2.1(a) (note that the rows of Θ have unit norm). Furthermore, an interesting property is that $\Theta\Theta^T$ has a structured pattern (see Figure 2.1(b)). As noticed in [11], when $K = 2^q$ the matrix $\Theta\Theta^T$ have non-zero entries only on the 2^{qk}-th upper and lower diagonals, with $k \geq 0$. As a consequence, the error on the estimation of $\|B\|_F$ will be induced only by the non-zero off-diagonal entries of B that occupy the 2^{qk}-th upper and lower diagonals, with $k \geq 1$. If the entries of B vanish away from the diagonal, the Frobenius norm is expected to be accurately estimated. Note that the P-SRHT matrices can be interpreted as a "randomized version" of the rescaled partial Hadamard matrices, and Figure 2.1(a) shows that the error $err(\Theta)$ associated to the P-SRHT matrix behaves almost like the rescaled partial Hadamard matrix. Also, P-SRHT matrices yield a structured pattern for $\Theta\Theta^T$, see Figure 2.1(c). The rescaled Rademacher matrices give higher errors $err(\Theta)$ and yield matrices $\Theta\Theta^T$ with no specific patterns, see Figure 2.1(d).

The advantage of using rescaled Rademacher matrices or P-SRHT matrices is that we can control the quality of the resulting projection $P_m(\xi)$ with high probability, provided a sufficiently large number of rows K for Θ (see Proposition 2.5). Table 2.1 shows the theoretical value for K in order to obtain the quasi-optimality result (2.16) with $\sqrt{(1+\varepsilon')/(1-\varepsilon')} = 10$ and $\delta = 0.1\%$. We see that K grows very slowly with respect to the matrix size n. Also, the dependence of K with respect to m is linear for the rescaled Rademacher matrices and quadratic for the P-SRHT



(a) $err(\Theta)$ as function of K.

(b) Distribution of the entries of $\Theta\Theta^T$ (in absolute value) where Θ is the rescaled partial Hadamard matrix with $K = 100$.

(c) Distribution of the entries of $\Theta\Theta^T$ (in absolute value) where Θ is a sample of the P-SRHT matrix with $K = 100$.

(d) Distribution of the entries of $\Theta\Theta^T$ (in absolute value) where Θ is a sample of the rescaled Rademacher matrix with $K = 100$.

Figure 2.1: Comparison between the rescaled partial Hadamard, the rescaled Rademacher and the P-SRHT matrix for the definition of matrix Θ, with $n = 600$.

matrices (see equations (2.14) and (2.15)). However, these theoretical bounds for K are very pessimistic, especially for the P-SRHT matrices. In practice, we observe that a very small value for K may provide very good results (see Section 5). Also, it is worth mentioning that our numerical experiments do not show significant differences between the rescaled partial Hadamard, the rescaled Rademacher and the P-SRHT matrices.

(a) Rescaled Rademacher distribution.

	$m = 2$	$m = 5$	$m = 10$	$m = 20$	$m = 50$
$n = 10^4$	239	363	567	972	2 185
$n = 10^6$	270	395	599	1 005	2 219
$n = 10^8$	301	427	632	1 038	2 253

(b) P-SRHT distribution.

	$m = 2$	$m = 5$	$m = 10$	$m = 20$	$m = 50$
$n = 10^4$	27 059	63 298	155 129	455 851	2 286 645
$n = 10^6$	30 597	69 129	164 750	473 011	2 326 301
$n = 10^8$	34 112	74 929	174 333	490 126	2 365 914

Table 2.1: Theoretical number of columns K for the random matrix Θ in order to ensure (2.16), with $\sqrt{(1+\varepsilon')/(1-\varepsilon')} = 10$ and $\delta = 0.1\%$. The constant C has been chosen in order to minimize K.

2.3 Ensuring the invertibility of the preconditioner for positive definite matrix

Here, we propose a modification of the interpolation which ensures that $P_m(\xi)$ is invertible when $A(\xi)$ is positive definite.

Since $A(\xi_i)$ is positive definite, $P_i = A(\xi_i)^{-1}$ is positive definite. We introduce the vectors $\gamma^- \in \mathbb{R}^m$ and $\gamma^+ \in \mathbb{R}^m$ whose components

$$\gamma_i^- = \inf_{w \in \mathbb{R}^n} \frac{\langle P_i w, w\rangle}{\|w\|^2} > 0 \quad \text{and} \quad \gamma_i^+ = \sup_{w \in \mathbb{R}^n} \frac{\langle P_i w, w\rangle}{\|w\|^2} < \infty$$

correspond respectively to the lowest and highest eigenvalues of the symmetric part of P_i. Then, for any $P = \sum_{i=1}^m \lambda_i P_i \in Y_m$,

$$\inf_{w \in \mathbb{R}^n} \frac{\langle Pw, w\rangle}{\|w\|^2} \geq \langle \lambda^+, \gamma^-\rangle - \langle \lambda^-, \gamma^+\rangle, \tag{2.20}$$

where $\lambda^+ \geq 0$ and $\lambda^- \geq 0$ are respectively the positive and negative parts of $\lambda = \lambda^+ - \lambda^- \in \mathbb{R}^m$. As a consequence, if the right hand side of (2.20) is strictly positive, then P is invertible. Furthermore, we have $\|P\| \leq \langle \lambda^+ + \lambda^-, C\rangle$, where $C \in \mathbb{R}^m$ is the vector of component $C_i = \|P_i\|$, where $\|P_i\|$ denotes the operator norm of P_i. If we assume that $\langle \lambda^+, \gamma^-\rangle - \langle \lambda^-, \gamma^+\rangle > 0$, the condition number of P satisfies

$$\kappa(P) = \|P\| \, \|P^{-1}\| \leq \|P\| \left(\inf_{w \in \mathbb{R}^n} \frac{\langle Pw, w\rangle}{\|w\|^2} \right)^{-1} \leq \frac{\langle \lambda^+ + \lambda^-, C\rangle}{\langle \lambda^+, \gamma^-\rangle - \langle \lambda^-, \gamma^+\rangle}.$$

It is then possible to bound $\kappa(P)$ by $\bar{\kappa}$ by imposing

$$\langle \lambda^+ + \lambda^-, C\rangle \leq \bar{\kappa}(\langle \lambda^+, \gamma^-\rangle - \langle \lambda^-, \gamma^+\rangle),$$

which is a linear inequality constraint on λ^+ and λ^-. We introduce two convex subsets of Y_m defined by

$$Y_m^{\bar{\kappa}} = \left\{ \sum_{i=1}^m \lambda_i^+ P_i - \sum_{i=1}^m \lambda_i^- P_i \ : \ \begin{array}{c} \lambda_i^+ \geq 0, \ \lambda_i^- \geq 0 \\ \langle \lambda^+, \gamma^-\rangle - \langle \lambda^-, \gamma^+\rangle \geq 0 \\ \langle \lambda^+, \bar{\kappa}\gamma^- - C\rangle - \langle \lambda^-, \bar{\kappa}\gamma^+ + C\rangle \geq 0 \end{array} \right\},$$

$$Y_m^+ = \left\{ \sum_{i=1}^m \lambda_i P_i \ : \ \lambda_i \geq 0 \right\}.$$

From (2.20), we have that any nonzero element of Y_m^+ is invertible, while any nonzero element of $Y_m^{\bar{\kappa}}$ is invertible and has a condition number lower than $\bar{\kappa}$. Under the condition $\bar{\kappa} \geq \max_i C_i/\gamma_i^-$, we have

$$Y_m^+ \subset Y_m^{\bar{\kappa}} \subset Y_m. \tag{2.22}$$

Then definitions (2.2) and (2.5) for the approximation $P_m(\xi)$ can be replaced respectively by

$$P_m(\xi) = \underset{P \in Y_m^+ \text{ or } Y_m^{\bar{\kappa}}}{\operatorname{argmin}} \ \|I - PA(\xi)\|_F, \tag{2.23a}$$

$$P_m(\xi) = \underset{P \in Y_m^+ \text{ or } Y_m^{\bar{\kappa}}}{\operatorname{argmin}} \ \|(I - PA(\xi))\Theta\|_F, \tag{2.23b}$$

which are quadratic optimization problems with linear inequality constraints. Furthermore, since $P_i \in Y_m^+$ for all i, all the resulting projections $P_m(\xi)$ interpolate $A(\xi)^{-1}$ at the points $\xi_1, \ldots, \xi_m$.

The following proposition shows that properties (2.3) and (2.4) still hold for the preconditioned operator.

Proposition 2.7. *The solution $P_m(\xi)$ of (2.23a) is such that $P_m(\xi)A(\xi)$ satisfies (2.3) and (2.4). Also, under the assumptions of Proposition 2.5 the solution $P_m(\xi)$ of (2.23b) is such that $P_m(\xi)A(\xi)$ satisfies (2.17) and (2.18) with a probability higher than $1 - \delta$.*

Proof: Since Y_m^+ (or $Y_m^{\bar{\kappa}}$) is a closed and convex positive cone, the solution $P_m(\xi)$ of (2.23a) is such that $trace((I - P_m(\xi)A(\xi))^T(P_m(\xi) - P)A(\xi)) \geq 0$ for all $P \in Y_m^+$ (or $Y_m^{\bar{\kappa}}$). Taking $P = 2P_m(\xi)$ and $P = 0$, we obtain that $\text{trace}((I - P_m(\xi)A(\xi))^T P_m(\xi)A(\xi)) = 0$, which implies $\|P_m(\xi)A(\xi)\|_F^2 = \text{trace}(P_m(\xi)A(\xi))$. We refer to the proof of Lemma 2.6 and Theorem 3.2 in [70] to deduce (2.3) and (2.4). Using the same arguments, we prove that the solution $P_m(\xi)$ of (2.23b) satisfies $\|P_m(\xi)A(\xi)\Theta\|_F^2 = \text{trace}(\Theta^T P_m(\xi)A(\xi)\Theta)$, and then that (2.17) and (2.18) hold with a probability higher than $1 - \delta$. ∎

2.4 Practical computation of the projection

Here, we detail how to efficiently compute $M^\Theta(\xi)$ and $S^\Theta(\xi)$ given in equation (2.6) in a multi-query context, i.e. for several different values of ξ. The same methodology can be applied for computing $M(\xi)$ and $S(\xi)$. We assume that the operator $A(\xi)$ has an affine expansion of the form

$$A(\xi) = \sum_{k=1}^{m_A} \Phi_k(\xi) A_k, \tag{2.24}$$

where the A_k are matrices in $\mathbb{R}^{n\times n}$ and the $\Phi_k : \Xi \to \mathbb{R}$ are real-valued functions. Then $M^\Theta(\xi)$ and $S^\Theta(\xi)$ also have the affine expansions

$$M_{i,j}^\Theta(\xi) = \sum_{k=1}^{m_A}\sum_{l=1}^{m_A} \Phi_k(\xi)\Phi_l(\xi)\ \text{trace}(\Theta^T A_k^T P_i^T P_j A_l \Theta), \tag{2.25a}$$

$$S_i^\Theta(\xi) = \sum_{k=1}^{m_A} \Phi_k(\xi)\ \text{trace}(\Theta^T P_i A_k \Theta), \tag{2.25b}$$

respectively. Computing the multiple terms of these expansions would require many computations of traces of implicit matrices and also, it would require the computation of the affine expansion of $A(\xi)$. Here, we use the methodology introduced in [29] for obtaining affine decompositions with a lower number of terms. These decompositions only require the knowledge of functions Φ_k in the affine decomposition (2.24), and evaluations of $M_{i,j}^\Theta(\xi)$ and $S_i^\Theta(\xi)$ (that means evaluations of $A(\xi)$) at some selected points. We briefly recall this methodology.

Suppose that $g : \Xi \to X$, with X a vector space, has an affine decomposition $g(\xi) = \sum_{k=1}^m \zeta_k(\xi) g_k$, with $\zeta_k : \Xi \to \mathbb{R}$ and $g_k \in X$. We first compute an interpolation of $\zeta(\xi) = (\zeta_1(\xi), \ldots, \zeta_m(\xi))$ under the form $\zeta(\xi) = \sum_{k=1}^{m_g} \Psi_k(\xi)\zeta(\xi_k^*)$, with

$m_g \leq m$, where $\xi_1^*, \ldots, \xi_{m_g}^*$ are interpolation points and $\Psi_1(\xi), \ldots, \Psi_{m_g}(\xi)$ the associated interpolation functions. Such an interpolation can be computed with the Empirical Interpolation Method [94] described in Algorithm 2. Then, we obtain an affine decomposition $g(\xi) = \sum_{k=1}^{m_g} \Psi_k(\xi) g(\xi_k^*)$ which can be computed from evaluations of g at interpolation points ξ_k^*.

Algorithm 2 Empirical Interpolation Method (EIM).

Require: $(\zeta_1(\cdot), \ldots, \zeta_m(\cdot))$

Ensure: $\Psi_1(\cdot), \ldots, \Psi_k(\cdot)$ and $\xi_1^*, \ldots, \xi_k^*$

1: Define $R_1(i, \xi) = \zeta_i(\xi)$ for all i, ξ
2: Initialize $e = 1$, $k = 0$
3: **while** $e \geq$ *tolerance* (in practice the machine precision) **do**
4: $\quad k = k + 1$
5: $\quad$ Find $(i_k^*, \xi_k^*) \in \operatorname{argmax}_{i,\xi} |R_k(i, \xi)|$
6: $\quad$ Set the error to $e = |R_k(i_k^*, \xi_k^*)|$
7: $\quad$ Actualize $R_{k+1}(i, \xi) = R_k(i, \xi) - R_k(i, \xi_k^*) R_k(i_k^*, \xi) / R_k(i_k^*, \xi_k^*)$ for all i, ξ
8: **end while**
9: Fill in the k-by-k matrix $Q : Q_{i,j} = \zeta_{i_i^*}(\xi_j^*)$ for all $1 \leq i, j \leq k$
10: Compute $\Psi_i(\xi) = \sum_{j=1}^{k} (Q^{-1})_{i,j} \zeta_{i_j^*}(\xi)$ for all ξ and $1 \leq i \leq k$

Applying the above procedure to both $M^\Theta(\xi)$ and $S^\Theta(\xi)$, we obtain

$$M^\Theta(\xi) \approx \sum_{k=1}^{m_M} \Psi_k(\xi)\ M^\Theta(\xi_k^*), \quad S^\Theta(\xi) \approx \sum_{k=1}^{m_S} \widetilde{\Psi}_k(\xi)\ S^\Theta(\widetilde{\xi}_k^*). \tag{2.26}$$

The first (so-called *offline*) step consists in computing the interpolation functions $\Psi_k(\xi)$ and $\widetilde{\Psi}_k(\xi)$ and associated interpolation points ξ_k^* and $\widetilde{\xi}_k^*$ using Algorithm 2 with input $\{\Phi_i \Phi_j\}_{1 \leq i,j \leq m_A}$ and $\{\Phi_i\}_{1 \leq i \leq m_A}$ respectively, and then in computing matrices $M^\Theta(\xi_k^*)$ and vectors $S^\Theta(\widetilde{\xi}_k^*)$ using Algorithm 1. The second (so-called *online*) step simply consists in computing the matrix $M^\Theta(\xi)$ and the vector $S^\Theta(\xi)$ for a given value of ξ using (2.26).

3 Preconditioners for projection-based model reduction

We consider a parameter-dependent linear equation

$$A(\xi) u(\xi) = b(\xi), \tag{2.27}$$

with $A(\xi) \in \mathbb{R}^{n\times n}$ and $b(\xi) = \mathbb{R}^n$. Projection-based model reduction consists in projecting the solution $u(\xi)$ onto an approximation space $V_r \subset V := \mathbb{R}^n$ of low dimension $r \ll n$. In this section, we show how the preconditioner $P_m(\xi)$ can be used for the definition of the projection and for the construction of the approximation space V_r.

V is endowed with the norm $\|\cdot\|_V$ defined by $\|\cdot\|_V^2 = \langle R_V\cdot,\cdot\rangle$, where R_V is a symmetric positive definite matrix and $\langle\cdot,\cdot\rangle$ is the canonical inner product of $\mathbb{R}^n$. We also introduce the dual norm $\|\cdot\|_{V'} = \|R_V^{-1}\cdot\|_V$ such that for any $v, w \in V$ we have $|\langle v, w\rangle| \leq \|v\|_V\ \|w\|_{V'}$.

3.1 Projection of the solution on a given reduced subspace

Here, we suppose that the approximation space V_r has been computed by some model order reduction method. The best approximation of $u(\xi)$ on V_r is $u_r^*(\xi) = \arg\min_{v\in V_r} \|u(\xi) - v\|_V$ and is characterized by the orthogonality condition

$$\langle u_r^*(\xi) - u(\xi), R_V v_r\rangle = 0, \quad \forall v_r \in V_r, \tag{2.28}$$

or equivalently by the Petrov-Galerkin orthogonality condition

$$\langle A(\xi)u_r^*(\xi) - b(\xi), A^{-T}(\xi)R_V v_r\rangle = 0, \quad \forall v_r \in V_r. \tag{2.29}$$

Obviously the computation of test functions $A^{-T}(\xi)R_V v_r$ for basis functions v_r of V_r is prohibitive. By replacing $A(\xi)^{-1}$ by $P_m(\xi)$, we obtain the feasible Petrov-Galerkin formulation

$$\langle A(\xi)u_r(\xi) - b(\xi), P_m^T(\xi)R_V v_r\rangle = 0, \quad \forall v_r \in V_r. \tag{2.30}$$

Denoting by $U \in \mathbb{R}^{n\times r}$ a matrix whose range is V_r, the solution of (2.30) is $u_r(\xi) = Ua(\xi)$ where the vector $a(\xi) \in \mathbb{R}^r$ is the solution of

$$(U^T R_V P_m(\xi)A(\xi)U)a(\xi) = U^T R_V P_m(\xi)b(\xi).$$

Note that (2.30) corresponds to the standard Galerkin projection when replacing $P_m(\xi)$ by R_V^{-1}.

We give now a quasi-optimality result for the approximation $u_r(\xi)$. This analysis relies on the notion of δ-proximality introduced in [41].

Proposition 3.1. *Let $\delta_{r,m}(\xi) \in [0,1]$ be defined by*

$$\delta_{r,m}(\xi) = \max_{v_r \in V_r} \min_{w_r \in V_r} \frac{\|v_r - R_V^{-1}(P_m(\xi)A(\xi))^T R_V w_r\|_V}{\|v_r\|_V}. \tag{2.31}$$

The solutions $u_r^*(\xi) \in V_r$ and $u_r(\xi) \in V_r$ of (2.28) and (2.30) satisfy

$$\|u_r^*(\xi) - u_r(\xi)\|_V \leq \delta_{r,m}(\xi)\|u(\xi) - u_r(\xi)\|_V. \tag{2.32}$$

Moreover, if $\delta_{r,m}(\xi) < 1$ holds, then

$$\|u(\xi) - u_r(\xi)\|_V \leq (1 - \delta_{r,m}(\xi)^2)^{-1/2}\|u(\xi) - u_r^*(\xi)\|_V. \tag{2.33}$$

Proof: The orthogonality condition (2.28) yields

$$\langle u_r^*(\xi) - u_r(\xi), R_V v_r\rangle = \langle u(\xi) - u_r(\xi), R_V v_r\rangle = \langle b(\xi) - A(\xi)u_r(\xi), A^{-T}(\xi)R_V v_r\rangle$$

for all $v_r \in V_r$. Using (2.30), we have that for any $w_r \in V_r$,

$$\begin{aligned}
\langle u_r^*(\xi) - u_r(\xi), R_V v_r\rangle &= \langle b(\xi) - A(\xi)u_r(\xi), A^{-T}(\xi)R_V v_r - P_m(\xi)^T R_V w_r\rangle, \\
&= \langle u(\xi) - u_r(\xi), R_V v_r - (P_m(\xi)A(\xi))^T R_V w_r\rangle, \\
&\leq \|u(\xi) - u_r(\xi)\|_V \; \|R_V v_r - (P_m(\xi)A(\xi))^T R_V w_r\|_{V'} \\
&= \|u(\xi) - u_r(\xi)\|_V \; \|v_r - R_V^{-1}(P_m(\xi)A(\xi))^T R_V w_r\|_V.
\end{aligned}$$

Taking the infimum over $w_r \in V_r$ and by the definition of $\delta_{r,m}(\xi)$, we obtain

$$\langle u_r^*(\xi) - u_r(\xi), R_V v_r\rangle \leq \delta_{r,m}(\xi)\|u(\xi) - u_r(\xi)\|_V \; \|v_r\|_V.$$

Then, noting that $u_r^*(\xi) - u_r(\xi) \in V_r$, we obtain

$$\|u_r^*(\xi) - u_r(\xi)\|_V = \sup_{v_r \in V_r} \frac{\langle u_r^*(\xi) - u_r(\xi), R_V v_r\rangle}{\|v_r\|_V} \leq \delta_{r,m}(\xi)\|u(\xi) - u_r(\xi)\|_V,$$

that is (2.32). Finally, using orthogonality condition (2.28), we have that

$$\begin{aligned}
\|u(\xi) - u_r(\xi)\|_V^2 &= \|u(\xi) - u_r^*(\xi)\|_V^2 + \|u_r^*(\xi) - u_r(\xi)\|_V^2, \\
&\leq \|u(\xi) - u_r^*(\xi)\|_V^2 + \delta_{r,m}(\xi)^2\|u(\xi) - u_r(\xi)\|_V^2,
\end{aligned}$$

from which we deduce (2.33) when $\delta_{r,m}(\xi) < 1$. ∎

An immediate consequence of Proposition 3.1 is that when $\delta_{r,m}(\xi) = 0$, the

Petrov-Galerkin projection $u_r(\xi)$ coincides with the orthogonal projection $u_r^*(\xi)$. Following [42], we show in the following proposition that $\delta_{r,m}(\xi)$ can be computed by solving an eigenvalue problem of size r.

Proposition 3.2. *We have $\delta_{r,m}(\xi) = \sqrt{1-\gamma}$, where γ is the lowest eigenvalue of the generalized eigenvalue problem $Cx = \gamma Dx$, with*

$$C = U^T B(B^T R_V^{-1} B)^{-1} B^T U \in \mathbb{R}^{r\times r},$$
$$D = U^T R_V U \in \mathbb{R}^{r\times r},$$

where $B = (P_m(\xi)A(\xi))^T R_V U \in \mathbb{R}^{n\times r}$ and where $U \in \mathbb{R}^{n\times r}$ is a matrix whose range is V_r.

Proof: Since the range of U is V_r, we have

$$\delta_{r,m}(\xi)^2 = \max_{a\in\mathbb{R}^r} \min_{b\in\mathbb{R}^r} \frac{\|Ua - R_V^{-1}Bb\|_V^2}{\|Ua\|_V^2}.$$

For any $a \in \mathbb{R}^r$, the minimizer b^* of $\|Ua - R_V^{-1}Bb\|_V^2$ over $b \in \mathbb{R}^r$ is given by $b^* = (B^T R_V^{-1} B)^{-1} B^T Ua$. Therefore, we have $\|Ua - R_V^{-1}Bb^*\|_V^2 = \|Ua\|_V^2 - \langle Ua, Bb^*\rangle$, and

$$\delta_{r,m}^2(\xi) = 1 - \inf_{a\in\mathbb{R}^r} \frac{\langle U^T B(B^T R_V^{-1} B)^{-1} B^T Ua, a\rangle}{\langle U^T R_V Ua, a\rangle},$$

which concludes the proof. ∎

3.2 Greedy construction of the solution reduced subspace

Following the idea of the Reduced Basis method [114, 123], a sequence of nested approximation spaces $\{V_r\}_{r\geq 1}$ in V can be constructed by a greedy algorithm such that $V_{r+1} = V_r + \mathrm{span}(u(\xi_{r+1}^{RB}))$, where ξ_{r+1}^{RB} is a point where the error of approximation of $u(\xi)$ in V_r is maximal. An ideal greedy algorithm using the best approximation in V_r and an exact evaluation of the projection error is such that

$$u_r^*(\xi) \text{ is the orthogonal projection of } u(\xi) \text{ on } V_r \text{ defined by (2.28)}, \tag{2.34a}$$
$$\xi_{r+1}^{RB} \in \operatorname*{argmax}_{\xi\in\Xi} \|u(\xi) - u_r^*(\xi)\|_V. \tag{2.34b}$$

This ideal greedy algorithm is not feasible in practice since $u(\xi)$ is not known. Therefore, we rather rely on a feasible weak greedy algorithm such that

$$u_r(\xi) \text{ is the Petrov-Galerkin projection of } u(\xi) \text{ on } V_r \text{ defined by (2.30)}, \tag{2.35a}$$

$$\xi_{r+1}^{RB} \in \operatorname*{argmax}_{\xi\in\Xi} \|P_m(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V. \tag{2.35b}$$

Assume that

$$\underline{\alpha}_m \, \|u(\xi) - u_r(\xi)\|_V \leq \|P_m(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V \leq \bar{\beta}_m \, \|u(\xi) - u_r(\xi)\|_V$$

holds with $\underline{\alpha}_m = \inf_{\xi\in\Xi} \alpha_m(\xi) > 0$ and $\bar{\beta}_m = \sup_{\xi\in\Xi} \beta_m(\xi) < \infty$, where $\alpha_m(\xi)$ and $\beta_m(\xi)$ are respectively the lowest and largest singular values of $P_m(\xi)A(\xi)$ with respect to the norm $\|\cdot\|_V$, respectively defined by the infimum and supremum of $\|P_m(\xi)A(\xi)v\|_V$ over $v \in V$ such that $\|v\|_V = 1$. Then, we easily prove that algorithm (2.35) is such that

$$\|u(\xi_{r+1}^{RB}) - u_r(\xi_{r+1}^{RB})\|_V \geq \gamma_m \max_{\xi\in\Xi} \|u(\xi) - u_r(\xi)\|_V, \tag{2.36}$$

where $\gamma_m = \underline{\alpha}_m / \bar{\beta}_m \leq 1$ measures how far the selection of the new point is from the ideal greedy selection. Under condition (2.36), convergence results for this weak greedy algorithm can be found in [18, 49].

We give now sharper bounds for the preconditoned residual norm that exploits the fact that the approximation $u_r(\xi)$ is the Petrov-Galerkin projection.

Proposition 3.3. *Let $u_r(\xi)$ be the Petrov-Galerkin projection of $u(\xi)$ on V_r defined by (2.29). Then we have*

$$\alpha_{r,m}(\xi) \, \|u(\xi) - u_r(\xi)\|_V \leq \|P_m(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V \leq \beta_{r,m}(\xi) \, \|u(\xi) - u_r(\xi)\|_V,$$

with

$$\alpha_{r,m}(\xi) = \inf_{v\in V} \sup_{w_r\in V_r} \frac{\|(P_m(\xi)A(\xi))^T R_V v\|_{V'}}{\|v - w_r\|_V},$$

$$\beta_{r,m}(\xi) = \sup_{v\in V} \inf_{w_r\in V_r} \frac{\|(P_m(\xi)A(\xi))^T R_V (v - w_r)\|_{V'}}{\|v\|_V}.$$

Proof: For any $v \in V$ and $w_r \in V_r$ and according to (2.30), we have

$$\begin{aligned}\langle u(\xi) - u_r(\xi), R_V v\rangle &= \langle b(\xi) - A(\xi)u_r(\xi), A^{-T}(\xi)R_V v - P_m^T(\xi)R_V w_r\rangle \\ &= \langle P_m(\xi)(b(\xi) - A(\xi)u_r(\xi)), (P_m(\xi)A(\xi))^{-T}R_V v - R_V w_r\rangle \\ &\leq \|R\|_V \; \|(P_m(\xi)A(\xi))^{-T}R_V v - R_V w_r\|_{V'},\end{aligned}$$

where $R(\xi) := P_m(\xi)(b(\xi) - A(\xi)u_r(\xi))$. Taking the infimum over $w_r \in V_r$, dividing by $\|v\|_V$ and taking the supremum over $v \in V$, we obtain

$$\begin{aligned}\|u(\xi) - u_r(\xi)\|_V &\leq \|R(\xi)\|_V \sup_{v\in V} \inf_{w_r\in V_r} \frac{\|(P_m(\xi)A(\xi))^{-T}R_V v - R_V w_r\|_{V'}}{\|v\|_V}, \\ &= \|R(\xi)\|_V \sup_{v\in V} \inf_{w_r\in V_r} \frac{\|v - w_r\|_V}{\|(P_m(\xi)A(\xi))^T R_V v\|_{V'}}, \\ &= \|R(\xi)\|_V \left(\inf_{v\in V} \sup_{w_r\in V_r} \frac{\|(P_m(\xi)A(\xi))^T R_V v\|_{V'}}{\|v - w_r\|_V}\right)^{-1},\end{aligned}$$

which proves the first inequality. Furthermore, for any $v \in V$ and $w_r \in V_r$, we have

$$\begin{aligned}\langle P_m(\xi)(b(\xi) - A(\xi)u_r(\xi)), R_V v\rangle &= \langle b(\xi) - A(\xi)u_r(\xi), P_m^T(\xi)R_V(v - w_r)\rangle \\ &\leq \|u(\xi) - u_r(\xi)\|_V \; \|(P_m(\xi)A(\xi))^T R_V(v - w_r)\|_{V'}.\end{aligned}$$

Taking the infimum over $w_r \in V_r$, dividing by $\|v\|_V$ and taking the supremum over $v \in V$, we obtain the second inequality. ∎

Since $V_r \subset V_{r+1}$, we have $\alpha_{r+1,m}(\xi) \geq \alpha_{r,m}(\xi) \geq \alpha_m(\xi)$ and $\beta_{r+1,m}(\xi) \leq \beta_{r,m}(\xi) \leq \beta_m(\xi)$. Equation (2.36) holds with γ_m replaced by the parameter $\gamma_{r,m} = \underline{\alpha}_{r,m}/\ \bar{\beta}_{r,m}$. Since $\gamma_{r,m}$ increases with r, a reasonable expectation is that the convergence properties of the weak greedy algorithm will improve when r increases.

Remark 3.4. *When replacing $P_m(\xi)$ by R_V^{-1}, the preconditioned residual norm $\|P_m(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V$ turns out to be the residual norm $\|A(\xi)u_r(\xi) - b(\xi)\|_{V'}$, which is a standard choice in the Reduced Basis method for the greedy selection of points (with R_V being associated with the natural norm on V or with a norm associated with the operator at some nominal parameter value). This can be interpreted as a basic preconditioning method with a parameter-independent preconditioner.*

4 Selection of the interpolation points

In this section, we propose strategies for the adaptive selection of the interpolation points. For a given set of interpolation points $\xi_1, \ldots, \xi_m$, three different methods are proposed for the selection of a new interpolation point ξ_{m+1}. The first method aims at reducing uniformly the error between the inverse operator and its interpolation. The resulting interpolation of the inverse is pertinent for preconditioning iterative solvers or estimating errors based on preconditioned residuals. The second method aims at improving Petrov-Galerkin projections of the solution of a parameter-dependent equation on a given approximation space. The third method aims at reducing the cost for the computation of the preconditioner by reusing operators computed when solving samples of a parameter-dependent equation.

4.1 Greedy approximation of the inverse of a parameter-dependent matrix

A natural idea is to select a new interpolation point where the preconditioner $P_m(\xi)$ is not a good approximation of $A(\xi)^{-1}$. Obviously, an ideal strategy for preconditioning would be to choose ξ_{m+1} where the condition number of $P_m(\xi)A(\xi)$ is maximal. The computation of the condition number for many values of ξ being computationaly expensive, one could use upper bounds of this condition number, e.g. computed using SCM [83].

Here, we propose the following selection method: given an approximation $P_m(\xi)$ associated with interpolation points $\xi_1, \ldots, \xi_m$, a new point ξ_{m+1} is selected such that

$$\xi_{m+1} \in \operatorname*{argmax}_{\xi \in \Xi} \|(I - P_m(\xi)A(\xi))\Theta\|_F, \tag{2.38}$$

where the matrix Θ is either the random rescaled Rademacher matrix, or the P-SRHT matrix (see Section 2.2). This adaptive selection of the interpolation points yields the construction of an increasing sequence of subspaces $Y_{m+1} = Y_m + \operatorname{span}(A(\xi_{m+1})^{-1})$ in $Y = \mathbb{R}^{n\times n}$. This algorithm is detailed in Algorithm (3). The following lemma interprets the above construction as a weak greedy algorithm.

Lemma 4.1. *Assume that $A(\xi)$ satisfies $\underline{\alpha}_0\|\cdot\| \leq \|A(\xi)\cdot\| \leq \bar{\beta}_0\|\cdot\|$ for all $\xi \in \Xi$, and let $P_m(\xi)$ be defined by (2.5). Under the assumptions that there exists*

$\varepsilon \in [0,1[$ such that

$$|\|(I - PA(\xi))\Theta\|_F^2 - \|I - PA(\xi)\|_F^2| \leq \epsilon \|I - PA(\xi)\|_F^2 \tag{2.39}$$

holds for all ξ and $P \in Y_m$, we have

$$\|P_m(\xi_{m+1}) - A(\xi_{m+1})^{-1}\|_F \geq \gamma_\varepsilon \max_{\xi \in \Xi} \min_{P \in Y_m} \|P - A(\xi)^{-1}\|_F, \tag{2.40}$$

with $\gamma_\varepsilon = \underline{\alpha}_0\sqrt{1-\varepsilon}/(\bar{\beta}_0\sqrt{1+\varepsilon})$, and with ξ_{m+1} defined by (2.38).

Proof: Since $\|BC\|_F \leq \|B\|_F\|C\|$ holds for any matrices B and C, with $\|C\|$ the operator norm of C, we have for all $P \in Y$,

$$\|A(\xi)^{-1} - P\|_F \leq \|I - PA(\xi)\|_F\|A(\xi)^{-1}\| \leq \underline{\alpha}_0^{-1}\|I - PA(\xi)\|_F,$$
$$\|I - PA(\xi)\|_F \leq \|A(\xi)^{-1} - P\|_F\|A(\xi)\| \leq \bar{\beta}_0\|A(\xi)^{-1} - P\|_F.$$

Then, thanks to (2.39) we have

$$\|A(\xi)^{-1} - P\|_F \leq (\underline{\alpha}_0\sqrt{1-\varepsilon})^{-1}\|(I - PA(\xi))\Theta\|_F$$
$$\text{and} \quad \|(I - PA(\xi))\Theta\|_F \leq \bar{\beta}_0\sqrt{1+\varepsilon}\|A(\xi)^{-1} - P\|_F,$$

which implies

$$\frac{1}{\bar{\beta}_0\sqrt{1+\varepsilon}}\|(I - PA(\xi))\Theta\|_F \leq \|A(\xi)^{-1} - P\|_F \leq \frac{1}{\underline{\alpha}_0\sqrt{1-\varepsilon}}\|(I - PA(\xi))\Theta\|_F.$$

We easily deduce that ξ_{m+1} is such that (2.40) holds. ∎

Remark 4.2. *We have different possibilities to show that assumption (2.39) of Lemma 4.1 holds with high probability. When considering Ξ as a training set of finite cardinality, we can extend the results of proposition 2.5 to any $\xi \in \Xi$ using a union bound. In that case, the probability that (2.39) holds will be higher than $1 - \delta(\#\Xi)$. Another possibility is to use the affine decomposition of $A(\xi)$, see equation (2.24), so that the space $M_L = \text{span}\{I - PA(\xi) : \xi \in \Xi, P \in Y_m\}$ is of dimension $L = 1 + m_A m$. Then using proposition 2.4 we can obtain (2.39) with high probability.*

The quality of the resulting spaces Y_m have to be compared with the Kolmogorov

m-width of the set $A^{-1}(\Xi) := \{A(\xi)^{-1} : \xi \in \Xi\} \subset Y$, defined by

$$d_m(A^{-1}(\Xi))_Y = \min_{\substack{Y_m \subset Y \\ \dim(Y_m) = m}} \sup_{\xi \in \Xi} \min_{P \in Y_m} \|A(\xi)^{-1} - P\|_F, \tag{2.41}$$

which evaluates how well the elements of $A^{-1}(\Xi)$ can be approximated on a m-dimensional subspace of matrices. (2.40) implies that the following results holds (see Corollary 3.3 in [49]):

$$\|A(\xi)^{-1} - P_m(\xi)\|_F = \begin{cases} \mathcal{O}(m^{-a}) & \text{if } d_m(A^{-1}(\Xi))_Y = \mathcal{O}(m^{-a}) \\ \mathcal{O}(e^{-\tilde{c}m^b}) & \text{if } d_m(A^{-1}(\Xi))_Y = \mathcal{O}(e^{-cm^b}) \end{cases},$$

where $\tilde{c} > 0$ is a constant which depends on c and b. That means that if the Kolmogorov m-width has an algebraic or exponential convergence rate, then the weak greedy algorithm yields an error $\|P_m(\xi) - A(\xi)^{-1}\|_F$ which has the same type of convergence. Therefore, the proposed interpolation method will present good convergence properties when $d_m(A^{-1}(\Xi))_Y$ rapidly decreases with m.

Algorithm 3 Greedy selection of interpolation points.

Require: $A(\xi), \Theta, M$.

Ensure: Interpolation points $\xi_1, \ldots, \xi_M$ and interpolation $P_M(\xi)$.

```
Initialize P_0(ξ) = I
for m = 0 to M − 1 do
    Compute the new point ξ_{m+1} according to (2.38)
    Compute a factorization of A(ξ_{m+1})
    Define A(ξ_{m+1})^{-1} as an implicit operator
    Update the space Y_{m+1} = Y_m + span(A(ξ_{m+1})^{-1})
    Compute P_{m+1}(ξ) = arg min_{P∈Y_{m+1}} ‖(I − PA(ξ))Θ‖_F
end for
```

Remark 4.3. *When the parameter set Ξ is $[-1,1]^d$ (or a product of compact intervals), an exponential decay can be obtained when $A(\xi)^{-1}$ admits an holomorphic extension to a domain in $\mathbb{C}^d$ containing Ξ (see [31]).*

Remark 4.4. *Note that here, there is no constraint on the minimization problem over Y_m (either optimal subspaces or subspaces constructed by the greedy pro-*

cedure), so that we have no guaranty that the resulting approximations Y_m are invertible (see Section 2.3).

4.2 Selection of points for improving the projection on a reduced space

We here suppose that we want to find an approximation of the solution $u(\xi)$ of a parameter-dependent equation (2.27) onto a low-dimensional approximation space V_r, using a Petrov-Galerkin orthogonality condition given by (2.30). The best approximation is considered as the orthogonal projection defined by (2.28). The quantity $\delta_{r,m}(\xi)$ defined by (2.31) controls the quality of the Petrov-Galerkin projection on V_r (see Proposition 3.1). As indicated in Proposition 3.2, $\delta_{r,m}(\xi)$ can be efficiently computed. Thus, we propose the following selection strategy which aims at improving the quality of the Petrov-Galerkin projection: given a preconditioner $P_m(\xi)$ associated with interpolation points $\xi_1, \ldots, \xi_m$, the next point ξ_{m+1} is selected such that

$$\xi_{m+1} \in \operatorname*{argmax}_{\xi \in \Xi} \delta_{r,m}(\xi). \tag{2.42}$$

The resulting construction is described by Algorithm 3 with the above selection of ξ_{m+1}. Note that this strategy is closely related with [42], where the authors propose a greedy construction of a parameter-independent test space for Petrov-Galerkin projection, with a selection of basis functions based on an error indicator similar to $\delta_{r,m}(\xi)$.

4.3 Recycling factorizations of operator's evaluations - Application to reduced basis method

When using a sample-based approach for solving a parameter-dependent equation (2.27), the linear system is solved for many values of the parameter ξ. When using a direct solver for solving a linear system for a given ξ, a factorization of the operator is usually available and can be used for improving a preconditioner for the solution of subsequent linear systems.

We here describe this idea in the particular context of greedy algorithms for Reduced Basis method, where the interpolation points $\xi_1, \ldots, \xi_r$ for the interpolation of the inverse $A(\xi)^{-1}$ are taken as the evaluation points $\xi_1^{RB}, \ldots, \xi_r^{RB}$ for the solution. At iteration r, having a preconditioner $P_r(\xi)$ and an approximation $u_r(\xi)$, a new

interpolation point is defined such that

$$\xi_{r+1}^{RB} \in \underset{\xi \in \Xi}{\operatorname{argmax}} \, \|P_r(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V.$$

Algorithm 4 describes this strategy.

Algorithm 4 Reduced Basis method with recycling of operator factorizations.

Require: $A(\xi), b(\xi)$, Θ, and R.

1: Initialize $u_0(\xi) = 0$, $P_0(\xi) = I$
2: **for** $r = 0$ to $R - 1$ **do**
3: Find $\xi_{r+1}^{RB} \in \arg\max_{\xi \in \Xi} \|P_r(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V$
4: Compute a factorization of $A(\xi_{r+1}^{RB})$
5: Solve the linear system $v_{r+1} = A(\xi_{r+1}^{RB})^{-1} b(\xi_{r+1}^{RB})$
6: Update the approximation subspace $V_{r+1} = V_r + \operatorname{span}(v_{r+1})$
7: Define the implicit operator $P_{r+1} = A(\xi_{r+1}^{RB})^{-1}$
8: Update the space Y_{r+1} (or Y_{r+1}^{+})
9: Compute the preconditioner : $P_{r+1}(\xi) = \underset{P \in Y_{r+1} (\text{or } Y_{r+1}^{+})}{\operatorname{argmin}} \|(I - PA(\xi))\Theta\|_F$
10: Compute the Petrov-Galerkin approximation $u_{r+1}(\xi)$ of $u(\xi)$ on V_{r+1} using equation (2.30)
11: **end for**
12: **return** Approximation $u_R(\xi)$.

5 Numerical results

5.1 Illustration on a one parameter-dependent model

In this section we compare the different interpolation methods on the following one parameter-dependent advection-diffusion-reaction equation:

$$-\Delta u + v(\xi) \cdot \nabla u + u = f$$

defined over a square domain $\Omega = [0, 1]^2$ with periodic boundary conditions. The advection vector field $v(\xi)$ is spatially constant and depends on the parameter ξ that takes values in $[0, 1]$: $v(\xi) = D\cos(2\pi\xi)e_1 + D\sin(2\pi\xi)e_2$, with $D = 50$ and (e_1, e_2) the canonical basis of $\mathbb{R}^2$. Ξ denotes a uniform grid of 250 points on $[0, 1]$. The source term f is represented in Figure 2.2(a). We introduce a finite element approximation space of dimension $n = 1600$ with piecewise linear approximations

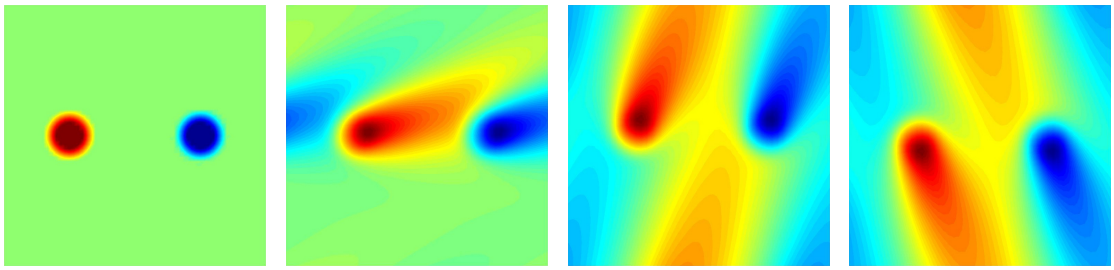
on a regular mesh of Ω. A Galerkin projection on this approximation space yields the linear system of equations $A(\xi)u(\xi) = b$, with

$$A(\xi) = A_0 + \cos(2\pi\xi)A_1 + \sin(2\pi\xi)A_2,$$

where the matrices A_0, A_1, A_2 and the vector b are given by

$$(A_0)_{i,j} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j + \phi_i\phi_j \ , \ (A_1)_{i,j} = \int_\Omega \phi_i(e_1 \cdot \nabla\phi_j)$$
$$(A_2)_{i,j} = \int_\Omega \phi_i(e_2 \cdot \nabla\phi_j) \ , \ (b)_i = \int_\Omega \phi_i f,$$

where $\{\phi_i\}_{i=1}^n$ is the basis of the finite element space. Figures 2.2(b), 2.2(c) and 2.2(d) show three samples of the solution.



(a) f (b) $u(0.05)$ (c) $u(0.2)$ (d) $u(0.8)$

Figure 2.2: Plot of the source term f (a) and 3 samples of the solution corresponding to parameter values $\xi = 0.05$ (b), $\xi = 0.2$ (c) and $\xi = 0.8$ (d) respectively.

5.1.1 Comparison of the interpolation strategies

We first choose arbitrarily 3 interpolation points ($\xi_1 = 0.05$, $\xi_2 = 0.2$ and $\xi_3 = 0.8$) and show the benefits of using the Frobenius norm projection for the definition of the preconditioner. For the comparison, we consider the Shepard and the nearest neighbor interpolation strategies. Let $\|\cdot\|_\Xi$ denote a norm on the parameter set Ξ. The Shepard interpolation method is an inverse weighted distance interpolation:

$$\lambda_i(\xi) = \begin{cases} \dfrac{\|\xi - \xi_i\|_\Xi^{-s}}{\sum_{j=1}^m \|\xi - \xi_j\|_\Xi^{-s}} & \text{if } \xi \neq \xi_i \\ 1 & \text{if } \xi = \xi_i \end{cases},$$

where $s > 0$ is a parameter. Here we take $s = 2$. The nearest neighbor interpolation method consists in choosing the value taken by the nearest interpolation point, that

means $\lambda_i(\xi) = 1$ for some $i \in \arg\min_j \|\xi - \xi_j\|_\Xi$, and $\lambda_j(\xi) = 0$ for all $j \neq i$.

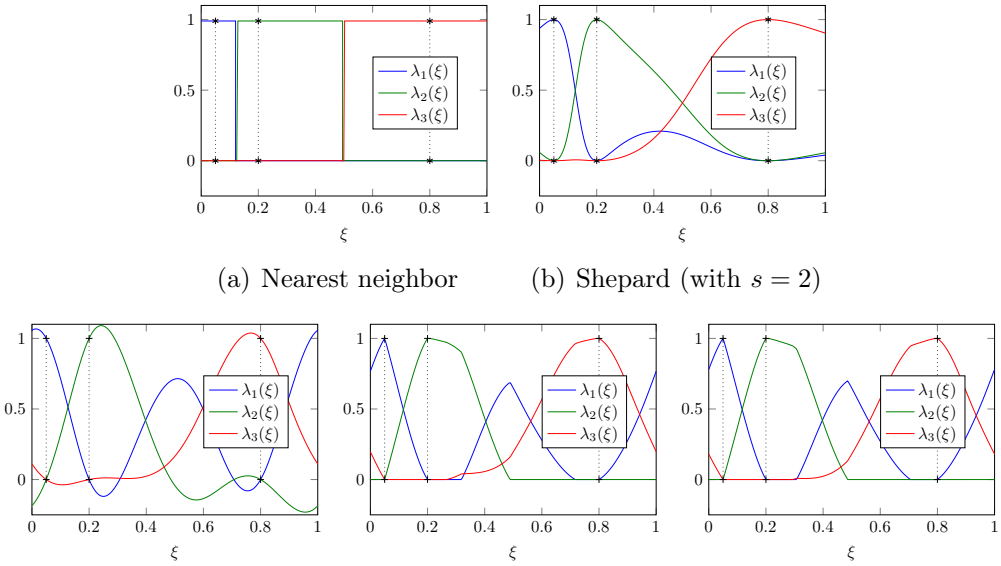
Concerning the Frobenius norm projection on Y_m (or Y_m^+), we first construct the affine decomposition of $M(\xi)$ and $S(\xi)$ as explained in Section 2.4. The interpolation points ξ_k^* (resp. $\widetilde{\xi}_k^*$) given by the EIM procedure for $M(\xi)$ (resp. $S(\xi)$) are $\{0.0; 0.25; 0.37; 0.56; 0.80\}$ (resp. $\{0.0; 0.25; 0.62\}$). The number of terms $m_M = 5$ in the resulting affine decomposition of $M(\xi)$ (see equation (2.26)) is less than the expected number $m_A^2 = 9$ (see equation (2.25a)). Considering the functions $\Phi_1(\xi) = 1$, $\Phi_2(\xi) = \cos(2\pi\xi)$, $\Phi_3(\xi) = \sin(2\pi\xi)$, and thanks to relation $\cos^2 = 1 - \sin^2$, the space

$$\mathrm{span}_{i,j}\{\Phi_i\Phi_j\} = \mathrm{span}\{1, \cos, \sin, \cos\sin, \cos^2, \sin^2\} = \mathrm{span}\{1, \cos, \sin, \cos\sin, \cos^2\}$$

is of dimension $m_M = 5$. The EIM procedure automatically detects the redundancy in the set of functions and reduces the number of terms in the decomposition (2.26). Then, since the dimension n of the discretization space is reasonable, we compute the matrices $M(\xi_k^*)$ and the vectors $S(\widetilde{\xi}_k^*)$ using equation (2.26).

The functions $\lambda_i(\xi)$ are plotted on Figure 2.3 for the proposed interpolation strategies. It is important to note that contrary to the Shepard or the nearest neighbor method, the Frobenius norm projection (on Y_m or Y_m^+) leads to periodic interpolation functions, *i.e.* $\lambda_i(\xi = 0) = \lambda_i(\xi = 1)$. This is consistent with the fact that the application $\xi \mapsto A(\xi)$ is 1-periodic. The Frobenius norm projection automatically detects such a feature.

Figure 2.4 shows the condition number $\kappa_m(\xi)$ of $P_m(\xi)A(\xi)$ with respect to ξ. We first note that for the constant preconditioner $P_1(\xi) = A(\xi_2)^{-1}$, the resulting condition number is higher than the one of the non preconditioned matrix $A(\xi)$ for $\xi \in [0.55; 0.95]$. We also note that the interpolation strategies based on the Frobenius norm projection lead to better preconditioners than the Shepard and nearest neighbor interpolation strategies. When considering the projection on Y_m^+ and $Y_m^{\bar{\kappa}}$ (with $\bar{\kappa} = 5 \times 10^4$ such that (2.22) holds), the resulting condition number is roughly the same, so as the interpolation functions of Figures 2.3(d) and 2.3(e). Since the projection on $Y_m^{\bar{\kappa}}$ requires the expensive computation of the constants γ^+, γ^- and C (see Section 2.3), we prefer to simply use the projection on Y_m^+ in order to ensure the preconditioner to be invertible. Finally, for this example, it is not necessary to impose any constraint since the projection on Y_m leads to the best preconditioner and this preconditioner appears to be invertible for any $\xi \in \Xi$.



(a) Nearest neighbor (b) Shepard (with $s = 2$)

(c) Projection on Y_m (d) Projection on Y_m^+ (e) Projection on $Y_m^{\bar{\kappa}}$

Figure 2.3: Interpolation functions $\lambda_i(\xi)$ for different interpolation methods.

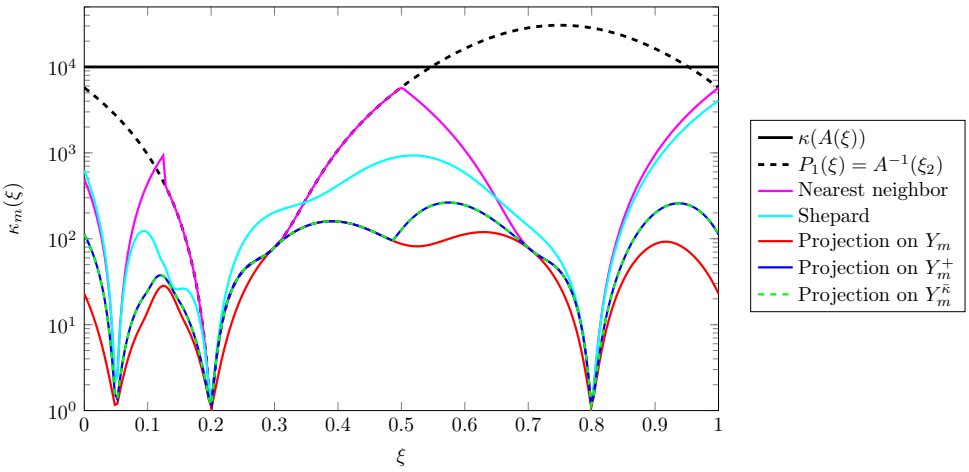


Figure 2.4: Condition number of $P_m(\xi)A(\xi)$ for different interpolation strategies. The condition number of $A(\xi)$ is given as a reference.

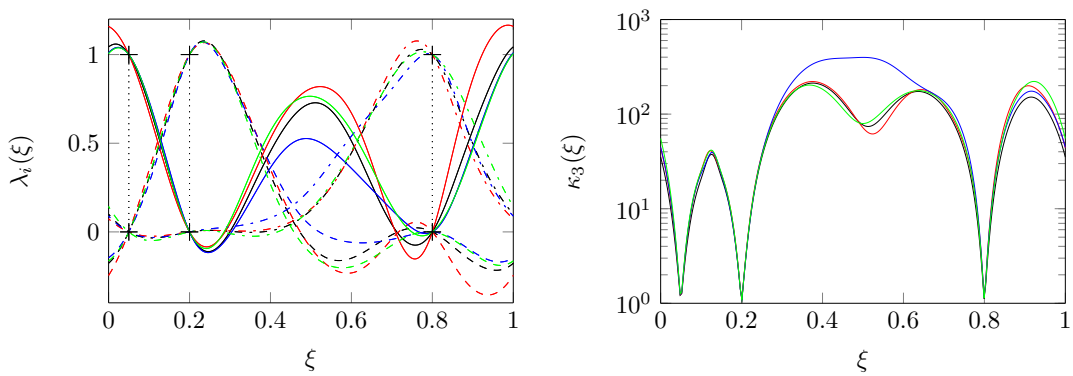
5.1.2 Using the Frobenius semi-norm

We analyze now the interpolation method defined by the Frobenius semi-norm projection on Y_m (2.5) for the different definitions of $\Theta \in \mathbb{R}^{n\times K}$ proposed in sections 2.2.2 and 2.2.1. According to Table 2.2, the error on the interpolation functions decreases slowly with K (roughly as $\mathcal{O}(K^{-1/2})$), and the use of the P-SRHT matrix leads to a slightly lower error. The interpolation functions are plotted on Figure

2.5(a) in the case where $K = 8$. Even if we have an error of 36% to 101% on the interpolation functions, the condition number given on Figure 2.5(b) remains close to the one computed with the Frobenius norm. Also, an important remark is that with $K = 8$ the computational effort for computing $M^{\Theta}(\xi_k^*)$ and $S^{\Theta}(\widetilde{\xi_k^*})$ is negligible compared to the one for $M(\xi_k^*)$ and $S(\widetilde{\xi_k^*})$.

K	8	16	32	64	128	256	512
Rescaled partial Hadamard	0.4131	0.3918	0.3221	0.1010	0.0573	0.0181	0.0255
Rescaled Rademacher (1)	0.5518	0.0973	0.2031	0.1046	0.1224	0.1111	0.0596
Rescaled Rademacher (2)	1.0120	0.6480	0.1683	0.1239	0.0597	0.0989	0.0514
Rescaled Rademacher (3)	0.7193	0.2014	0.1241	0.1051	0.1235	0.1369	0.0519
P-SRHT (1)	0.4343	0.2081	0.2297	0.0741	0.0723	0.0669	0.0114
P-SRHT (2)	0.3624	0.2753	0.0931	0.1285	0.0622	0.0619	0.0249
P-SRHT (3)	0.8133	0.4227	0.1138	0.0741	0.0824	0.0469	0.0197

Table 2.2: Relative error $\sup_\xi \|\lambda(\xi) - \lambda^{\Theta}(\xi)\|_{\mathbb{R}^3} / \sup_\xi \|\lambda(\xi)\|_{\mathbb{R}^3}$: $\lambda^{\Theta}(\xi)$ (resp. $\lambda(\xi)$) are the interpolation functions associated to the Frobenius semi-norm projection (resp. the Frobenius norm projection) on Y_m, with Θ either the rescaled partial Hadamard matrix, the random rescaled Rademacher matrix or the P-SRHT matrix (3 different samples for random matrices).



(a) Interpolation functions.

(b) Condition number of $P_3(\xi)A(\xi)$.

Figure 2.5: Comparison between the Frobenius norm projection on Y_3 (black lines) and the Frobenius semi-norm projection on Y_3, using for Θ either a sample of the rescaled Rademacher matrix (blue lines), the rescaled partial Hadamard matrix (red lines) or a sample of the P-SRHT matrix (green lines) with $K = 8$.

5.1.3 Greedy selection of the interpolation points

We now consider the greedy selection of the interpolation points presented in Section 4. We start with an initial point $\xi_1 = 0$ and the next points are defined by (2.38), where matrix Θ is a realization of the P-SRHT matrix with $K = 128$ columns. $P_m(\xi)$ is the projection on Y_m using the Frobenius semi-norm defined by (2.5). The first 3 steps of the algorithm are illustrated on Figure 2.6. We observe that at each iteration, the new interpolation point ξ_{m+1} is close to the point where the condition number of $P_m(\xi)A(\xi)$ is maximal. Table 2.3 presents the maximal value over $\xi \in \Xi$ of the residual, and of the condition number of $P_m(\xi)A(\xi)$. Both quantities are rapidly decreasing with m. This shows that this algorithm, initially designed to minimize $\|(I - P_m(\xi)A(\xi))\Theta\|_F$, seems to be also efficient for the construction of preconditioners, in the sense that the condition number decreases rapidly.

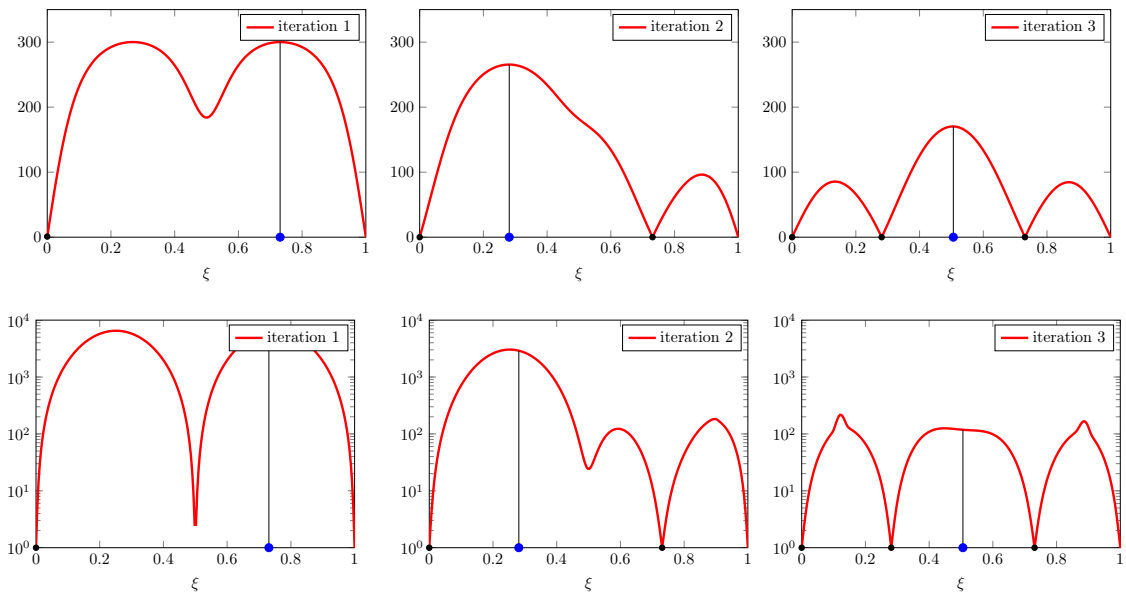


Figure 2.6: Greedy selection of the interpolation points: the first row is the residual $\|(I - P_k(\xi)A(\xi))\Theta\|_F$ (the blue points correspond to the maximum of the residual) with Θ a realization of the P-SRHT matrix with $K = 128$ columns, and the second row is the condition number of $P_m(\xi)A(\xi)$.

5.2 Multi-parameter-dependent equation

We introduce a benchmark proposed within the OPUS project (see http://www.opus-project.fr). Two electronic components Ω_{IC} (see Figure 2.7) submitted to a cooling

iteration m	0	1	2	5	10	20	30
$\sup_\xi \kappa(P_m(\xi)A(\xi))$	10001	6501	3037	165,7	51,6	16,7	7,3
$\sup_\xi \|(I - P_m(\xi)A(\xi))\Theta\|_F$	-	300	265	80,5	35,4	10,0	7,6

Table 2.3: Convergence of the greedy algorithm: supremum over $\xi \in \Xi$ of the condition number (first row) and of the Frobenius semi-norm residual (second row).

air flow in the domain Ω_{Air} are fixed on a printed circuit board Ω_{PCB}. The temperature field defined over $\Omega = \Omega_{IC} \cup \Omega_{PBC} \cup \Omega_{Air} \subset \mathbb{R}^2$ satisfies the advection-diffusion equation:

$$-\nabla \cdot (\kappa(\xi)\nabla u) + D(\xi)v \cdot \nabla u = f. \tag{2.43}$$

The diffusion coefficient $\kappa(\xi)$ is equal to κ_{PCB} on Ω_{PCB}, κ_{Air} on Ω_{Air} and κ_{IC} on Ω_{IC}. The right hand side f is equal to $Q = 10^6$ on Ω_{IC} and 0 elsewhere. The boundary conditions are $u = 0$ on Γ_d, $e_2 \cdot \nabla u = 0$ on Γ_u (e_1, e_2 are the canonical vectors of $\mathbb{R}^2$), and $u_{|\Gamma_l} = u_{|\Gamma_r}$ (periodic boundary condition). At the interface $\Gamma_C = \partial\Omega_{IC} \cap \partial\Omega_{PCB}$ there is a thermal contact conductance, meaning that the temperature field u admits a jump over Γ_C which satisfies

$$\kappa_{IC}(e_1 \cdot \nabla u_{|\Omega_{IC}}) = \kappa_{PCB}(e_1 \cdot \nabla u_{|\Omega_{PCB}}) = r(u_{|\Omega_{IC}} - u_{|\Omega_{PCB}}) \quad \text{on } \Gamma_C.$$

The advection field v is given by $v(x, y) = e_2 g(x)$, where $g(x) = 0$ if $x \leq e_{PCB} + e_{IC}$ and

$$g(x) = \frac{3}{2(e - e_{IC})}\left(1 - \left(\frac{2x - (e + e_{IC} + 2e_{PCB})}{e - e_{IC}}\right)^2\right)$$

otherwise.

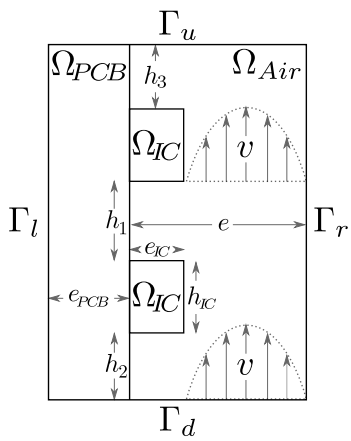
We have 4 parameters: the width $e := \xi_1$ of the domain Ω_{Air}, the thermal conductance parameter $r := \xi_2$, the diffusion coefficient $\kappa_{IC} := \xi_3$ of the components and the amplitude of the advection field $D := \xi_4$. Since the domain $\Omega = \Omega(e)$ depends on the parameter $\xi_1 \in [e_{min}, e_{max}]$, we introduce a geometric transformation $(x, y) = \phi_{\xi_1}(x_0, y_0)$ that maps a reference domain $\Omega_0 = \Omega(e_{max})$ to $\Omega(\xi_1)$:

$$\phi_{\xi_1}(x_0, y_0) = \left(\begin{array}{c} \left\{\begin{array}{cc} x_0 & \text{if } x_0 \leq e_0 \\ e_0 + (x_0 - e_0)\frac{\xi_1 - e_{IC}}{e_{max} - e_{IC}} & \text{otherwise.} \end{array}\right\} \\ y_0 \end{array}\right),$$

with $e_0 = e_{PCB} + e_{IC}$. This method is described in [114]: since the geometric transformation ϕ_{ξ_1} satisfies the so-called *Affine Geometry Precondition*, the operator of

equation (2.43) formulated on the reference domain admits an affine representation.

For the spatial discretization we use a finite element approximation with $n = 2.8 \times 10^4$ degrees of freedom (piecewise linear approximation). We rely on a Galerkin method with SUPG stabilization (see [24]). Ξ is a set of 10^4 independent samples drawn according the loguniform probability laws of the parameters given on Figure 2.7.



Geometry (m)	
$e = \xi_1$	$\sim \log\mathcal{U}(2.5 \times 10^{-3}, 5 \times 10^{-2})$
e_{PCB}	2×10^{-3}
e_{IC}	2×10^{-3}
h_{IC}	2×10^{-2}
h_1	3×10^{-2}
h_2	2×10^{-2}
h_3	4×10^{-2}
Thermal conductance ($Wm^{-2}K^{-1}$)	
$r = \xi_2$	$\sim \log\mathcal{U}(1 \times 10^{-1}, 1 \times 10^{2})$
Diffusion coefficients ($Wm^{-1}K^{-1}$)	
κ_{PCB}	2×10^{-1}
κ_{Air}	3×10^{-2}
$\kappa_{IC} = \xi_3$	$\sim \log\mathcal{U}(2 \times 10^{-1}, 1.5 \times 10^{2})$
Advection field ($Wm^{-2}K^{-1}$)	
$D = \xi_4$	$\sim \log\mathcal{U}(5 \times 10^{-4}, 1 \times 10^{-2})$

Figure 2.7: Geometry and parameters of the benchmark OPUS.

5.2.1 Preconditioner for the projection on a given reduced space

We consider here a POD basis V_r of dimension $r = 50$ computed with 100 snapshots of the solution (a basis of V_r is obtained by the first 50 dominant singular vectors of a matrix of 100 random snapshots of $u(\xi)$). Then we compute the Petrov-Galerkin projection as presented in Section 3.1. The efficiency of the preconditioner can be measured with the quantity $\delta_{r,m}(\xi)$: the associated quasi-optimality constant $(1 - \delta_{r,m}(\xi)^2)^{-1/2}$ should be as close to one as possible (see equation (2.33)). We introduce the quantile q_p of probability p associated to the quasi-optimality constant $(1 - \delta_{r,m}(\xi)^2)^{-1/2}$ defined as the smallest value $q_p \geq 1$ satisfying

$$\mathbb{P}\big(\{\xi \in \Xi : (1 - \delta_{r,m}(\xi)^2)^{-1/2} \leq q_p\}\big) \geq p,$$

where $\mathbb{P}(A) = \#A/\#\Xi$ for $A \subset \Xi$. Table 2.4 shows the evolution of the quantile with respect to the number of interpolation points for the preconditioner. Here the goal is to compare the different strategies for the selection of the interpolation points:

(a) the greedy selection (2.42) based on the quantity $\delta_{r,m}(\xi)$,

(b) the greedy selection (2.38) based on the Frobenius semi-norm residual, with Θ a P-SRHT matrix with $K = 256$ columns, and

(c) a random Latin Hypercube sample (LHS).

The projection on Y_m (or Y_m^+) is then defined with the Frobenius semi-norm using for Θ a P-SRHT matrix with $K = 330$ columns.

When considering a small number of interpolation points $m \leq 3$, the projection on Y_m^+ provides lower quantiles for the quasi-optimality constant compared to the projection on Y_m. The positivity constraint is useful for small m. But for high values of m (see $m = 15$) the positivity constraint is no longer necessary and the projection on Y_m provides lower quantiles.

Concerning the choice of the interpolation points, the strategy (a) shows the faster decay of the quantiles q_p, especially for $p = 50\%$. The strategy (b) shows also good results, but the quantile q_p for $p = 100\%$ are still high compared to (a). These results show the benefits of the greedy selection based on the quasi-optimality constant. Finally the strategy (c) shows bad results (high values of the quantiles), especially for small m.

5.2.2 Preconditioner for Reduced Basis method

We now consider the preconditioned Reduced Basis method for the construction of the approximation space V_r, as presented in Section 3.2. Figures 2.9 and 2.10 show the convergence of the error with respect to the rank r of $u_r(\xi)$ for different constructions of the preconditioner $P_m(\xi)$. Two measures of the error are given: $\sup_{\xi\in\Xi} \|u(\xi) - u_r(\xi)\|_V / \|u(\xi)\|_V$, and the quantile of probability 0.97 for $\|u(\xi) - u_r(\xi)\|_V / \|u(\xi)\|_V$. The curve "Ideal greedy" corresponds to the algorithm defined by (2.34) which provides a reference for the ideally conditioned algorithm, *i.e.* with $\kappa_m(\xi) = 1$. Figure 2.8 shows the corresponding first interpolation points for the solution.

The greedy selection of the interpolation points based on (2.38) (see Figure 2.9) allows to almost recover the convergence curve of the ideal greedy algorithm when using the projection on Y_m with $m = 15$. For the recycling strategy, the approximation is rather bad for $r = m \leq 10$ meaning that the space Y_r (or Y_r^+) is not really adapted for the construction of a good preconditioner over the whole parametric domain. However, for higher values of r, the preconditioner is getting better and

	Projection on Y_m								
	Greedy selection based on						(c) Latin Hypercube sampling		
	(a) $\delta_{r,m}(\xi)$			(b) Frob. residual					
	50%	90%	100%	50%	90%	100%	50%	90%	100%
$m = 0$	21.3	64.1	94.1	21.3	64.1	94.1	21.3	64.1	94.1
$m = 1$	18.3	74.1	286.7	10.2	36.1	161.6	18.3	104.1	231.8
$m = 2$	11.9	22.6	42.1	9.8	53.3	374.0	11.5	113.0	533.9
$m = 3$	11.1	49.2	200.4	7.8	31.2	60.2	18.3	138.7	738.5
$m = 5$	5.2	10.8	18.4	6.8	18.6	24.5	8.7	121.1	651.4
$m = 10$	3.1	9.0	13.2	5.3	22.3	62.1	4.0	21.6	345.7
$m = 15$	2.2	6.3	10.4	3.5	6.5	11.5	2.7	7.8	48.6

	Projection on Y_m^+								
	Greedy selection based on						(c) Latin Hypercube sampling		
	(a) $\delta_{r,m}(\xi)$			(b) Frob. residual					
	50%	90%	100%	50%	90%	100%	50%	90%	100%
$m = 0$	21.3	64.1	94.1	21.3	64.1	94.1	21.3	64.1	94.1
$m = 1$	18.3	74.1	286.7	10.2	36.1	161.6	18.3	104.1	231.8
$m = 2$	11.9	22.6	42.1	8.9	35.5	78.6	10.4	41.5	112.5
$m = 3$	9.7	24.4	48.0	7.9	27.7	57.9	12.1	48.8	114.1
$m = 5$	6.4	15.0	25.5	6.9	26.8	65.1	5.7	11.6	17.5
$m = 10$	4.6	9.5	16.8	7.3	18.9	38.0	4.3	10.0	18.5
$m = 15$	4.3	7.1	11.2	6.4	10.1	18.0	4.2	9.0	19.3

Table 2.4: Quantiles q_p of the quasi-optimality constant associated to the Petrov-Galerkin projection on the POD subspace V_r for $p = 50\%$, 90% and 100%. The row $m = 0$ corresponds to $P_0(\xi) = R_V^{-1}$, that is the standard Galerkin projection.

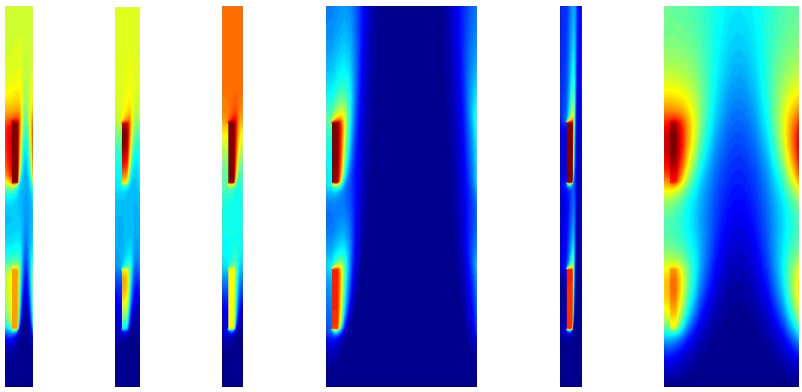
better. For $r \geq 20$, we almost reach the convergence of the ideal greedy algorithm. We conclude that this recycling strategy, with a computational cost which is comparable to the standard non preconditioned Reduced Basis greedy algorithm, allows obtaining asymptotically the performance of the ideal greedy algorithm. Note that the positivity constraint yields a better preconditioner for small values of r but is no longer necessary for large r.

Let us finally consider the effectivity index

$$\eta_r(\xi) = \|P_r(\xi)(A(\xi)u_r(\xi) - b(\xi))\|_V / \|u(\xi) - u_r(\xi)\|_V,$$

which evaluates the quality of the preconditioned residual norm for error estimation. We introduce the confidence interval $I_r(p)$ defined as the smallest interval which satisfies

$$\mathbb{P}(\{\xi \in \Xi \ : \ \eta_r(\xi) \in I_r(p)\}) \geq p.$$



(a) (b) (c) (d) (e) (f)

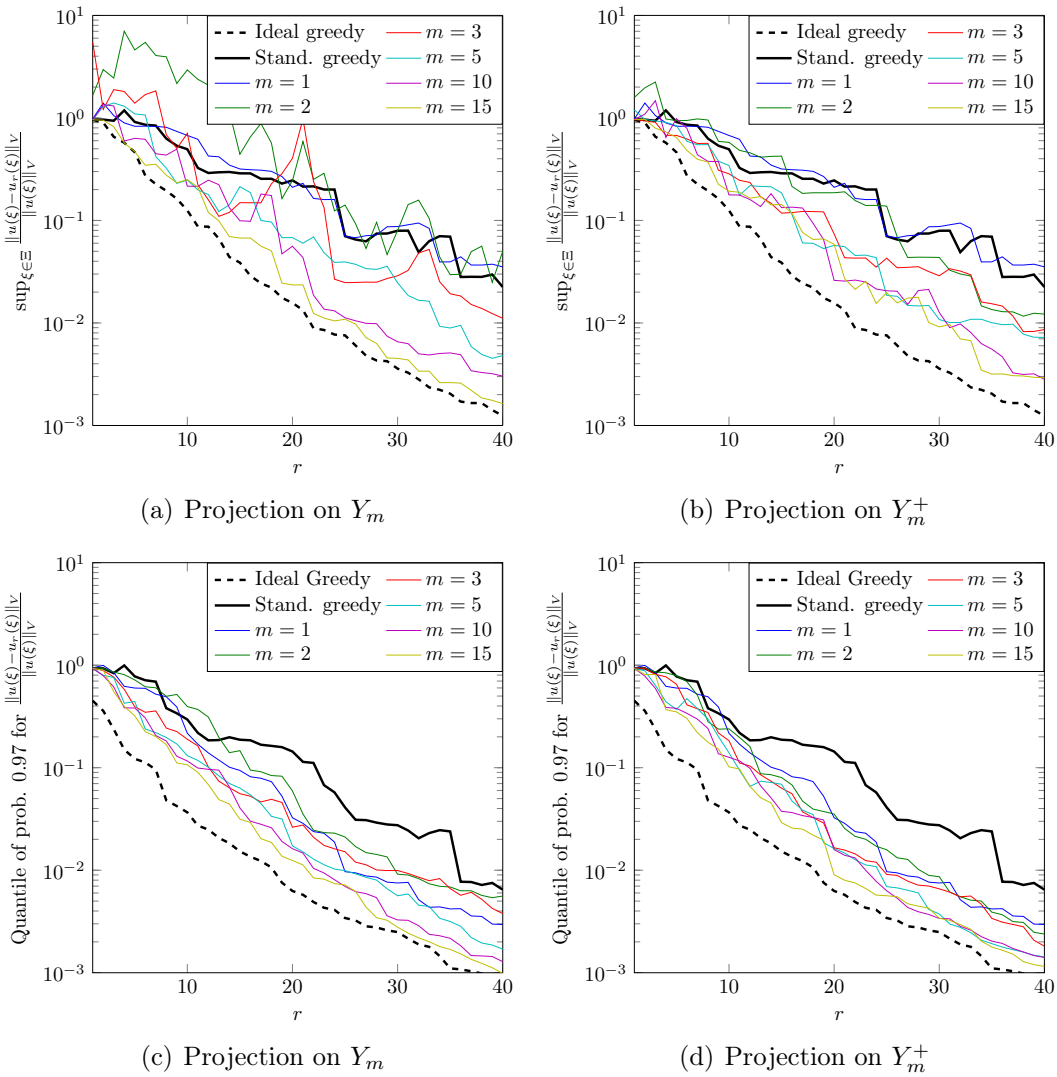
	(a) $r = 1$	(b) $r = 2$	(c) $r = 3$	(d) $r = 4$	(e) $r = 5$	(f) $r = 6$
e	$7.1 \cdot 10^{-3}$	$6.1 \cdot 10^{-3}$	$5.0 \cdot 10^{-3}$	$5.0 \cdot 10^{-2}$	$5.1 \cdot 10^{-3}$	$4.4 \cdot 10^{-2}$
r	$9.9 \cdot 10^{1}$	$2.3 \cdot 10^{0}$	$4.1 \cdot 10^{-1}$	$2.8 \cdot 10^{0}$	$8.6 \cdot 10^{-1}$	$4.8 \cdot 10^{1}$
κ_{IC}	$1.1 \cdot 10^{2}$	$2.1 \cdot 10^{-1}$	$3.2 \cdot 10^{1}$	$6.0 \cdot 10^{0}$	$1.1 \cdot 10^{2}$	$3.2 \cdot 10^{-1}$
D	$1.7 \cdot 10^{-3}$	$7.4 \cdot 10^{-4}$	$6.2 \cdot 10^{-4}$	$9.9 \cdot 10^{-3}$	$8.6 \cdot 10^{-3}$	$6.0 \cdot 10^{-4}$

Figure 2.8: First six interpolation points of the ideal reduced basis method and corresponding reduced basis functions.

On Figure 2.11 we see that the confidence intervals are shrinking around 1 when r increases, meaning that the preconditioned residual norm becomes a better and better error estimator when r increases. Again, the positivity constraint is needed for small values of r, but we obtain a better error estimation without imposing this constraint for $r \geq 20$. On the contrary, the standard residual norm leads to effectivity indices that spread from 10^{-1} to 10^{1} with no improvement as r increases, meaning that we can have a factor 10^2 between the error estimator $\|A(\xi)u_r(\xi) - b(\xi)\|_{V'}$ and the true error $\|u_r(\xi) - u(\xi)\|_V$.

5.2.3 Preconditioner for iterative solvers

In this section, we show the benefits of using the preconditioner for iterative linear solvers. Here, there is no model order reduction. For all ξ in a sample set Ξ_t of cardinality $t = 10^4$, we solve the linear system $A(\xi)u(\xi) = b(\xi)$ by two iterative solvers: the generalized minimal residual method (GMRES, [116]), and the conjugate gradient squared (CGS, [118]). These solvers are well adapted since the operator $A(\xi)$ is non-symmetric. In order to improve the convergence, we precondition the system to the left with the preconditioner $P_m(\xi)$ defined by the Frobenius semi-norm projection (where Θ is a realization of a P-SRHT matrix with $K = 330$ columns) either on Y_m or on Y_m^+. The interpolation points are defined by the greedy procedure de-



(a) Projection on Y_m

(b) Projection on Y_m^+

(c) Projection on Y_m

(d) Projection on Y_m^+

Figure 2.9: Convergence of the preconditioned reduced basis method using the greedy selection of interpolation points for the preconditioner. Supremum over Ξ (top) and quantile of probability 97% (bottom) of the relative error $\|u(\xi) - u_r(\xi)\|_V / \|u(\xi)\|_V$ with respect to r. Comparison of preconditioned reduced basis algorithms with ideal and standard greedy algorithms.

scribed by algorithm 3: the results are given on Figure 2.12. For the comparison, we also consider a Latin Hypercube Sample for the interpolation points, see Figure 2.13.

We see that the number of iterations for reaching a given precision is decreasing with the number of interpolation points m for the preconditioner. This illustrates the fact that the preconditoner $P_m(\xi)$ becomes a uniformly good approximation of

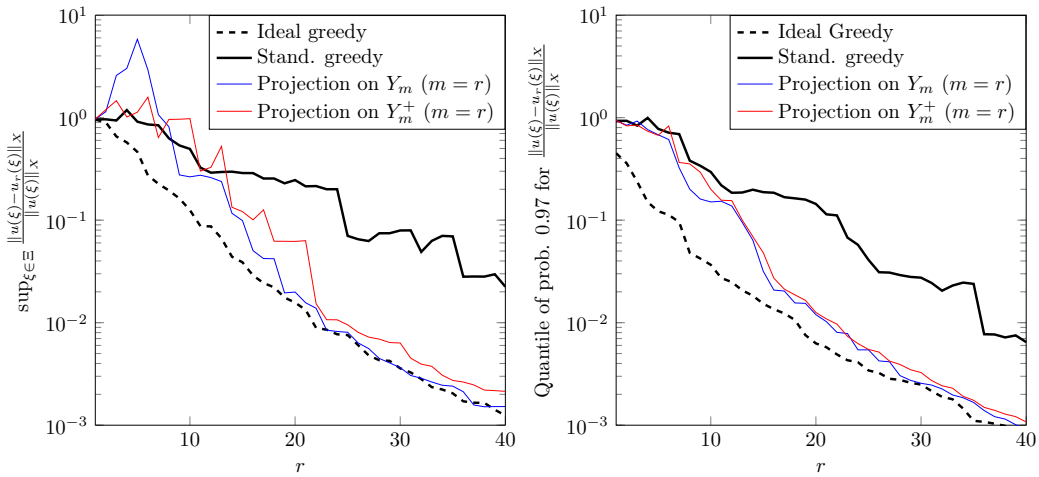
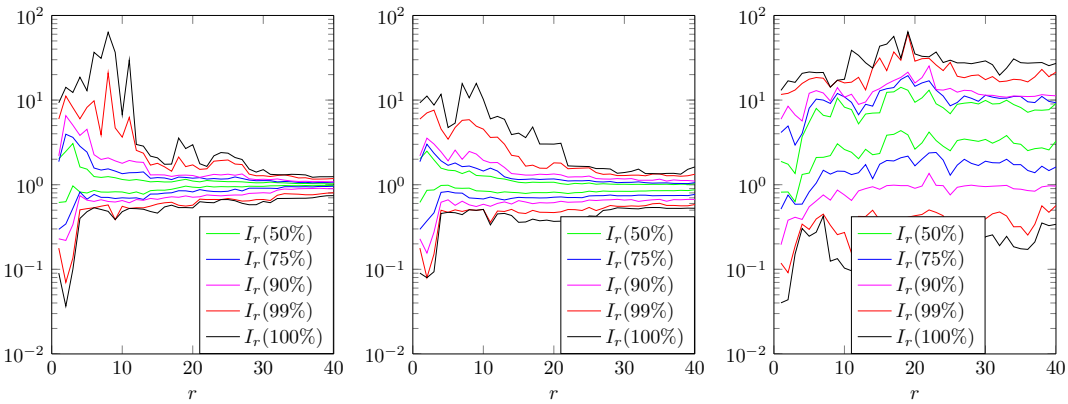


Figure 2.10: Preconditioned Reduced basis methods with recycling. Supremum over Ξ (left) and quantile of probability 97% (right) of the relative error $\|u(\xi) - u_r(\xi)\|_V / \|u(\xi)\|_V$ with respect to r. Comparison of preconditioned reduced basis algorithms with ideal and standard greedy algorithms.

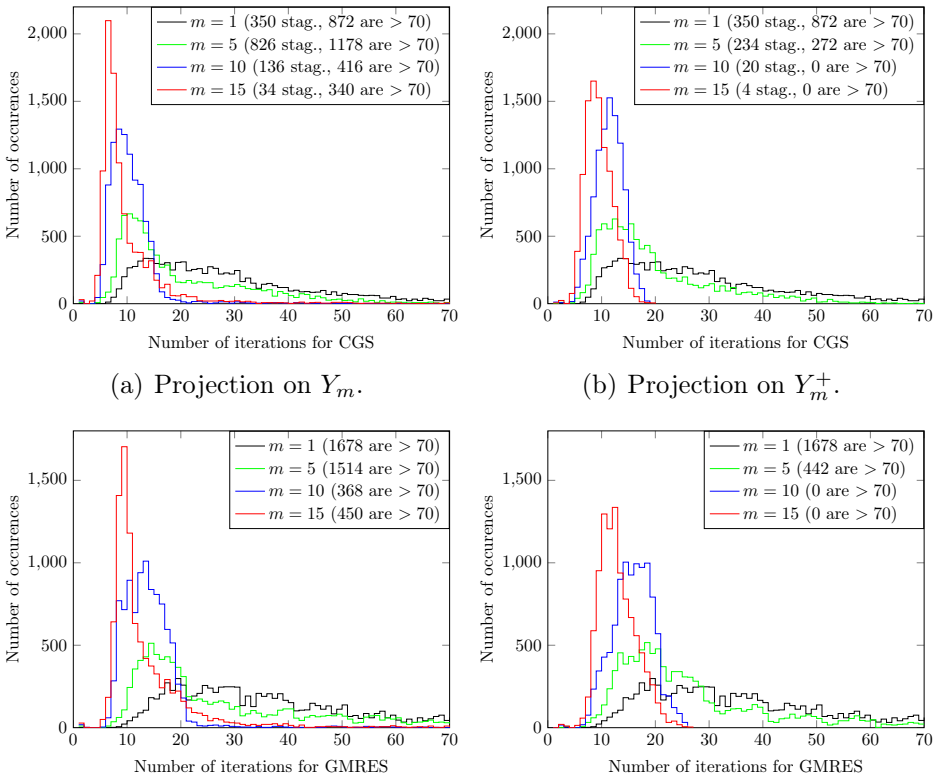


(a) Recycling, proj. on Y_m. (b) Recycling, proj. on Y_m^+. (c) Dual residual norm.

Figure 2.11: Confidence intervals of the effectivity index during the iterations of the Reduced Basis greedy construction. Comparison between preconditioned algorithms with recycling of operators factorizations (a,b) and the non preconditioned greedy algorithm (c).

$A(\xi)^{-1}$ when adding interpolation points. We observe that for $m \geq 10$, the number of iterations is slightly smaller for the greedy selection of the interpolation points compared to the Latin Hypercube Sample. This shows that the greedy selection yields a uniformly better preconditioner, but also that the Latin Hypercube Sample is a fairly good strategy. Let us also note that for $m = 5$ the preconditioner is more

efficient with the Latin Hypercube Sample selection of interpolation points. This shows that the greedy selection is not the optimal way to choose the interpolation points. Finally when using the positivity constraint (projection on Y_m^+), the number of sample $\xi \in \Xi_t$ for which the iterative algorithms stagnate is significantly smaller. In that context, this constraint is particularly relevant.



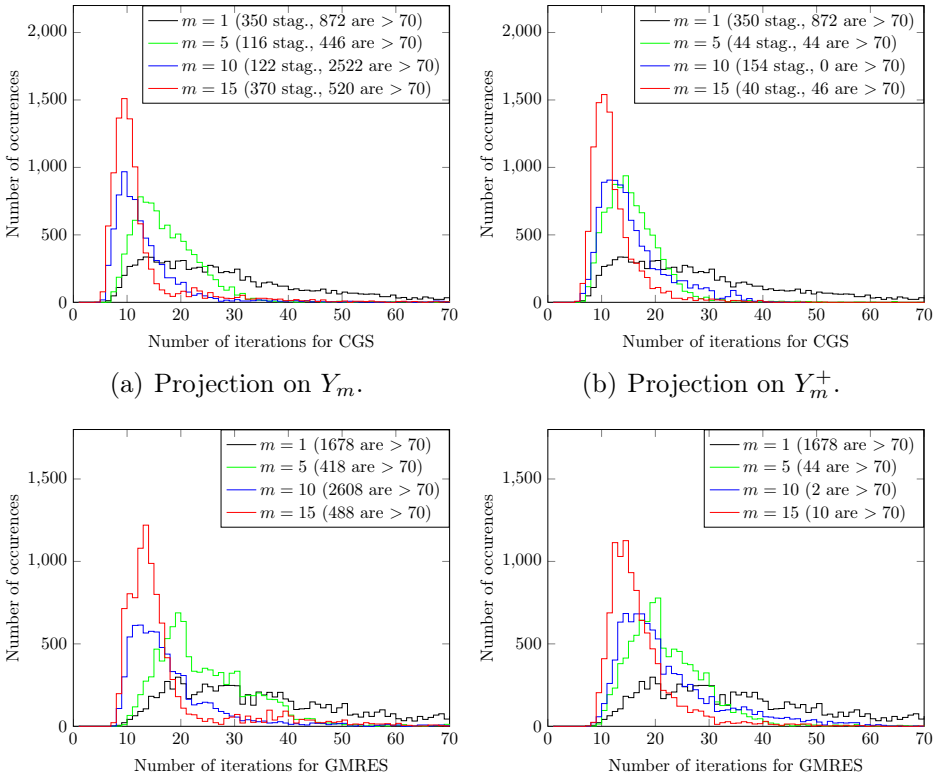
(a) Projection on Y_m. (b) Projection on Y_m^+.

(c) Projection on Y_m. (d) Projection on Y_m^+.

Figure 2.12: Histogram of the number of iterations of the iterative solver (first row: CGS, second row: GMRES) to reach the tolerance 10^{-8}. The interpolation points of the preconditioner are defined by the greedy algorithm 3. The preconditioner is either defined as the projection on Y_m or on Y_m^+ (with positivity constraints)

6 Conclusion

We have proposed a method for the interpolation of the inverse of a parameter-dependent matrix. The interpolation is defined by the projection of the identity in the sense of the Frobenius norm. Approximations of the Frobenius norm (based on the use of Hadamard matrices or random matrices) have been introduced to make



(a) Projection on Y_m. (b) Projection on Y_m^+.

(c) Projection on Y_m. (d) Projection on Y_m^+.

Figure 2.13: Histogram of the number of iterations of the iterative solver (first row: CGS, second row: GMRES) to reach the tolerance 10^{-8}. The interpolation points of the preconditioner are given by a Latin Hypercube sample. The preconditioner is either defined as the projection on Y_m or on Y_m^+ (with positivity constraints)

computationally feasible the projection in the case of large matrices.

Then, different strategies have been proposed for the selection of interpolation points depending of the objective: (i) the construction of an optimal approximation of the inverse operator for preconditioning iterative solvers or for improving error estimators based on preconditioned residuals, (ii) the improvement of the quality of Petrov-Galerkin projections of the solution of a parameter-dependent equation on a given reduced approximation space, or (iii) the recycling of operator factorizations when solving a parameter-dependent equation with a sample-based approach.

The performance of the obtained parameter-dependent preconditioners has been illustrated in the context of projection-based model reduction techniques such as

the Proper Orthogonal Decomposition and the Reduced Basis method, and also for the solution of linear system using iterative solvers.

In this chapter, we have restricted the presentation to the case of real matrices but the methodology can be naturally extended to the case of complex matrices.

Chapter 3

Projection-based model order reduction for estimating vector-valued variables of interest

This chapter focuses on the estimation of a variable of interest $s(\xi)$ that is a linear function of the solution $u(\xi)$ of a parameter-dependent equation. We propose and compare different projection-based methods that allow to consider functional-valued or vector-valued variables of interest. In particular we highlight the role played by three reduced spaces: the approximation space and the test space associated to the primal variable, and the approximation space associated to the dual variable. Then, in the spirit of the Reduced Basis method, we propose greedy algorithms for the construction of these reduced spaces.

Contents

1 Introduction 75

2 Analysis of different projection methods for the estimation of a variable of interest 76

2.1 Petrov-Galerkin projection 76

2.2 Primal-dual approach 79

2.3 Saddle point problem 82

3 Goal-oriented projections for parameter-dependent equations 86

3.1 Error estimates for vector-valued variables of interest 87

3.2 Greedy construction of the approximation spaces 89

4 Numerical results 91

4.1 Applications 91

4.2 Comparison of the projection methods 94

4.3 Greedy construction of the reduced spaces 100

5 Conclusion 104

1 Introduction

The main objective of the *Reduced basis method* is to construct a low dimensional approximation space, also called the *reduced space*, that can uniformly approximate the solution manifold $\mathcal{M} = \{u(\xi), \xi \in \Xi\}$ associated to a parameter-dependent equation. However, in many applications one is not interested in the solution itself but only in a variable of interest $s(\xi)$ which is a functional of the solution $u(\xi)$. Here we assume that $s(\xi)$ linearly depends on $u(\xi)$. When $s(\xi)$ takes scalar values, efficient methods exist to provide accurate estimation of the variable of interest. The idea is to construct an approximation of the so-called *dual solution* which is used to improve the estimation of $s(\xi)$. We refer to [110] for a general survey on primal-dual methods and to [75] for the application to the Reduced Basis method. In the present chapter, we extend this methodology to functional-valued or vector-valued variables of interest, *i.e.* when $s(\xi)$ belongs to a vector space. For example, this allows to consider variables of interest defined as the restriction of the solution on a part of the boundary, or to consider simultaneously multiple scalar variables of interest.

In section 2, we analyze different projection methods for computing approximations of the solution and of the variables of interest. In particular, we generalize the primal-dual method to vector-valued variables of interest, and we propose a new method based on a saddle point problem. We show that the error on the variable of interest depends on three reduced spaces: (a) the primal approximation space (*i.e.* the approximation space of the primal solution $u(\xi)$), (b) the primal test space that is used for the (Petrov-)Galerkin projection of $u(\xi)$, and (c) the dual approximation space for the solution of the dual problem. Section 3 is concerned with the construction of these reduced spaces. Following the methodology of the Reduced Basis method, we propose greedy algorithms for the spaces (a) and (c). For the test space (b), we propose to use the preconditioners introduced in the previous chapter (see also [127]). Finally in Section 4, numerical experiments illustrate the properties of the projection methods and the greedy algorithms. Of particular interest are the complexities of the *offline* phase, meaning the computational cost for the construction of the reduced spaces, and of the *online* phase, that is the cost to compute the estimation of $s(\xi)$.

2 Analysis of different projection methods for the estimation of a variable of interest

In this section, we omit the dependence on ξ for the sake of clarity. Let V and W be two Hilbert spaces. We consider the linear equation $Au = b$ where $A \in \mathcal{L}(V, W')$, and $b \in W'$. Let us introduce a variable of interest $s \in Z$ defined by $s = Lu$, where $L \in \mathcal{L}(V, Z)$ is a continuous linear operator and Z is a Hilbert space.

Let us introduce some notations. Any Hilbert space H is endowed with a norm $\|\cdot\|_H$ defined by the relation $\|\cdot\|_H^2 = \langle R_H \cdot, \cdot \rangle$, where the Riesz map $R_H \in \mathcal{L}(H, H')$ is a continuous symmetric positive definite operator and $\langle \cdot, \cdot \rangle$ is the duality pair. The dual space H' is endowed with the dual norm $\|\cdot\|_{H'}$ with the associated Riesz map $R_{H'} = R_H^{-1}$. Then the relations $\|v\|_H = \|R_V v\|_{H'}$ and $|\langle v, w \rangle| \leq \|v\|_H \|w\|_{H'}$ hold for any $v \in H$ and $w \in H'$. For any continuous operator C from a Hilbert space H_1 to another Hilbert space H_2, the notation C^* denotes the adjoint of C, such that $\langle Cv_1, v_2 \rangle = \langle v_1, C^* v_2 \rangle$ holds for any $v_1 \in H_1$ and $v_2 \in H_2$.

2.1 Petrov-Galerkin projection

Suppose that we are given a subspace $V_r \subset V$ of dimension r in which we seek an approximation of u. The orthogonal projection u_r^* of u on V_r, given by $u_r^* = \arg\min_{v \in V_r} \|u - v\|_V$, is characterized by

$$\langle u - u_r^*, R_V v \rangle = 0, \quad \forall v \in V_r. \tag{3.1}$$

Let us consider $u_r \in V_r$ defined by the following Petrov-Galerkin projection

$$\langle Au_r - b, y \rangle = 0 \quad \forall y \in W_r, \tag{3.2}$$

where $W_r \subset W$ is a test space of dimension r. The following proposition gives an error bound for the approximation of the variable of interest.

Proposition 2.1. *The solution u_r of equation (3.2) satisfies*

$$\|u - u_r\|_V \leq \frac{1}{\sqrt{1 - (\delta_{V_r, W_r})^2}} \min_{v \in V_r} \|u - v\|_V, \tag{3.3}$$

where

$$\delta_{V_r,W_r} = \max_{0\neq v\in V_r} \min_{y\in W_r} \frac{\|v - R_V^{-1}A^*y\|_V}{\|v\|_V} \tag{3.4}$$

is assumed to be strictly inferior to 1. Furthermore,

$$\|s - Lu_r\|_Z \leq \delta^L_{W_r}\|u - u_r\|_V, \tag{3.5}$$

with

$$\delta^L_{W_r} = \sup_{0\neq z'\in Z'} \min_{y\in W_r} \frac{\|L^*z' - A^*y\|_{V'}}{\|z'\|_{Z'}}. \tag{3.6}$$

Moreover, we have

$$\|s - Lu_r\|_Z \leq \frac{\delta^L_{W_r}}{\sqrt{1-(\delta_{V_r,W_r})^2}} \min_{v\in V_r}\|u - v\|_V. \tag{3.7}$$

Proof: We recall that u_r^* denotes the orthogonal projection of u on V_r. For any $v \in V_r$ and $y \in W_r$, we have

$$\begin{aligned}\langle u_r^* - u_r, R_V v\rangle &\overset{(3.1)}{=} \langle u - u_r, R_V v\rangle = \langle b - Au_r, A^{-*}R_V v\rangle \\ &\overset{(3.2)}{=} \langle b - Au_r, A^{-*}R_V v - y\rangle = \langle u - u_r, R_V v - A^*y\rangle \\ &\leq \|u - u_r\|_V\|R_V v - A^*y\|_{V'}.\end{aligned}$$

Taking the minimum over $y \in W_r$, dividing by $\|v\|_V$ and taking the maximum over $v \in V_r$, we obtain $\|u_r^* - u_r\|_V \leq \delta_{V_r,W_r}\|u - u_r\|_V$, where δ_{V_r,W_r} is defined by (3.4). Thanks to the orthogonality condition (3.1) we have $\|u - u_r\|_V^2 = \|u - u_r^*\|_V^2 + \|u_r^* - u_r\|_V^2$, which under the assumption $\delta_{V_r,W_r} < 1$ gives (3.3). Furthermore for any $z' \in Z'$ and $y \in W_r$, we have

$$\begin{aligned}\langle s - Lu_r, z'\rangle = \langle b - Au_r, A^{-*}L^*z'\rangle &\overset{(3.2)}{=} \langle b - Au_r, A^{-*}L^*z' - y\rangle \\ &\leq \|u - u_r\|_V\|L^*z' - A^*y\|_{V'}.\end{aligned}$$

Taking the minimum over $y \in W_r$, dividing by $\|z'\|_{Z'}$ and taking the supremum over $z' \in Z'$, we obtain (3.5). Finally, combining (3.3) and (3.5), we obtain (3.7). ■

The error bound (3.7) for the variable of interest is the product of three terms:

(a) $\min_{v\in V_r}\|u - v\|_V$ which suggests that the approximation space V_r should be defined such that u can be well approximated in V_r,

(b) $(1-(\delta_{V_r,W_r})^2)^{-1/2}$ which suggests that the test space W_r should be chosen such that any element of V_r can be well approximated by an element of $R_V^{-1}A^*W_r$, and

(c) $\delta^L_{W_r}$ which suggests that any element of range(L^*) should be well approximated by an element of A^*W_r.

As already noticed in [114, section 11.1], W_r plays a double role: a test space for the definition of u_r (point (b)) and an approximation space for the range of $A^{-*}L^*$ (point (c)). In the next section, we present the classical primal-dual approach used for the estimation of a variable of interest. We show that points (b) and (c) are treated separately.

Remark 2.2 (Comparison with the Céa's Lemma). *Let us assume that $V = W$ and that the operator A is continuous and coercive, meaning that there exist $\alpha > 0$ and $\beta < \infty$ such that $\|Av\|_{V'} \leq \beta\|v\|_V$ and $\langle Av, v\rangle \geq \alpha\|v\|_V^2$ hold for all $v \in V$. When considering the Galerkin projection, that is $V_r = W_r$, Céa's Lemma states that*

$$\|u - u_r\|_V \leq \frac{\beta}{\alpha} \min_{v\in V_r} \|u - v\|_V.$$

The inequality (3.3) is sharper than the above one. Indeed, for any $v, y \in V_r$ we have

$$\min_{\lambda\in\mathbb{R}} \|v - \lambda R_V^{-1}A^*y\|_V^2 = \|v\|_V^2 - \frac{\langle v, A^*y\rangle^2}{\|A^*y\|_V^2}.$$

Then, taking the minimum over $y \in V_r$ and dividing by $\|v\|_V^2$ we can write:

$$\min_{y\in V_r} \frac{\|v - R_V^{-1}A^*y\|_V^2}{\|v\|_V^2} \overset{y=v}{\leq} 1 - \frac{\langle Av, v\rangle^2}{\|A^*v\|_V^2\|v\|_V^2} \leq 1 - \frac{\alpha^2}{\beta^2}.$$

Then by definition (3.4) of δ_{V_r,W_r} we have $\delta_{V_r,W_r} \leq \sqrt{1-\alpha^2/\beta^2}$, that gives

$$\frac{1}{\sqrt{1-(\delta_{V_r,W_r})^2}} \leq \frac{\beta}{\alpha}.$$

Remark 2.3 (SPD and compliant case). *We suppose that A is symmetric positive definite (SPD). In that case we can set $V = W$, and the natural norm for the space V is the one induced by the operator with $R_V = A$. Then, the ideal test space is $W_r = V_r$. That yields $\delta_{V_r,W_r} = 0$, and $u_r = u_r^*$ corresponds to the standard Galerkin projection.*

Furthermore, if the variable of interest s is scalar-valued, we have $Z = \mathbb{R}$ and $\mathcal{L}(V, Z) = V'$. The compliant case corresponds to $Lv = \langle b, v\rangle$ for any $v \in V$. Then we have $\delta^L_{W_r} = \min_{v\in V_r} \|u - v\|_V = \|u - u_r\|_V$, and thanks to (3.7), we recover the so called "squared effect":

$$\|s - Lu_r\|_Z = |s - Lu_r| \leq \|u - u_r\|_V^2.$$

2.2 Primal-dual approach

We now extend the classical primal-dual approach for the estimation of a vector-valued variable of interest. Let us introduce the dual variable $Q \in \mathcal{L}(Z', W)$ defined by $A^*Q = L^*$. The relation

$$s = Lu = LA^{-1}b = (A^{-*}L^*)^*b = Q^*b$$

shows that the variable of interest can be exactly determined if either the primal variable u or the dual variable Q is known. The following proposition shows how to compute an estimation of the variable of interest provided approximations of both primal and dual variables are available. This proposition also contains an error analysis that is a generalization of the classical error bound for scalar-valued variables of interest (see [110]) to vector-valued variables of interest.

Proposition 2.4 (A generalization of a classical error bound). *Suppose that we dispose of approximations $\widetilde{u}$ of u and $\widetilde{Q}$ of Q. Then*

$$\widetilde{s} = L\widetilde{u} + \widetilde{Q}^*(b - A\widetilde{u}) \tag{3.8}$$

provides an approximation of s which satisfies

$$\|s - \widetilde{s}\|_Z \leq \|u - \widetilde{u}\|_V \|L^* - A^*\widetilde{Q}\|_{Z'\to V'}, \tag{3.9}$$

where

$$\|L^* - A^*\widetilde{Q}\|_{Z'\to V'} = \sup_{0\neq z'\in Z'} \frac{\|(L^* - A^*\widetilde{Q})z'\|_{V'}}{\|z'\|_{Z'}}. \tag{3.10}$$

Proof: For any $z' \in Z'$, we have

$$\begin{aligned}\langle s - \widetilde{s}, z'\rangle &= \langle Lu - L\widetilde{u} - \widetilde{Q}^*(b - A\widetilde{u}), z'\rangle = \langle (L - \widetilde{Q}^*A)(u - \widetilde{u}), z'\rangle \\ &= \langle u - \widetilde{u}, (L^* - A^*\widetilde{Q})z'\rangle \leq \|u - \widetilde{u}\|_V \|(L^* - A^*\widetilde{Q})z'\|_{V'}.\end{aligned}$$

Dividing by $\|z'\|_{Z'}$ and taking the supremum over $z' \in Z'$, we obtain (3.9). ∎

The approximation $\widetilde{u}$ can be defined as the Petrov-Galerkin projection u_r of u on a given approximation space V_r with a given test space W_r, see equation (3.2). For the approximation $\widetilde{Q}$ of $Q \in \mathcal{L}(Z', W)$, the bound (3.9) suggests that $\|L^* - A^*\widetilde{Q}\|_{Z' \to V'}$ should be small. We then propose to seek a solution of

$$\inf_{\widetilde{Q} \in \mathcal{L}(Z', W_k^Q)} \|L^* - A^*\widetilde{Q}\|_{Z' \to V'}, \tag{3.11}$$

where $W_k^Q \subset W$ is a given approximation space (different from W_r). The next proposition shows how to construct a solution of (3.11).

Proposition 2.5. *The linear operator $Q_k : Z' \to W_k^Q$ defined for $z' \in Z'$ by*

$$Q_k z' = \arg \min_{y_k \in W_k^Q} \|L^* z' - A^* y_k\|_{V'} \tag{3.12}$$

is a solution of (3.11). Moreover $Q_k z' \in W_k^Q$ is characterized by

$$\langle L^* z' - A^* Q_k z', R_V^{-1} A^* y_k \rangle = 0, \qquad \forall y_k \in W_k^Q. \tag{3.13}$$

Proof: We first note that Q_k defined by (3.12) is a linear operator in $\mathcal{L}(Z', W_k^Q)$, and equation (3.13) is the stationarity condition of the minimization problem (3.12). Furthermore for any $\widetilde{Q} \in \mathcal{L}(Z', W_k^Q)$ and $z' \in Z'$ we have

$$\frac{\|L^* z' - A^* Q_k z'\|_{V'}}{\|z'\|_{Z'}} \overset{(3.12)}{\leq} \frac{\|L^* z' - A^* \widetilde{Q} z'\|_{V'}}{\|z'\|_{Z'}} \leq \|L^* - A^* \widetilde{Q}\|_{Z' \to V'}.$$

Then, taking the supremum over $z' \in Z'$ and the infimum over $\widetilde{Q} \in \mathcal{L}(Z', W_k^Q)$, we obtain that $Q_k \in \mathcal{L}(Z', W_k^Q)$ is a solution of (3.11).

∎

In practice, for computing the approximation of the variable of interest (3.8) with $\widetilde{Q} = Q_k$, we only need $Q_k^*(b - Au_r)$. The following lemma shows how this can be performed without computing the operator Q_k.

Lemma 2.6. *Let Q_k be defined by (3.12). Then for $r = b - Au_r \in W'$,*

$$Q_k^* r = L R_V^{-1} A^* y_k^*, \tag{3.14}$$

where $y_k^* \in W_k^Q$ *is defined by*

$$\langle AR_V^{-1}A^*y_k^*, y_k\rangle = \langle r, y_k\rangle, \qquad \forall y_k \in W_k^Q. \tag{3.15}$$

Proof: For any $z' \in Z'$, since $Q_k z' \in W_k^Q$, we have

$$\langle Q_k z', AR_V^{-1}A^*y_k^*\rangle \overset{(3.15)}{=} \langle Q_k z', r\rangle. \tag{3.16}$$

Furthermore, by definition of Q_k we have

$$\langle Q_k z', AR_V^{-1}A^*y_k^*\rangle \overset{(3.13)}{=} \langle L^*z', R_V^{-1}A^*y_k^*\rangle. \tag{3.17}$$

Combining (3.16) and (3.17), we obtain $\langle z', Q_k^* r\rangle = \langle z', LR_V^{-1}A^*y_k^*\rangle$, which concludes the proof. ∎

We give now a sharper bound of the error on the variable of interest. The idea is to take advantage of the orthogonality relation (3.13).

Proposition 2.7. *The approximation* $\widetilde{s}$ *defined by* (3.8), *where* $\widetilde{u} = u_r$ *is the Petrov-Galerkin projection defined by* (3.2) *and* $\widetilde{Q} = Q_k$ *is defined by* (3.13), *satisfies*

$$\|s - \widetilde{s}\|_Z \leq \delta^L_{W_k^Q} \min_{y_k \in W_k^Q} \|u - u_r - R_V^{-1}A^*y_k\|_V, \tag{3.18}$$

where

$$\delta^L_{W_k^Q} = \sup_{0 \neq z' \in Z'} \min_{y \in W_k^Q} \frac{\|L^*z' - A^*y\|_{V'}}{\|z'\|_{Z'}}. \tag{3.19}$$

Moreover,

$$\|s - \widetilde{s}\|_Z \leq \frac{\delta^L_{W_k^Q}}{\sqrt{1 - (\delta_{V_r,W_r})^2}} \min_{v \in V_r} \|u - v\|_V. \tag{3.20}$$

Proof: For any $z' \in Z'$, and for any $y_k \in W_k^Q$ we have

$$\begin{aligned}\langle s - \widetilde{s}, z' \rangle &= \langle u - u_r, (L^* - A^* Q_k) z' \rangle \\ &\overset{(3.13)}{=} \langle u - u_r - R_V^{-1} A^* y_k, (L^* - A^* Q_k) z' \rangle \\ &\leq \|u - u_r - R_V^{-1} A^* y_k\|_V \|(L^* - A^* Q_k) z'\|_{V'}. \end{aligned} \tag{3.21}$$

Since Q_k satisfies (3.12), the last term of (3.21) becomes

$$\|(L^* - A^* Q_k) z'\|_{V'} = \min_{y_k \in W_k^Q} \|L^* z' - A^* y_k\|_{V'}.$$

Dividing by $\|z'\|_Z$ and taking the supremum over $z' \in Z'$ in (3.21), we obtain

$$\|s - \widetilde{s}\|_Z \leq \|u - u_r - R_V^{-1} A^* y_k\|_V \delta^L_{W_k^Q}$$

Then, taking the minimum over $y_k \in W_k^Q$, we obtain (3.18). Finally, taking $y_k = 0$ in (3.18) and thanks to (3.3), we obtain (3.20). ∎

2.3 Saddle point problem

In this section we extend the idea of [42] for the approximation of variables of interest. Let us define the Riesz map $R_W = A R_V^{-1} A^*$ for the norm over W, so that the relation $\|y\|_W = \|A^* y\|_V$ holds for any $y \in W$. The orthogonal projection u_r^* of u on V_r can be defined by

$$\begin{aligned}\|u - u_r^*\|_V &= \min_{v \in V_r} \|u - v\|_V \\ &= \min_{v \in V_r} \max_{w \in V} \frac{|\langle u - v, R_V w \rangle|}{\|w\|_V} \\ &= \min_{v \in V_r} \max_{w \in V} \frac{|\langle Av - b, A^{-*} R_V w \rangle|}{\|w\|_V} \\ &= \min_{v \in V_r} \max_{y \in W} \frac{|\langle Av - b, y \rangle|}{\|y\|_W}. \end{aligned}$$

We consider a space $W_p \subset W$ of dimension p. Replacing W by W_p, we obtain a saddle point problem

$$\min_{v \in V_r} \max_{y \in W_p} \frac{|\langle Av - b, y \rangle|}{\|y\|_W}. \tag{3.22}$$

The following proposition shows how to find a solution of this saddle point problem.

Proposition 2.8. *The solution* $(u_{r,p}, y_{r,p}) \in V_r \times W_p$ *of equations*

$$\langle R_W y_{r,p}, y\rangle + \langle A u_{r,p}, y\rangle = \langle b, y\rangle \quad \forall y \in W_p \tag{3.23}$$

$$\langle A^* y_{r,p}, v\rangle = 0 \quad \forall v \in V_r \tag{3.24}$$

is a solution of the saddle point problem (3.22).

Proof: Let $J(v, y) = |\langle Av - b, y\rangle|/\|y\|_W$. We will show that $J(u_{r,p}, y) \leq J(u_{r,p}, y_{r,p}) \leq J(v, y_{r,p})$ for any $y \in W_p$ and $v \in V_r$.

- (Left inequality) Let $r = b - Au_{r,p}$. The solution $\widetilde{y}_p \in W_p$ of the minimization problem $\min_{y\in W_p} \|R_W^{-1} r - y\|_W^2$ satisfies $\langle R_W \widetilde{y}_p - r, y\rangle = 0$ for any $y \in W_p$, that is (3.23). Then we have $y_{r,p} = \widetilde{y}_p$, and

$$\min_{y\in W_p} \|R_W^{-1} r - y\|_W^2 = \|R_W^{-1} r - y_{r,p}\|_W^2 = \|R_W^{-1} r\|_W^2 - \frac{\langle r, y_{r,p}\rangle^2}{\|y_{r,p}\|_W^2}. \tag{3.25}$$

  Furthermore, since W_p is a cone, we have for any $y \in W_p$

$$\min_{y\in W_p} \|R_W^{-1} r - y\|_W^2 \leq \min_{\alpha\in\mathbb{R}} \|R_W^{-1} r - \alpha y\|_W^2 = \|R_W^{-1} r\|_W^2 - \frac{\langle r, y\rangle^2}{\|y\|_W^2}. \tag{3.26}$$

  Combining (3.25) and (3.26), we obtain $J(u_{r,p}, y) \leq J(u_{r,p}, y_{r,p})$.

- (Right inequality) We simply note that thanks to (3.24), we have $|\langle Av - b, y_{r,p}\rangle| = |\langle b, y_{r,p}\rangle|$ for any $v \in V_r$. Then we can write $J(u_{r,p}, y_{r,p}) = J(v, y_{r,p})$ for all $v \in V_r$.

■

Equations (3.23)–(3.24) correspond to a linear system of size $(r + p)$. If $p < r$, the orthogonality condition (3.24) implies that $y_{r,p} = 0$, and the equation (3.23) turns out to be underdetermined. Then we need $p \geq r$.

The following proposition provides an error bound for the solution and for the variable of interest.

Proposition 2.9. *Let* $(u_{r,p}, y_{r,p}) \in V_r \times W_p$ *be the solution of* (3.23)–(3.24). *Then we have*

$$\|u - u_{r,p}\|_V \leq \frac{1}{\sqrt{1 - (\delta_{V_r,W_p})^2}} \min_{v\in V_r} \|u - v\|_V, \tag{3.27}$$

where

$$\delta_{V_r,W_p} = \max_{0\neq v\in V_r} \min_{y\in W_p} \frac{\|v - R_V^{-1}A^*y\|_V}{\|v\|_V}. \tag{3.28}$$

The quantity $\widetilde{s}$ *defined by*

$$\widetilde{s} = Lu_{r,p} + LR_V^{-1}A^*y_{r,p} \tag{3.29}$$

provides an approximation of s *such that*

$$\|s - \widetilde{s}\|_Z \leq \delta^L_{W_p}\|u - u_{r,p} - R_V^{-1}A^*y_{r,p}\|_V, \tag{3.30}$$

where

$$\delta^L_{W_p} = \sup_{0\neq z'\in Z'} \min_{y\in W_p} \frac{\|L^*z' - A^*y\|_{V'}}{\|z'\|_{Z'}}. \tag{3.31}$$

Also we have

$$\|s - \widetilde{s}\|_Z \leq \frac{\delta^L_{W_p}}{\sqrt{1-(\delta_{V_r,W_p})^2}} \min_{v\in V_r}\|u - v\|_V. \tag{3.32}$$

Proof: For any $v \in V_r$ and $y \in W_p$ we have

$$\begin{aligned}
\langle u_r^* - u_{r,p}, R_V v\rangle &\overset{(3.1)}{=} \langle u - u_{r,p}, R_V v\rangle = \langle b - Au_{r,p}, A^{-*}R_V v\rangle \\
&\overset{(3.23)}{=} \langle b - Au_{r,p}, A^{-*}R_V v - y\rangle + \langle R_W y_{r,p}, y\rangle \\
&\overset{(3.24)}{=} \langle b - Au_{r,p}, A^{-*}R_V v - y\rangle - \langle R_W y_{r,p}, A^{-*}R_V v - y\rangle \\
&= \langle u - u_{r,p} - R_V^{-1}A^*y_{r,p}, R_V v - A^*y\rangle \\
&\leq \|u - u_{r,p} - R_V^{-1}A^*y_{r,p}\|_V \|R_V v - A^*y\|_{V'}.
\end{aligned} \tag{3.33}$$

As mentioned in the proof of proposition 2.8, $\|R_W^{-1}(b - Au_{r,p}) - y_{r,p}\|_W = \min_{y\in W_p}\|R_W^{-1}(b-Au_{r,p})-y\|_W$. Then we have $\|R_W^{-1}(b-Au_{r,p})-y_{r,p}\|_W \leq \|R_W^{-1}(b-Au_{r,p})\|_W$, which by definition of the norm $\|\cdot\|_W$, gives

$$\|u - u_{r,p} - R_V^{-1}A^*y_{r,p}\|_V \leq \|u - u_{r,p}\|_V. \tag{3.34}$$

Using (3.34) in (3.33), taking the minimum over $y \in W_k$, dividing by $\|v\|_V$ and taking the maximum over $v \in V_r$, we obtain

$$\|u_r^* - u_{r,p}\|_V \leq \delta_{V_r,W_p}\|u - u_{r,p}\|. \tag{3.35}$$

Using (3.1), we obtain (3.27). Now, for any $z' \in Z'$ and $y \in W_p$, we have

$$\begin{aligned}\langle s - \widetilde{s}, z' \rangle &= \langle u - u_{r,p} - R_V^{-1} A^* y_{r,p}, L^* z' \rangle \\ &\overset{(3.23)}{=} \langle u - u_{r,p} - R_V^{-1} A^* y_{r,p}, L^* z' - A^* y \rangle \\ &\leq \| u - u_{r,p} - R_V^{-1} A^* y_{r,p} \|_V \| L^* z' - A^* y \|_{V'}.\end{aligned}$$

Taking the minimum over $y \in W_p$, dividing by $\|z'\|_{Z'}$ and taking the supremum over $z' \in Z'$, we obtain (3.30). Finally, thanks to (3.30), (3.34) and (3.27), we obtain (3.32). ∎

We observe that W_p plays a double role: a test space for the Petrov-Galerkin projection of the primal variable (see equation (3.27)) and an approximation space for the dual variable (see equation (3.32)). Then we will consider spaces of the form

$$W_p = W_r + W_k^Q. \tag{3.36}$$

This implies $\delta^L_{W_p} \leq \delta^L_{W_k^Q}$ and $\delta_{V_r,W_p} \leq \delta_{V_r,W_r}$, so that the bound for the variable of interest (3.32) is better than one of the primal-dual method (3.20). So we expect the approximation $u_{r,p}$ to be closer to the solution u compared to the Petrov-Galerkin projection u_r. Also, the approximation of the quantity of interest should be improved.

Remark 2.10 (SPD case). *Following remark 2.3, we consider the case where A is symmetric definite positive. Once again we choose $R_V = A$ and $W_r = V_r$. Thanks to (3.36), $V_r \subset W_p$ so that $\delta_{V_r,W_p} = 0$. Then $u_{r,p}$ is the Galerkin projection of u on V_r. Furthermore if we restrict y to $W_p \cap V_r^\perp$ in (3.23), then (3.23)–(3.24) imply that*

$$\begin{aligned}\langle A y_{r,p}, y \rangle &= \langle b - A u_{r,p}, y \rangle \qquad \forall y \in W_p \cap V_r^\perp, \\ \langle A y_{r,p}, v \rangle &= 0 = \langle b - A u_{r,p}, v \rangle \quad \forall v \in V_r.\end{aligned}$$

Since $V_r \subset W_p$ we have $W_p = V_r + (W_p \cap V_r^\perp)$. Then, summing the last two equations, we obtain $\langle A y_{r,p}, y \rangle = \langle b - A u_{r,p}, y \rangle$ for any $y \in W_p$. Let $t_{r,p} = y_{r,p} + u_{r,p} \in W_p$. Then $\langle A t_{r,p}, y \rangle = \langle b, y \rangle$ for any $y \in W_p$. This condition uniquely defines $t_{r,p}$. Furthermore, the approximation of the variable of interest (3.29) is given by $\widetilde{s} = L t_{r,p}$. Finally, the saddle point method for the SPD case can be simply interpreted as a Galerkin projection over the enriched space $V_r + W_k^Q$.

3 Goal-oriented projections for parameter-dependent equations

We now consider a parameter-dependent equation $A(\xi)u(\xi) = b(\xi)$ where ξ denotes a parameter taking values in a set $\Xi \subset \mathbb{R}^d$. The variable of interest is defined by $s(\xi) = L(\xi)u(\xi)$.

In the previous section, we presented different possibilities for the estimation of the variable of interest by means of projection methods that rely on three spaces: the primal approximation space V_r, the primal test space W_r and the dual approximation space W_k^Q (we recall that we introduce the space $W_p = W_r + W_k^Q$ for the saddle point method). Following the methodology of the Reduced Basis method, we will construct these spaces during the so-called *offline phase* such that for any parameter value $\xi \in \Xi$ the estimation $\widetilde{s}(\xi)$ can be rapidly computed during the *online phase*. For this last requirement, a key ingredient is that the spaces V_r, W_r and W_k^Q have low dimension: we then use the terminology *reduced spaces*.

We first address the problem of the construction of the test space W_r. Assuming that the primal approximation space V_r is given, we know from the previous section that W_r should be chosen such that δ_{V_r,W_r} is as close to zero as possible, see propositions 2.1, 2.7 and 2.9. When the operator $A(\xi)$ is symmetric positive definite (SPD), we set $W_r = V_r$, which is the optimal test space with respect to the norm induced by the operator, see remark 2.3. Otherwise, the optimal test space is $W_r = W_r(\xi) = A^{-*}(\xi)R_V V_r$, which is not feasible in practice since it requires the computation of $A^{-*}(\xi)v_r'$ for any $v_r' \in \{R_V v_v : v_r \in V_r\}$ and for any parameter $\xi \in \Xi$. Following the idea proposed in [127], we will consider a (parameter-dependent) test space of the form

$$W_r = W_r(\xi) = P_m(\xi)^* R_V V_r, \tag{3.37}$$

where $P_m(\xi)$ is an interpolation of the inverse of $A(\xi)$ using m interpolation points, that will be specified later on. The underlying idea is to obtain a test space as close as possible to the ideal test space. By convention $P_0(\xi) = R_V^{-1}$ yields the Galerkin projection, *i.e.* $W_r = V_r$.

Remark 3.1. *In the literature, $W_r = V_r$ (*i.e.* the Galerkin projection) is a common choice. However, this choice may lead to inaccurate projection of the primal variable when the operator is ill-conditioned. In the case of non coercive (or*

weakly coercive) operators, the test space is generally defined by $W_r = W_r(\xi) = R_V^{-1}A(\xi)V_r$ (here we have $W = V$), where $R_V^{-1}A(\xi)$ is called the "supremizer operator", see for example [115]. Note that this test space is parameter-dependent. The resulting Petrov-Galerkin projection $u_r(\xi)$ defined by (3.2) corresponds to the minimizer of the norm of the residual $\|A(\xi)v_r - b(\xi)\|_{V'}$ for $v_r \in V_r$. The common approach using the so-called "supremizer operator" for the definition of a Petrov-Galerkin projection is no more than a minimal residual method. Up to our knowledge, the only attempt to construct quasi-optimal test space for general operators can be found in [42]: the authors use the saddle point method presented in section 2.3, and construct a quasi-optimal test space W_p.

Remark 3.2. *In the literature, the choice $W_r = V_r$ is (almost) systematic. This may lead to inaccurate projection of the primal variable when the operator is ill-conditioned. Up to our knowledge, the only attempt to construct quasi-optimal test space for non SPD operator can be found in [42]: the authors use the saddle point method presented in section 2.3, and construct a quasi-optimal test space W_p.*

We discuss now the construction of the approximation spaces V_r and W_k^Q. In the literature, and for scalar-valued variables of interest, these reduced spaces are typically defined as the span of snapshots of the primal and dual solutions. These snapshots can be selected at random, using samples drawn from a certain probability measure over ξ, see *e.g.* [111]. Another popular method is to select the snapshots in a greedy way, in order to minimize the error $\|s(\xi) - \widetilde{s}(\xi)\|_Z$ uniformly over Ξ. This method requires an estimate of the error on the variable of interest. In the same spirit, we introduce error estimates for vector-valued variables of interest in section 3.1, and we propose greedy algorithms for the construction of V_r and W_k^Q in section 3.2.

3.1 Error estimates for vector-valued variables of interest

In this section, we propose practical error estimates for the variable of interest. The commonly used strategy is to start from the error bound

$$\|s(\xi) - \widetilde{s}(\xi)\|_Z \leq \|u(\xi) - \widetilde{u}(\xi)\|_V \|L(\xi)^* - A(\xi)^*\widetilde{Q}(\xi)\|_{Z'\to V'},$$

provided by proposition 2.4. This suggests to measure the norm of the residuals associated to the primal and dual variables. In practice, we distinguish two cases:

- In the case where the operator $A(\xi)$ is SPD, it is natural to choose the norm $\|\cdot\|_V$ as the one induced by the operator, *i.e.* $R_V = R_V(\xi) = A(\xi)$. Note that the norm $\|\cdot\|_V$ is then parameter-dependent. Therefore, neither the primal residual norm nor the dual residual norm can be computed without the knowledge of the primal and the dual solutions. The classical way to circumvent this issue is to introduce a parameter-independent norm $\|\cdot\|_{V_0}$, that is in general the "natural" norm associated to the space V, and to measure the residuals with this norm. We assume here that the operator $A(\xi)$ satisfies $\alpha(\xi)\|\cdot\|_{V_0} \leq \|A(\xi)\cdot\|_{V_0'}$, where $\alpha(\xi) > 0$ (for SPD operator, $\alpha(\xi)$ is nothing but the coercivity constant). Then by definition of the norm $\|\cdot\|_V$ we can write

$$\|v\|_V^2 = \langle A(\xi)v, v\rangle \leq \|A(\xi)v\|_{V_0'}\|v\|_{V_0} \leq \alpha(\xi)^{-1}\|A(\xi)v\|_{V_0'}^2 \qquad \forall v \in V.$$

  Then we have $\|u(\xi) - \widetilde{u}(\xi)\|_V \leq \alpha(\xi)^{-1/2}\|A(\xi)\widetilde{u}(\xi) - b(\xi)\|_{V_0'}$. The same trick can be used for the dual variable, leading to $\|L(\xi)^* - A(\xi)^*\widetilde{Q}(\xi)\|_{Z'\to V'} \leq \alpha(\xi)^{-1/2}\|L(\xi)^* - A(\xi)^*\widetilde{Q}(\xi)\|_{Z'\to V_0'}$. Then we obtain

$$\|s(\xi) - \widetilde{s}(\xi)\|_Z \leq \frac{\|A(\xi)\widetilde{u}(\xi) - b(\xi)\|_{V_0'}\|L(\xi)^* - A(\xi)^*\widetilde{Q}(\xi)\|_{Z'\to V_0'}}{\alpha(\xi)} =: \Delta(\xi) \tag{3.38}$$

  where $\Delta(\xi)$ is a certified error bound for the variable of interest.

- In the general case, the operator cannot be used for the definition of the norm $\|\cdot\|_V$. Then we consider the natural norm over V, *i.e.* $\|\cdot\|_V = \|\cdot\|_{V_0}$. As a consequence, the norm of the dual residual is computable, but the computation of the error $\|u(\xi) - \widetilde{u}(\xi)\|_{V_0}$ requires the knowledge of the primal solution $u(\xi)$, which is not feasible in practice. Once again, we assume that the operator satisfies $\alpha(\xi)\|\cdot\|_{V_0} \leq \|A(\xi)\cdot\|_{V_0'}$ so that we can write $\|u(\xi) - \widetilde{u}(\xi)\|_{V_0} \leq \alpha(\xi)^{-1}\|A(\xi)\widetilde{u}(\xi) - b(\xi)\|_{V_0'}$. Then we naturally end up with the same bound (3.38) for the variable of interest.

We now derive new error bounds in the case where $\widetilde{s}(\xi)$ is provided by the saddle point method. Let us start from the error bound

$$\|s(\xi) - \widetilde{s}(\xi)\|_Z \leq \sup_{0\neq z'\in Z'} \min_{y\in W_p} \frac{\|L(\xi)^*z' - A(\xi)^*y\|_{V'}}{\|z'\|_{Z'}}\|u(\xi) - u_{r,p}(\xi) - R_V^{-1}A(\xi)^*y_{r,p}(\xi)\|_V$$

provided by proposition 2.9. Once again, we distinguish two cases:

- For the case where the operator $A(\xi)$ is SPD, we consider for the norm $\|\cdot\|_V$ the operator norm. According to remark 2.10, the quantity $t_{r,p}(\xi) = u_{r,p}(\xi) -$

$R_V^{-1}(\xi)A(\xi)^* y_{r,p}(\xi) = u_{r,p}(\xi) + y_{r,p}(\xi)$ is nothing but the Galerkin projection of $u(\xi)$ onto the space $W_p = W_r + W_k^Q$, with $W_r = V_r$. Then for any $\widetilde{t}_{r,p} \in W_p$ we can write

$$\|u(\xi) - t_{r,p}(\xi)\|_V^2 \leq \|u(\xi) - \widetilde{t}_{r,p}\|_V^2 \leq \alpha(\xi)^{-1}\|A(\xi)\widetilde{t}_{r,p} - b(\xi)\|_{V_0'}^2,$$

where the norm $\|\cdot\|_{V_0}$ is the natural norm of V. Then, taking the minimum over $\widetilde{t}_{r,p} \in W_p$ we obtain $\|u(\xi) - t_{r,p}(\xi)\|_V \leq \alpha(\xi)^{-1/2} \min_{\widetilde{t}_{r,p}\in W_p} \|A(\xi)\widetilde{t}_{r,p} - b(\xi)\|_{V_0'}$. The same methodology can be used for the dual variable. We then obtain the following error bound:

$$\|s(\xi) - \widetilde{s}(\xi)\|_Z \leq \frac{1}{\alpha(\xi)} \sup_{0\neq z'\in Z'} \min_{y\in W_p} \frac{\|L(\xi)^* z' - A(\xi)^* y\|_{V_0'}}{\|z'\|_{Z'}} \min_{\widetilde{t}_{r,p}\in W_p} \|A(\xi)\widetilde{t}_{r,p} - b(\xi)\|_{V_0'} =: \Delta(\xi). \tag{3.39}$$

Note that the main difference between this error estimate and the previous one (3.38) is the minimization problem over W_p in both primal and dual residuals: this leads to additional computational costs, but a sharper error bound will be obtained, as illustrated by the numerical examples in the next section.

- For the general case, we consider $\|\cdot\|_V = \|\cdot\|_{V_0}$. Once again, using the relation $\|\cdot\|_{V_0} \leq \alpha(\xi)^{-1}\|A(\xi)\cdot\|_{V_0'}$, we obtain the following error estimate:

$$\|s(\xi) - \widetilde{s}(\xi)\|_Z \leq \frac{1}{\alpha(\xi)} \sup_{0\neq z'\in Z'} \min_{y\in W_p} \frac{\|L(\xi)^* z' - A(\xi)^* y\|_{V_0'}}{\|z'\|_{Z'}} \|A(\xi)t_{r,p}(\xi) - b(\xi)\|_{V_0'} =: \Delta(\xi), \tag{3.40}$$

where $t_{r,p}(\xi) = u_{r,p}(\xi) + R_{V_0}^{-1}A(\xi)^* y_{r,p}(\xi)$.

All the proposed error estimates rely on the knowledge of $\alpha(\xi)$. In the case where $\alpha(\xi)$ can not be easily computed, we can replace it by a lower bound $\alpha^{LB}(\xi) \leq \alpha(\xi)$, that is for example provided by a SCM procedure [83]. This option will not be considered here. Another option is to remove it from the definitions of $\Delta(\xi)$: in this case, the estimator is no longer certified.

3.2 Greedy construction of the approximation spaces

Here, we propose greedy algorithms for the construction of the reduced spaces V_r and W_k^Q. At each iteration, we look for the largest value of the error estimate $\Delta(\xi)$:

$$\xi^* \in \arg\max_{\xi\in\Xi} \Delta(\xi). \tag{3.41}$$

Then we can either *simultaneously* enrich the primal approximation space

$$V_{r+1} = V_r + \mathrm{span}(u(\xi^*)) \tag{3.42}$$

and the dual approximation space

$$W^Q_{k+\dim(Z)} = W^Q_k + \mathrm{range}(Q(\xi^*)), \tag{3.43}$$

or *alternatively* enrich W^Q_k and V_r.

Remark 3.3. *In the literature, and for scalar-valued variables of interest, the classical approaches are either a* separated *construction of V_r and W^Q_k (using two independent greedy algorithms, see for example [73, 114]), or a* simultaneous *construction, see* e.g. *[112]. The latter option can take advantage of a single factorization of the operator $A(\xi^*)$ to compute both the primal and dual variables. The* alternate *construction is not commonly used. This possibility is mentioned in remark 2.47 of the tutorial [75].*

Since we are considering vector-valued variables of interest, the dimension of $\mathrm{range}(Q(\xi^*))$ equals the dimension of the space Z, which is assumed to be equal to $l < \infty$. Then the enrichment (3.43) may lead to a rapid increase of the dimension of the space W^Q_k. Another option is to add only one vector at each iteration:

$$W^Q_{k+1} = W^Q_k + \mathrm{span}(Q(\xi^*)z') \tag{3.44}$$

where $z' \in Z'$ is such that:

$$z' \in \arg\max_{\widetilde{z}' \in Z'} \frac{\|(L(\xi^*)^* - A(\xi^*)^*\widetilde{Q}(\xi))\widetilde{z}'\|_{V_0'}}{\|\widetilde{z}'\|_{Z'}} \quad \text{(primal-dual method)}, \tag{3.45}$$

$$z' \in \arg\max_{\widetilde{z}' \in Z'} \min_{y \in W^Q_k} \frac{\|L(\xi^*)^*\widetilde{z}' - A(\xi^*)^* y\|_{V_0'}}{\|\widetilde{z}'\|_{Z'}} \quad \text{(saddle point method)}. \tag{3.46}$$

Contrarily to the full enrichment (3.43), this partial enrichment does not necessarily lead to a zero error at the point ξ^* for the next iterations. Then we expect that (3.44) will deteriorate the convergence properties of the algorithm, but W^Q_{k+1} will keep a low dimension, which is the essence of the reduced basis methods. It is worth to mention that in [42], the authors proposed the same kind of partial enrichment for the test space W_p (without considering any variable of interest).

For the definition (3.37) of the test space W_r, one needs to build the preconditioner $P_m(\xi)$ by interpolation of the inverse of $A(\xi)$. Following the idea proposed in [127] (see chapter 2), the interpolation points can be the ones where solutions (primal and dual) have been computed, *i.e.* the points given by (3.41). The resulting algorithms are summarized in Algorithm 5 and Algorithm 6 for the simultaneous and the alternate constructions of V_r and W^Q_k.

Algorithm 5 Simultaneous construction of V_r and W_k^Q

Require: Error estimator $\Delta(\cdot)$, a training set Ξ, maximum iteration I

Initialize the spaces $V_r = \{0\}$ and $W_k^Q = \{0\}$, and $r, k = 0$
for $i = 1$ to I **do**
  Find $\xi_i \in \arg\max_{\xi \in \Xi} \Delta(\xi)$
  Compute a factorization of $A(\xi_i)$ and update the preconditioner if needed
  Solve $u(\xi_i) = A(\xi_i)^{-1} b(\xi_i)$
  Update $V_{r+1} = V_r + \text{span}(u(\xi_i))$, $r \leftarrow r + 1$
  if Full dual enrichment **then**
    Solve $Q(\xi_i) = A(\xi_i)^{-*} L(\xi_i)^*$
    Update $W_{k+l}^Q = W_k^Q + \text{range}(Q(\xi_i))$, and $k \leftarrow k + l$
  else if Partial dual enrichment **then**
    Find z' according to (3.46) or (3.45)
    Solve $y(\xi_i) = A(\xi_i)^{-*}(L(\xi_i)^* z')$
    Update $W_{k+1}^Q = W_k^Q + \text{span}(y(\xi_i))$, and $k \leftarrow k + 1$
  end if
end for

4 Numerical results

This section is concerned with numerical applications of the methods proposed in Sections 2 and 3. After introducing two parameter-dependent problems, we compare the projection methods for the estimation of a variable of interest and then we study the behavior of the proposed greedy algorithms for the construction of the reduced spaces.

4.1 Applications

4.1.1 Application 1 : a symmetric positive definite problem

We consider a linear elasticity equation $\text{div}(K(\xi) : \varepsilon(\mathbf{u}(\xi))) = 0$ over a domain Ω that has the shape of a bridge, see Figure 3.1(a), where $\mathbf{u}(\xi) : \Omega \mapsto \mathbb{R}^3$ is the displacement field. The notation $\varepsilon(\mathbf{u}(\xi)) = \nabla^{sym}\mathbf{u}(\xi) \in \mathbb{R}^{3\times 3}$ corresponds to the infinitesimal strain tensor. The Hooke tensor $K(\xi)$ is such that

$$K(\xi) : \varepsilon(\mathbf{u}(\xi)) = \frac{E(\xi)}{1+\nu}\left(\varepsilon(\mathbf{u}(\xi)) + \frac{\nu}{1-2\nu}\text{trace}(\varepsilon(\mathbf{u}(\xi)))I_3\right),$$

where $\nu = 0.3$ is the Poisson coefficient and $E(\xi)$ is the Young modulus defined by $E(\xi) = 1_{\Omega_0} + \sum_{i=1}^{6} \xi_i 1_{\Omega_i}$, 1_{Ω_i} being the indicator function of the subdomain Ω_i,

Algorithm 6 Alternate construction of V_r and W_k^Q

Require: Error estimator $\Delta(\cdot)$, a training set Ξ, maximum iteration I

```
Initialize the spaces V_r = {0} and W_k^Q = {0}, and r, k = 0
for i = 1 to I do
    Find ξ_i ∈ arg max_{ξ∈Ξ} Δ(ξ)
    Compute a factorization of A(ξ_i) and update the preconditioner if needed
    if i is even then
        Solve u(ξ_i) = A(ξ_i)^{-1} b(ξ_i)
        Update V_{r+1} = V_r + span(u(ξ_i)), and r ← r + 1
    else if i is odd then
        if Full dual enrichment then
            Solve Q(ξ_i) = A(ξ_i)^{-*} L(ξ_i)^*
            Update W_{k+l}^Q = W_k^Q + range(Q(ξ_i)), and k ← k + l
        else if Partial dual enrichment then
            Find z' according to (3.46) or (3.45)
            Solve y(ξ_i) = A(ξ_i)^{-*}(L(ξ_i)^* z')
            Update W_{k+1}^Q = W_k^Q + span(y(ξ_i)), and k ← k + 1
        end if
    end if
end for
```

see Figure 3.1(b). The components of $\xi = (\xi_1, \cdots, \xi_6)$ are independent and log-uniformly distributed over $[10^{-1}, 10]$. We impose homogeneous Dirichlet boundary condition $\mathbf{u}(\xi) = 0$ on the red lines Γ_D, a unit vertical surface load on the green square Γ_{load}, and a zeros surface load on the complementary part of the boundary (see Figure 3.1(a)). We consider the Galerkin approximation $u^h(\xi)$ of $\mathbf{u}(\xi)$ on $V^h = \text{span}(\phi_i)_{i=1}^n \subset \{v \in H^1(\Omega)^3 : v_{|\partial\Omega_D} = 0\}$ that is a $\mathbb{P}_1$ finite element approximation space of dimension $n = 8916$ associated to the mesh given on Figure 3.1(b). The vector $u(\xi) \in V = \mathbb{R}^n$ such that $u^h(\xi) = \sum_{i=1}^n u_i(\xi)\phi_i$ is the solution of the linear system $A(\xi)u(\xi) = b$ of size n, with

$$A(\xi) = A^{(0)} + \sum_{k=1}^{6} \xi_i A^{(k)} \ , \quad A_{i,j}^{(k)} = \int_{\Omega_k} \nabla\phi_i : K_0 : \nabla\phi_j \ \mathrm{d}\Omega \ , \quad b_i = \int_{\Gamma_{\text{load}}} -e_3 \cdot \phi_i \ \mathrm{d}\Gamma, \tag{3.47}$$

where K_0 denotes the Hooke tensor with Poisson coefficient $\nu = 0.3$ and Young modulus $E = 1$. The operator norm $\|\cdot\|_V$ on the space V is given by

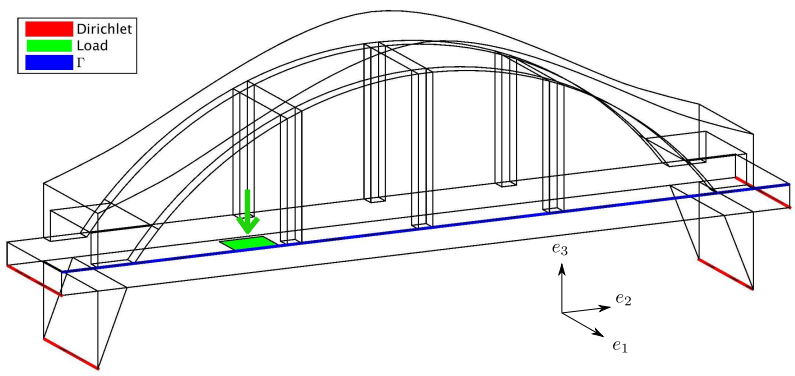
$$\|v\|_V^2 = \int_\Omega \varepsilon(v^h) : K(\xi) : \varepsilon(v^h) \mathrm{d}\Omega$$

for all $v \in V$, where $v^h(\xi) = \sum_{i=1}^n v_i(\xi)\phi_i$. We also consider the parameter-independent norm $\|\cdot\|_{V_0}$ defined as the operator norm for $\xi = (1, \cdots, 1)$. This corresponds to a Young modulus equals to 1 over Ω.

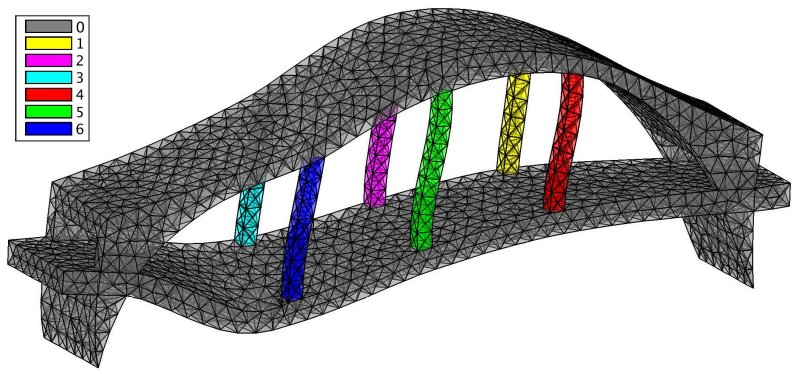
Let us consider $s^h(\xi) = u^h_{|\Gamma}(\xi) \cdot e_3$ that is the vertical displacement of the Galerkin approximation on the blue line Γ, see Figure 3.1(a). We can write $s^h(\xi) = \sum_{j=1}^l s_j(\xi)\psi_j$ where $\{\psi_j\}_{j=1}^l$ is a basis of the space $\{v^h_{|\Gamma} \cdot e_3,\ v^h \in V^h\}$ that is here of dimension $l = 44$. Then there exists $L \in \mathbb{R}^{l\times n}$ such that

$$s(\xi) = Lu(\xi)$$

where $s(\xi) = (s_1(\xi), \cdots, s_l(\xi)) \in Z = \mathbb{R}^l$ is the variable of interest. The norm $\|\cdot\|_Z$ is defined as the canonical norm of $\mathbb{R}^l$.



(a) Geometry, boundary condition and variable of interest.



(b) Realization of a solution and mesh of the domain Ω. The colors corresponds to the different sub-domains Ω_i for $i = 0, \cdots, 6$.

Figure 3.1: Application 1: schematic representation of the problem and a realization of the solution.

4.1.2 Application 2: a non symmetric problem

We consider the benchmark OPUS already presented in Chapter 2 (see section 5.2). The algebraic parameter-dependent equation $A(\xi)u(\xi) = b(\xi)$ corresponds to the finite element discretization of an advection-diffusion equation, where $\xi = (\xi_1, \cdots, \xi_4)$ is a random vector. The space $V = \mathbb{R}^n$ with $n = 2.8 \times 10^4$ is endowed with the norm $\|\cdot\|_V = \|\cdot\|_{V_0}$ that corresponds to the $H^1(\Omega)$-norm[1]. The variable of interest is the mean temperature of each components:

$$s_1(\xi) = \frac{1}{|\Omega_{IC_1}|}\int_{\Omega_{IC_1}} u^h(\xi)\mathrm{d}\Omega \quad , \quad s_2(\xi) = \frac{1}{|\Omega_{IC_2}|}\int_{\Omega_{IC_2}} u^h(\xi)\mathrm{d}\Omega. \tag{3.48}$$

Then we can write $s(\xi) = Lu(\xi)$ for an appropriate $L \in \mathbb{R}^{l\times n}$, with $l = 2$. Here we have $Z = \mathbb{R}^2$ and Z is equipped with its canonical norm.

4.2 Comparison of the projection methods

Here the goal is to compare the projection methods proposed in section 2 for the estimation of $s(\xi)$. Here the approximation spaces V_r, W_k^Q and the test space W_r are given. For the sake of simplicity, we assume now that V_r, W_k^Q and W_r are matrices containing the basis vectors of the corresponding subspace.

4.2.1 Application 1

We first detail how we build V_r, W_k^Q and W_r. The matrix V_r contains $r = 20$ snapshots of the solution: $V_r = (u(\xi_1), \cdots, u(\xi_{20}))$. The test space is $W_r = V_r$, which corresponds to the Galerkin projection. The matrix W_k^Q contains 2 snapshots of the dual variable $Q(\xi) = A(\xi)^{-1}L^T \in \mathbb{R}^{n\times l}$. Then $k = 2l = 88$. Finally, according to (3.36) the matrix $W_p = \big(W_r, W_k^Q\big)$ is the concatenation of the matrices W_r and W_k^Q.

We consider a training set $\Xi_t \subset \Xi$ of cardinality $t = 10^4$. For each $\xi \in \Xi_t$ we compute the exact quantity of interest $s(\xi)$ and the approximation $\widetilde{s}(\xi)$ by the following methods.

- *Primal only*: solve the linear system $\big(V_r^T A(\xi)V_r\big)U_r(\xi) = \big(V_r^T b\big)$ of size r and compute $\widetilde{s}(\xi) = \big(LV_r\big)U_r(\xi)$.
- *Dual only*: solve the linear system $\big((W_k^Q)^T A(\xi)W_k^Q\big)Y_k(\xi) = \big((W_k^Q)^T b\big)$ of size k and compute $\widetilde{s}(\xi) = \big(LW_k^Q\big)Y_k(\xi)$ (this method corresponds to the *primal-*

[1] meaning $\|v\|_{V_0} = \|v^h\|_{H^1(\Omega)}$ for all $v \in V$, where $v^h = \sum_{i=1}^n v_i\psi_i$

dual method where we removed the primal approximation, *i.e.* $V_r = W_r = \{0\}$)

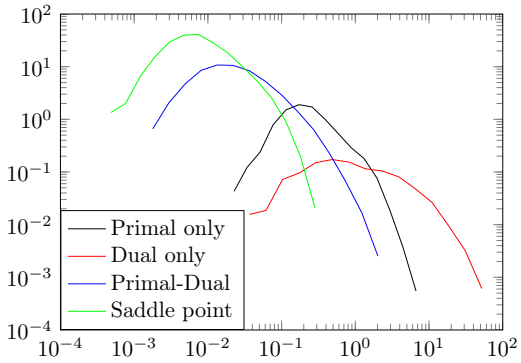
- *Primal-dual*: solve the linear system of the *Primal only* method, solve the linear system $\big((W_k^Q)^T A(\xi) W_k^Q\big) Y_k(\xi) = \big((W_k^Q)^T b\big) - \big((W_k^Q)^T A(\xi) V_r\big) U_r(\xi)$ of size k and compute $\widetilde{s}(\xi) = \big(LV_r\big) U_r(\xi) + \big(LW_k^Q\big) Y_k(\xi)$.

- *Saddle point*: using remark 2.10, solve the linear system $\big(W_p^T A(\xi) W_p\big) T_p(\xi) = \big(W_p^T b\big)$ of size $p = k + r$, and compute $\widetilde{s}(\xi) = \big(LW_p\big) T_p(\xi)$.

Thanks to relation (3.47), the matrix $A(\xi)$ admits an affine decomposition with respect to the parameter ξ. This allows the classical *offline/online* decomposition for a rapid assembling of the reduced systems for any parameter ξ.

Figure 3.2 gives the probability density function (PDF), the L^∞ norm and L^2 norm of the error $\|s(\xi) - \widetilde{s}(\xi)\|_Z$ estimated over the training set Ξ_t. We see that the *primal-dual* method provides errors for the quantity of interest that corresponds to the product of the errors of the *primal only* and the *dual only*, which reflects the "squared effect". Moreover the *saddle point* method provides errors that are almost 10 times lower than the *primal-dual* method. This impressive improvement can be explained by the fact that the proposed problem is "almost compliant", in the sense that the primal and dual solutions are similar: the primal solution is associated to a vertical force on the green square of Figure 3.1(a), and the dual solution is associated to vertical loading on Γ. To illustrate this, let us consider a "less compliant" application where the variable of interest is defined as the horizontal displacement (with respect to the direction e_2, see figure 3.1(a)) of the solution on the blue line Γ, *e.g.* $s_2(\xi) = L_2 u(\xi) = u_{|\Gamma}(\xi) \cdot e_2$. The same numerical analysis is applied, and the results are given on Figure 3.3. We can draw similar conclusions compared to the original application, but the *saddle point* method provides a solution that is "only" 2 times better (instead of 10 times) than the *primal-dual* method.

We consider now the effectivity indices $\eta(\xi) = \Delta(\xi)/\|s(\xi) - \widetilde{s}(\xi)\|_Z$ associated to the primal-dual error estimate defined by (3.38), and to the saddle-point error estimate defined by (3.39). For the considered application, the coercivity constant $\alpha(\xi)$ is exactly given for free using the *min-theta* method, see Proposition 2.35 in [75]. Figure 3.4 presents statistics for $\eta(\xi)$: the PDF, the mean, the max-min ratio and the normalized standard deviation. We first observe on figure 3.4(a) that the effectivity index is always greater that 1: this illustrates the fact that the error estimates are certified. Moreover, the error estimate of the saddle point method is better than the one of the primal-dual method since the max-min ratio and the

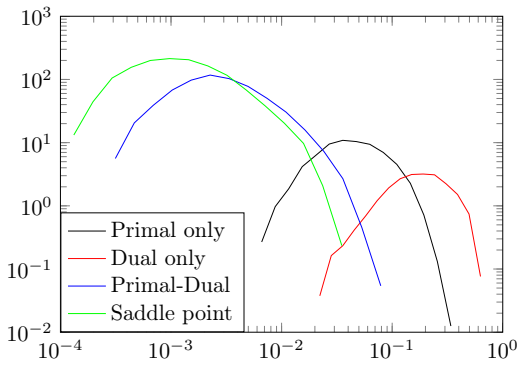


(a) PDF of the error.

	L^∞-norm	L^2-norm
Primal only	8.16×10^0	9.72×10^{-1}
Dual only	6.73×10^1	1.22×10^1
Primal-dual	2.60×10^0	2.47×10^{-1}
Saddle point	3.56×10^{-1}	4.69×10^{-2}

(b) L^∞ and L^2 norm of the error.

Figure 3.2: Application 1: Probability density function, L^∞ norm and L^2 norm of the error $\|s(\xi) - \widetilde{s}(\xi)\|_Z$ estimated on a training set of cardinality 10^4.



(a) PDF of the error.

	L^∞-norm	L^2-norm
Primal only	3.90×10^{-1}	9.03×10^{-2}
Dual only	7.09×10^{-1}	2.87×10^{-1}
Primal-dual	9.61×10^{-2}	1.34×10^{-2}
Saddle point	4.28×10^{-2}	7.07×10^{-3}

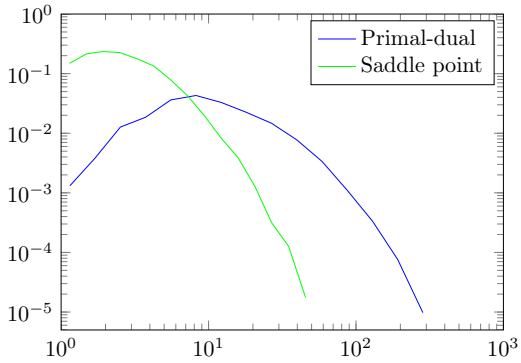
(b) L^∞ and L^2 norm of the error.

Figure 3.3: Application 1 with a different variable of interest ("less compliant case"): Probability density function, L^∞ norm and L^2 norm of the error $\|s_2(\xi) - \widetilde{s}_2(\xi)\|_Z$ estimated on a training set of cardinality 10^4.

standard deviation of the corresponding effectivity index are smaller. Finally we note that the mean value is closer to one for the saddle point method.

4.2.2 Application 2

For this application, $V_r = (u(\xi_1), \cdots, u(\xi_{50}))$ contains 50 snapshots of the primal solution (then $r = 50$), and $W_k^Q = (Q(\xi_1), \cdots, Q(\xi_{25}))$ contains 25 snapshots of the dual solution so that its dimension is $k = 25l = 50$. The test space W_r is defined according to (3.37), where $P_m(\xi)$ is an interpolation of $A(\xi)^{-1}$ using m interpolation points selected by a greedy procedure based on the residual $\|I - P_m(\xi)A(\xi)\|_F$ (see Chapter 2). The interpolation is defined by a Frobenius semi-norm projection



(a) PDF of $\eta(\xi)$ for the primal-dual method and the saddle point method.

	Primal-dual	Saddle point
$\mathbb{E}(\eta(\xi))$	26.9	4.42
$\frac{\max \eta(\xi)}{\min \eta(\xi)}$	366.7	51.8
$\frac{\mathrm{Var}(\eta(\xi))^{1/2}}{\mathbb{E}(\eta(\xi))}$	1.34	0.808

(b) Statistics of the effectivity index $\eta(\xi)$ for the primal-dual method and saddle point method.

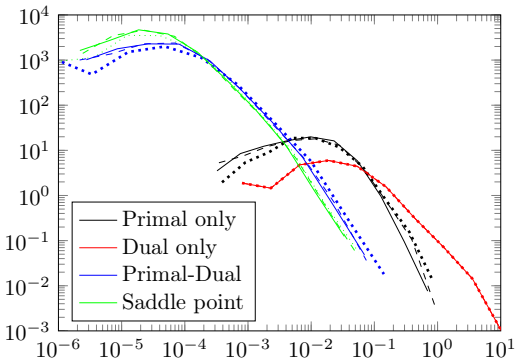
Figure 3.4: Application 1: Probability density function, mean, min-max ratio and normalized standard deviation of the effectivity index $\eta(\xi) = \Delta(\xi)/\|s(\xi) - \widetilde{s}(\xi)\|_Z$ estimated on a training set of cardinality 10^4. Here, $\Delta(\xi)$ is defined by (3.38) for the primal-dual method and by (3.39) for the saddle point method.

(with positivity constraint) using a P-SRHT matrix with 400 columns. The matrix associated to the test space is given by $W_r(\xi) = P_m^T(\xi) R_V V_r$.

Once again, we consider a training set Ξ_t of cardinality $t = 10^4$. For any $\xi \in \Xi_t$ we compute the exact quantity of interest $s(\xi)$ and the approximation $\widetilde{s}(\xi)$ by the following methods:

- *Primal only*: solve the linear system $\big(W_r^T(\xi)A(\xi)V_r\big)U_r(\xi) = \big(W_r(\xi)b\big)$ of size r and compute $\widetilde{s}(\xi) = \big(LV_r\big)U_r(\xi)$.
- *Dual only*: solve the linear system $\big((W_k^Q)^T A(\xi) R_V^{-1} A(\xi)^* W_k^Q\big)Y_k(\xi) = \big((W_k^Q)^T b\big)$ of size k and compute $\widetilde{s}(\xi) = \big(LR_V^{-1}A(\xi)^* W_k^Q\big)Y_k(\xi)$.
- *Primal-dual*: solve the linear system of the Primal only method, solve the linear system $\big((W_k^Q)^T A(\xi) R_V^{-1} A(\xi)^* W_k^Q\big)Y_k(\xi) = \big((W_k^Q)^T b\big) - \big((W_k^Q)^T A(\xi)V_r\big)U_r(\xi)$ of size k and compute $\widetilde{s}(\xi) = \big(LV_r\big)U_r(\xi) + \big(LR_V^{-1}A(\xi)^* W_k^Q\big)Y_k(\xi)$.
- *Saddle point*: solve the linear system of size $p + r$:
$$\begin{pmatrix} \big(W_p^T(\xi)A(\xi)R_V^{-1}A(\xi)^* W_p(\xi)\big) & \big(W_p^T(\xi)A(\xi)V_r\big) \\ \big(W_p^T(\xi)A(\xi)V_r\big)^T & 0 \end{pmatrix} \begin{pmatrix} Y_{r,p}(\xi) \\ U_{r,p}(\xi) \end{pmatrix} = \begin{pmatrix} \big(W_p(\xi)^T b\big) \\ 0 \end{pmatrix}$$
with $W_p(\xi) = \big(W_r(\xi), W_k^Q\big)$, and compute
$$\widetilde{s}(\xi) = \big(LV_r\big)U_{r,p}(\xi) + \big(LR_V^{-1}A(\xi)^* W_p(\xi)\big)Q_{r,p}(\xi).$$

The numerical results are given in Figure 3.5. Once again, the saddle point method leads to the lowest error on the variable of interest. Also, we see that a good preconditioner (for example with $m = 30$) improves the accuracy for the saddle point method, the primal only method and the primal-dual method. However, this improvement is not really significant for the considered application: the errors are barely divided by 2 compared to the Galerkin projection ($m = 0$). In fact, the preconditioner improves the quality of the test space, and the choice $W_r = V_r$ (that yields the standard Galerkin projection) is sufficiently accurate for this example and for the chosen norm on V.



(a) PDF of the error. Three different preconditioners $P_m(\xi)$ are used: $m = 0$ (dotted lines), $m = 10$ (dashed lines) and $m = 30$ (continuous lines).

Primal only	L^∞-norm	L^2-norm
$m = 0$	1.284×10^0	1.245×10^{-1}
$m = 5$	1.203×10^0	9.637×10^{-2}
$m = 10$	1.458×10^0	1.064×10^{-1}
$m = 20$	1.068×10^0	8.386×10^{-2}
$m = 30$	1.066×10^0	7.955×10^{-1}

Primal-dual	L^∞-norm	L^2-norm
$m = 0$	2.751×10^{-1}	1.085×10^{-2}
$m = 5$	1.308×10^{-1}	5.708×10^{-3}
$m = 10$	1.333×10^{-1}	5.807×10^{-3}
$m = 20$	1.232×10^{-1}	5.465×10^{-3}
$m = 30$	1.224×10^{-1}	5.408×10^{-3}

Saddle point	L^∞-norm	L^2-norm
$m = 0$	1.023×10^{-1}	4.347×10^{-3}
$m = 5$	9.715×10^{-2}	3.389×10^{-3}
$m = 10$	9.573×10^{-2}	3.867×10^{-3}
$m = 20$	6.022×10^{-2}	2.996×10^{-3}
$m = 30$	5.705×10^{-2}	2.896×10^{-3}

(b) L^∞ and L^2 norm of the error.

Figure 3.5: Application 2: Probability density function, L^∞ norm and L^2 norm of the error $\|s(\xi) - \widetilde{s}(\xi)\|_Z$ estimated over a training set of cardinality 10^4.

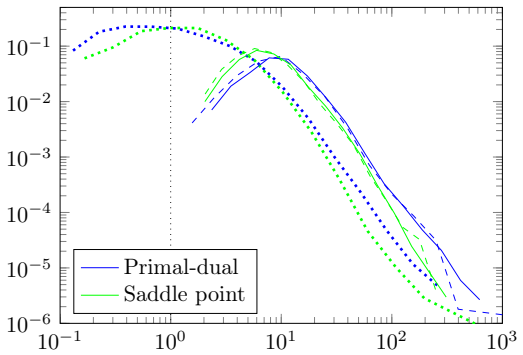
We discuss now the quality of the error estimate $\Delta(\xi)$ for the variable of interest. Since in this application the constant $\alpha(\xi)$ can not be easily computed, we consider surrogates for (3.38) and (3.40) using a preconditoner $P_m(\xi)$. We consider

$$\Delta(\xi) = \|P_m(\xi)(A(\xi)\widetilde{u}(\xi) - b(\xi))\|_{V_0} \|L(\xi)^* - A(\xi)^* \widetilde{Q}(\xi)\|_{Z' \to V_0'}, \tag{3.49}$$

for the primal-dual method and

$$\Delta(\xi) = \|P_m(\xi)(A(\xi)t_{r,p}(\xi) - b(\xi))\|_{V_0} \sup_{0 \neq z' \in Z'} \min_{y \in W_p} \frac{\|L(\xi)^* z' - A(\xi)^* y\|_{V_0'}}{\|z'\|_{Z'}} \tag{3.50}$$

for the saddle point method. Figure 3.6 shows statistics of the effectivity index $\eta(\xi) = \Delta(\xi)/\|s(\xi) - \widetilde{s}(\xi)\|_Z$ for different numbers m of interpolation points for the preconditioner. We see that the max-min ratio and the normalized standard deviation are decreasing with m: this indicates an improvement of the error estimate. Furthermore, the mean value of $\eta(\xi)$ seems to converge (with respect to m) around 19.5 for the primal-dual method, which is higher compared to the value 13.8 for the saddle point method. In fact, with a good preconditioner, $\|P_m(\xi)(A(\xi)\widetilde{u}_r(\xi) - b(\xi))\|_{V_0}$ (or $\|P_m(\xi)(A(\xi)t_{r,p}(\xi) - b(\xi))\|_{V_0}$) is expected to be a good approximation of the primal error $\|u(\xi) - \widetilde{u}_r(\xi)\|_{V_0}$ (or $\|u(\xi) - t_{r,p}(\xi)\|_{V_0}$), but this does not ensure that the effectivity index $\eta(\xi)$ will converge to 1.



(a) PDF of $\eta(\xi)$ for the primal-dual methods and the saddle point methods. Three different preconditioners $P_m(\xi)$ are presented: $m = 0$ (dotted lines), $m = 10$ (dashed lines) and $m = 30$ (continuous lines)

		$\mathbb{E}(\eta(\xi))$	$\frac{\max \eta(\xi)}{\min \eta(\xi)}$	$\frac{\mathrm{Var}(\eta(\xi))^{1/2}}{\mathbb{E}(\eta(\xi))}$
Primal-dual	$m = 0$	5.545	3.52×10^3	2.246
	$m = 5$	16.03	8.68×10^2	1.920
	$m = 10$	18.69	1.01×10^3	1.925
	$m = 20$	19.20	5.77×10^2	1.504
	$m = 30$	19.59	3.95×10^2	1.615
Saddle point	$m = 0$	4.726	6.93×10^3	3.597
	$m = 5$	12.61	1.80×10^2	1.429
	$m = 10$	13.27	1.72×10^2	1.160
	$m = 20$	13.97	1.89×10^2	1.090
	$m = 30$	13.84	2.17×10^2	1.113

(b) Statistics of the effectivity index $\eta(\xi)$ for the primal-dual method and the saddle point method.

Figure 3.6: Application 2: PDF, mean, max-min ratio and normalized standard deviation of the effectivity index $\eta(\xi) = \Delta(\xi)/\|s(\xi) - \widetilde{s}(\xi)\|_Z$. Here, $\Delta(\xi)$ is defined by (3.49) for the primal-dual method and by (3.50) for the saddle point method.

4.2.3 Partial conclusions and remarks

In both numerical examples, the saddle point method provides the most accurate estimation for the variable of interest. Let us note that the saddle point problem requires the solution of a dense linear system of size $(r+k)$ for the SPD case, and of size $(2r+k)$ for the general case. When using Gaussian elimination for the solution of those systems, the complexity is either in $C(r+k)^3$ or $C(2r+k)^3$ (with $C = 2/3$), which is larger than the complexity of the primal-dual method $C(r^3 + k^3)$. However, in the case where the primal and dual approximation spaces have the same

dimension $r = k$, the saddle point method is only 4 times (SPD case) or 13.5 times (general case) more expensive.

Furthermore, we showed that the preconditioner slightly improves the quality of the estimation $\widetilde{s}(\xi)$, and of the error estimate $\Delta(\xi)$. Since the construction of the preconditioner yield a significant increase in computational and memory costs (see [127]), preconditioning is clearly not needed for these applications.

4.3 Greedy construction of the reduced spaces

We now consider the greedy construction of the reduced spaces, see Algorithms 5 and 6. For the two considered applications, we show the convergence of the error estimate with respect to the complexity of the offline and of the online phase. For the sake of simplicity, we measure the complexity of the offline phase with the number of operator factorizations (this corresponds to the number of iteration I of Algorithms 5 and 6). Of course exact estimation of the offline complexity should take into account many other steps (for example, the computation of $\Delta(\xi)$, of the preconditioner etc), but the operator factorization is, for large scale applications, the main source of computation cost. For the online complexity, we only consider the computational cost for the solution of one reduced system, see Section 4.2.3. Here we do not take into account the complexity for assembling the reduced systems although it may be a significant part of the complexity for "not so reduced" systems of equations.

4.3.1 Application 1

Figure 3.7 shows the convergence of $\max_\xi \Delta(\xi)$ with respect to the offline and online complexities. On Figure 3.7(a), we see that the saddle point method (dashed lines) always provides lower values for the error estimate compared to the primal-dual method (continuous lines). But as already mentioned, the saddle point method requires the solution of larger reduced systems during the *online* phase. With this amount of computational cost, the primal-dual method can sometimes provide lower error estimate (see the blue and red curves of Figure 3.7(b)) for the same *online* complexity.

The simultaneous construction of V_r and W_k^Q with full dual enrichment (3.43) (green curves) yields a very fast convergence of the error estimate during the *offline* phase, see Figure 3.7(a)). But the rapid increase of $\dim(W_k^Q)$ leads to high *online* complexity, so that this strategy becomes non competitive during the *online* phase,

see Figure 3.7(b).

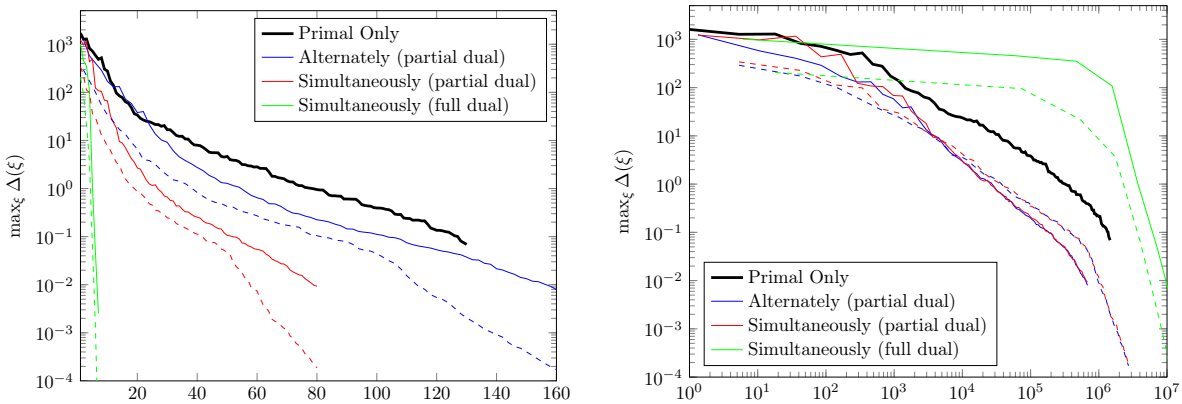
We compare now the alternate and the simultaneous construction of V_r and W_k^Q with partial dual enrichment (3.44) (red and blue curves on Figure 3.7). The initial idea of the alternate construction is to build reduced spaces of better quality. Indeed, since the evaluation points of the primal solution are different from the one of the dual solution, the reduced spaces are expected to contain more relevant information for the approximation of the variable of interest. In practice, we observe on Figure 3.7(a) that the alternate construction is (two times) more expensive during the *offline* phase, but the resulting error estimate behaves very similarly to the simultaneous strategy, see Figure 3.7(b). We conclude that the alternate strategy is not relevant for this application.

Furthermore, let us note that after the 50-th iteration of the greedy algorithm, the rate of convergence of the dashed red curve of Figure 3.7(a) (*i.e.* the simultaneous construction with partial dual enrichment using the saddle point method) rapidly increase. A possible explanation for that is that the dimension of the dual approximation space is large enough to reproduce correctly the dual variable, that requires a dimension higher than $l = 44$. The same observation can be done for the alternate strategy, *i.e.* the dashed blue curve, but after the iteration 100 (that corresponds to $\dim(W_k^Q) \geq 50$). Also, we note that the primal-dual method does not present this behavior.

4.3.2 Application 2

For the application 2, we first test Algorithms 5 and 6 with the use of preconditioner. We recall that the preconditioner $P_m(\xi)$ is defined by the Frobenius semi-norm projection (with positivity constraint) using a P-SRHT matrix with 400 columns (see [127] and Chapter 2), and that the interpolation points for the preconditioner are the ones where solutions (primal and dual) have been computed, see Algorithms 5 and 6. The preconditioner is used for the definition of the test space $W_r(\xi)$, see equation (3.37), and for the error estimate $\Delta(\xi)$, see equation (3.49) for the primal-dual method and (3.50) for the saddle point method. The numerical results are given on Figure 3.8. We can draw the same conclusions as for application 1.

- During the offline phase, the saddle point method provides lower errors (Figure 3.8(a)). But the corresponding reduced systems are larger, and we see that the primal-dual method provides lower errors for the same online complexity,



(a) Maximum value of the error estimate $\Delta(\xi)$ with respect to the number of operator factorizations.

(b) Maximum value of the error estimate $\Delta(\xi)$ with respect to the complexity for solving one reduced system.

Figure 3.7: Application 1: error estimate $\max_\xi \Delta(\xi)$ with respect to the offline complexity (Figure 3.7(a)) and the online complexity (Figure 3.7(b)). The continuous lines correspond to the primal-dual method, and the dashed lines correspond to the saddle point method. The primal only curves serve as reference.

see Figure 3.8(b). For this test case, the benefits (in term of accuracy) of the saddle point method does not compensate the amount of additional online computational costs.

- The full dual enrichment yields a fast convergence during the offline phase, but the rapid increase of W_k^Q is disadvantageous regarding the online complexity. However, since the dimension of the variable of interest is "only" $l = 2$, the full dual enrichment is still an acceptable strategy (compared to the previous test case).

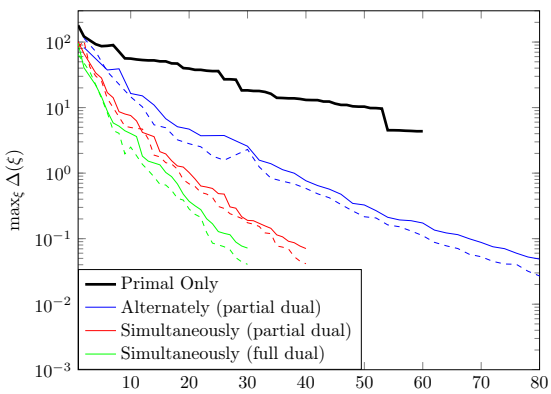
- Here, the alternate strategy (blue curves) seems to yield slightly better reduced spaces compared to the simultaneous strategy, see Figure 3.8(b). But this leads to higher offline costs, see Figure 3.8(a).

We also run numerical tests without using the preconditioner. In that case, we replace $P_m(\xi)$ by R_V^{-1}. Figure 3.9 shows that the numerical results are very similar to those of Figure 3.8. To illustrate the benefits of using the preconditioner, let us consider the effectivity index $\eta(\xi) = \Delta(\xi)/\|s(\xi) - \widetilde{s}(\xi)\|_Z$ associated to the error estimate for the variable of interest. Figure 3.10 shows the confidence interval $I(p)$ of probability p for $\eta(\xi)$ defined as the smallest interval which satisfies

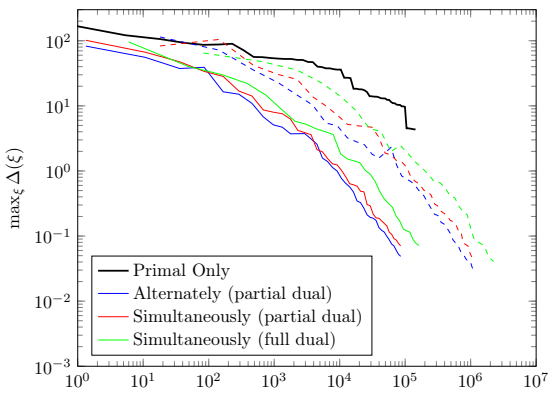
$$\mathbb{P}(\xi \in \Xi_t : \eta(\xi) \in I(p)) \geq p,$$

where $\mathbb{P}(A) = \#A/\#\Xi_t$ for $A \subset \Xi_r$ (Ξ_t being the training set). When using the preconditioner, we see on Figure 3.10 that the effectivity index is improved during the greedy iteration process in the sense that the confidence intervals are getting smaller and smaller. Also, we note that after the iteration 15, the effectivity index is always above 1: this indicates that the error estimate tends to be certified. Furthermore, after iteration 20 we do not observe any further improvement, so that is seems not useful to continue enriching the preconditioner.

Let us finally note that the use of the preconditioner yields significant computational cost. Indeed, we have to store operator factorizations, and the computation of the Frobenius semi-norm projections requires additional problems to solve. For the present application, even if the effectivity index of the error estimate is improved, the benefits of using the preconditioner remains questionable.

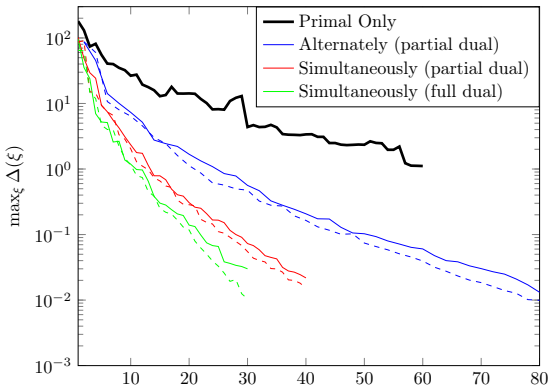


(a) Maximum value of the error estimate $\Delta(\xi)$ with respect to the number of operator factorization.

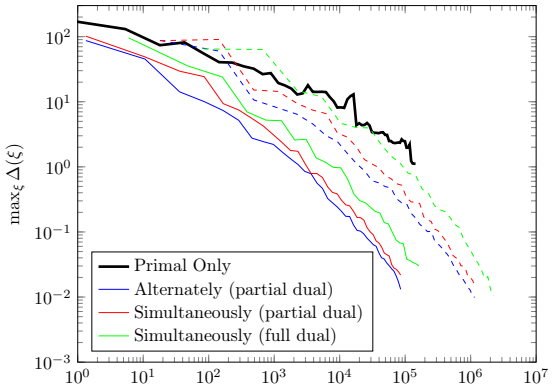


(b) Maximum value of the error estimate $\Delta(\xi)$ with respect to the complexity for solving one reduced system.

Figure 3.8: Application 2 when using the preconditioner: error estimate $\max_\xi \Delta(\xi)$ with respect to the offline complexity (Figure 3.8(a)) and the online complexity (Figure 3.8(b)). The continuous line corresponds to the primal-dual method, and the dashed line corresponds to the saddle point method. The primal only curve serves as reference.

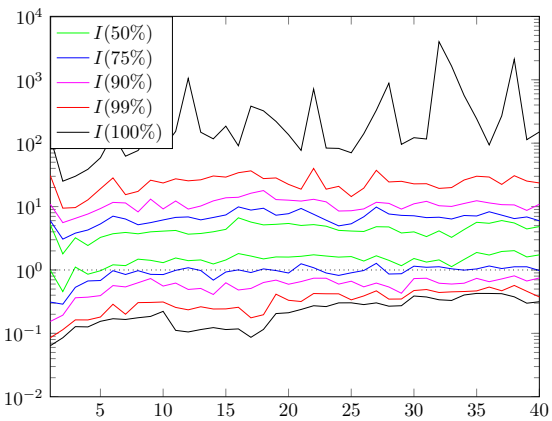


(a) Maximum value of the error estimate $\Delta(\xi)$ with respect to the number of operator factorization.

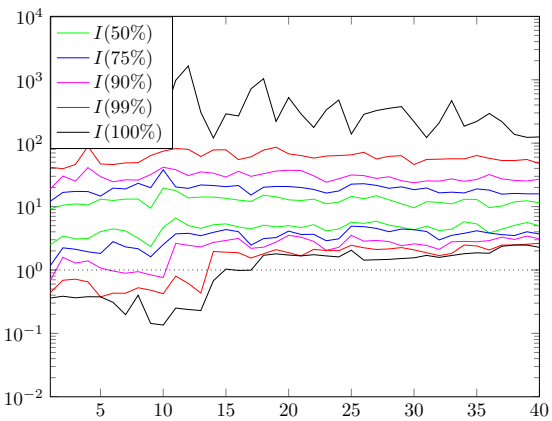


(b) Maximum value of the error estimate $\Delta(\xi)$ with respect to the complexity for solving one reduced system.

Figure 3.9: Application 2 when not using the preconditioner: error estimate $\max_\xi \Delta(\xi)$ with respect to the offline complexity (Figure 3.9(a)) and the online complexity (Figure 3.9(b)). The continuous line corresponds to the primal-dual method, and the dashed line corresponds to the saddle point method. The primal only curve serves as reference.



(a) Without preconditioner.



(b) With preconditioner.

Figure 3.10: Application 2: evolution with respect to the greedy iteration process of the confidence interval $I(p)$ for the effectivity index $\eta(\xi) = \Delta(\xi)/\|s(\xi) - \widetilde{s}(\xi)\|_Z$ for saddle point method.

5 Conclusion

We have proposed and analyzed projection-based methods for the estimation of vector-valued variables of interest in the context of parameter-dependent equations.

This includes a generalization of the classical primal-dual method to vector-valued variables of interest, and also a new method based on a saddle point problem. Numerical results showed that the saddle point method always improves the quality of the approximation compared to the primal-dual method. In the spirit of the Reduced Basis method, we have proposed greedy algorithms for the goal-oriented construction of the reduced spaces. Finally, the use of preconditioners defined by interpolation of the operator inverse yields better reduced test spaces, and better error estimates. However, and for the considered applications, we do not observe sufficient improvement regarding the additional costs for constructing such preconditioners.

Chapter 4

Ideal minimal residual formulation for tensor approximation

This Chapter is based on the article [17].

We propose a method for the approximation of the solution of high-dimensional weakly coercive problems formulated in tensor spaces using low-rank approximation formats. The method can be seen as a perturbation of a minimal residual method with a measure of the residual corresponding to the error in a specified solution norm. The residual norm can be designed such that the resulting low-rank approximations are optimal with respect to particular norms of interest, thus allowing to take into account a particular objective in the definition of reduced order approximations of high-dimensional problems. We introduce and analyze an iterative algorithm that is able to provide an approximation of the optimal approximation of the solution in a given low-rank subset, without any a priori information on this solution. We also introduce a weak greedy algorithm which uses this perturbed minimal residual method for the computation of successive greedy corrections in small tensor subsets. We prove its convergence under some conditions on the parameters of the algorithm. The proposed numerical method is applied to the solution of a stochastic partial differential equation which is discretized using standard Galerkin methods in tensor product spaces.

Contents

1 Introduction . . . 109

2 Functional framework for weakly coercive problems . . . 112

2.1 Notations . . . 112

2.2 Weakly coercive problems . . . 112

3 Approximation in low-rank tensor subsets . . . 113

3.1 Hilbert tensor spaces . . . 113

3.2 Classical low-rank tensor subsets . . . 114

3.3 Best approximation in tensor subsets . . . 115

4 Minimal residual based approximation . . . 116

4.1 Best approximation with respect to residual norms . . . 116

4.2 Ideal choice of the residual norm . . . 117

4.3 Gradient-type algorithm . . . 118

5 Perturbation of the ideal approximation . . . 120

5.1 Approximation of the ideal approach . . . 120

5.2 Quasi-optimal approximations in $\mathcal{M}_r(X)$ . . . 121

5.3 Perturbed gradient-type algorithm . . . 121

5.4 Error indicator . . . 123

6 Computational aspects . . . 124

6.1 Best approximation in tensor subsets . . . 124

6.2 Construction of an approximation of $\Lambda^\delta(r)$ . . . 125

6.3 Summary of the algorithm . . . 129

7 Greedy algorithm . . . 129

7.1 A weak greedy algorithm . . . 130

7.2 Convergence analysis . . . 131

8 Numerical example . . . 135

8.1 Stochastic reaction-advection-diffusion problem . . . 135

8.2 Comparison of minimal residual methods . . . 136

8.3 Properties of the algorithms . . . 141

8.4 Higher dimensional case . . . 144

9 Conclusion . . . 148

1 Introduction

Low-rank tensor approximation methods are receiving growing attention in computational science for the numerical solution of high-dimensional problems formulated in tensor spaces (see the recent surveys [35,72,85,87] and monograph [76]). Typical problems include the solution of high-dimensional partial differential equations arising in stochastic calculus, or the solution of stochastic or parametric partial differential equations using functional approaches, where functions of multiple (random) parameters have to be approximated. These problems take the general form

$$A(u) = b, \quad u \in X = X_1 \otimes \cdots \otimes X_d, \tag{4.1}$$

where A is an operator defined on the tensor space X. Low-rank tensor methods then consist in searching an approximation of the solution u in a subset $\mathcal{M}_r(X)$ of tensors with bounded ranks. The elements of $\mathcal{M}_r(X)$ can take the form

$$\sum_{i_1} \cdots \sum_{i_d} \alpha_{i_1 \cdots i_d} w^1_{i_1} \otimes \cdots \otimes w^d_{i_d}, \; w^\mu_{i_\mu} \in X_\mu, \tag{4.2}$$

where the set of coefficients $(\alpha_{i_1 \cdots i_d})$ possesses some specific structure. Classical low-rank tensor subsets include canonical tensors, Tucker tensors, Tensor Train tensors [80,106], Hierarchical Tucker tensors [78] or more general tree-based Hierarchical Tucker tensors [57]. In practice, many tensors arising in applications are observed to be efficiently approximable by elements of the mentioned subsets. Low-rank approximation methods are closely related to *a priori* model reduction methods in that they provide approximate representations of the solution on low-dimensional reduced bases $\{w^1_{i_1} \otimes \cdots \otimes w^d_{i_d}\}$ that are not selected *a priori*.

The best approximation of $u \in X$ in a given low-rank tensor subset $\mathcal{M}_r(X)$ with respect to a particular norm $\|\cdot\|_X$ in X is the solution of

$$\min_{v \in \mathcal{M}_r(X)} \|u - v\|_X. \tag{4.3}$$

Low-rank tensor subsets are neither linear subspaces nor convex sets. However, they usually satisfy topological properties that make the above best approximation problem meaningful and allows the application of standard optimization algorithms [54, 113, 122]. Of course, in the context of the solution of high-dimensional problems, the solution u of problem (4.1) is not available, and the best approximation problem (4.3) cannot be solved directly. Tensor approximation methods then typically rely on the definition of approximations based on the residual of equation

(4.1), which is a computable quantity. Different strategies have been proposed for the construction of low-rank approximations of the solution of equations in tensor format. The first family of methods consists in using classical iterative algorithms for linear or nonlinear systems of equations with low-rank tensor algebra (using low-rank tensor compression) for standard algebraic operations [8,86,89,98]. The second family of methods consists in directly computing an approximation of u in $\mathcal{M}_r(X)$ by minimizing some residual norm [16, 51, 100]:

$$\min_{v \in \mathcal{M}_r(X)} \|Av - b\|_\star. \tag{4.4}$$

In the context of approximation, where one is interested in obtaining an approximation with a given precision rather than obtaining the best low-rank approximation, constructive greedy algorithms have been proposed that consist in computing successive corrections in a small low-rank tensor subset, typically the set of rank-one tensors [3, 90, 100]. These greedy algorithms have been analyzed in several papers [2, 26, 27, 58–60] and a series of improved algorithms have been introduced in order to increase the quality of suboptimal greedy constructions [59,67,91,102,103].

Although minimal residual based approaches are well founded, they generally provide low-rank approximations that can be very far from optimal approximations with respect to the natural norm $\|\cdot\|_X$, at least when using usual measures of the residual. If we are interested in obtaining an optimal approximation with respect to the norm $\|\cdot\|_X$, e.g. taking into account some particular quantity of interest, an ideal approach would be to define the residual norm such that

$$\|Av - b\|_\star = \|u - v\|_X,$$

where $\|\cdot\|_X$ is the desired solution norm, that corresponds to solve an ideally conditioned problem. Minimizing the residual norm would therefore be equivalent to solving the best approximation problem (4.3). However, the computation of such a residual norm is in general equivalent to the solution of the initial problem (4.1).

In this chapter, we propose a method for the approximation of the ideal approach. This method applies to a general class of weakly coercive problems. It relies on the use of approximations $r_\delta(v)$ of the residual $r(v) = Av - b$ such that $\|r_\delta(v)\|_\star$ approximates the ideal residual norm $\|r(v)\|_\star = \|u - v\|_X$. The resulting method allows for the construction of low-rank tensor approximations which are quasi-optimal with respect to a norm $\|\cdot\|_X$ that can be designed according to some quantity of interest. We first introduce and analyze an algorithm for minimizing

the approximate residual norm $\|r_\delta(v)\|_\star$ in a given subset $\mathcal{M}_r(X)$. This algorithm can be seen as an extension of the algorithms introduced in [39, 41] to the context of nonlinear approximation in subsets $\mathcal{M}_r(X)$. It consists in a perturbation of a gradient algorithm for minimizing in $\mathcal{M}_r(X)$ the ideal residual norm $\|r(v)\|_\star$, using approximations $r_\delta(v)$ of the residual $r(v)$. An ideal algorithm would consist in computing an approximation $r_\delta(v)$ such that

$$(1-\delta)\|u-v\|_X \leq \|r_\delta(v)\|_\star \leq (1+\delta)\|u-v\|_X, \tag{4.5}$$

for some precision δ, that requires the use of guaranteed error estimators. In the present chapter, (4.5) is not exactly satisfied since we only use heuristic error estimates. However, these estimates seem to provide an acceptable measure of the error for the considered applications. The resulting algorithm can be interpreted as a preconditioned gradient algorithm with an implicit preconditioner that approximates the ideal preconditioner. Also, we propose a weak greedy algorithm for the adaptive construction of an approximation of the solution of problem (4.1), using the perturbed ideal minimal residual approach for the computation of greedy corrections. A convergence proof is provided under some conditions on the parameters of the algorithm.

The outline of the chapter is as follows. In section 2, we introduce a functional framework for weakly coercive problems. In section 3, we briefly recall some definitions and basic properties of tensor spaces and low-rank tensor subsets. In section 4, we present a natural minimal residual based method for the approximation in a nonlinear subset $\mathcal{M}_r(X)$, and we analyze a simple gradient algorithm in $\mathcal{M}_r(X)$. We discuss the conditioning issues that restrict the applicability of such algorithms when usual residual norms are used, and the interest of using an ideal measure of the residual. In section 5, we introduce the perturbed ideal minimal residual approach. A gradient-type algorithm is introduced and analyzed and we prove the convergence of this algorithm towards a neighborhood of the best approximation in $\mathcal{M}_r(X)$. Practical computational aspects are detailed in section 6. In section 7, we analyze a weak greedy algorithm using the perturbed ideal minimal residual method for the computation of greedy corrections. In section 8, a detailed numerical example will illustrate the proposed method. The example is a stochastic reaction-advection-diffusion problem which is discretized using Galerkin stochastic methods. In particular, this example will illustrate the possibility to introduce norms that are adapted to some quantities of interest and the ability of the method to provide (quasi-)best low-rank approximations in that context.

2 Functional framework for weakly coercive problems

2.1 Notations

For a given Hilbert space H, we denote by $\langle \cdot, \cdot \rangle_H$ the inner product in H and by $\| \cdot \|_H$ the associated norm. We denote by H' the topological dual of H and by $\langle \cdot, \cdot \rangle_{H',H}$ the duality pairing between H and H'. For $v \in H$ and $\varphi \in H'$, we denote $\varphi(v) = \langle \varphi, v \rangle_{H',H}$. We denote by $R_H : H \to H'$ the Riesz isomorphism defined by

$$\langle v, w \rangle_H = \langle v, R_H w \rangle_{H,H'} = \langle R_H v, w \rangle_{H',H} = \langle R_H v, R_H w \rangle_{H'} \quad \forall v, w \in H.$$

2.2 Weakly coercive problems

We denote by X (resp. Y) a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_X$ (resp. $\langle \cdot, \cdot \rangle_Y$) and associated norm $\| \cdot \|_X$ (resp. $\| \cdot \|_Y$). Let $a : X \times Y \to \mathbb{R}$ be a bilinear form and let $b \in Y'$ be a continuous linear form on Y. We consider the variational problem: find $u \in X$ such that

$$a(u, v) = b(v) \quad \forall v \in Y. \tag{4.6}$$

We assume that a is continuous and weakly coercive, that means that there exist constants α and β such that

$$\sup_{v \in X} \sup_{w \in Y} \frac{a(v, w)}{\|v\|_X \|w\|_Y} = \beta < +\infty, \tag{4.7}$$

$$\inf_{v \in X} \sup_{w \in Y} \frac{a(v, w)}{\|v\|_X \|w\|_Y} = \alpha > 0, \tag{4.8}$$

and

$$\sup_{v \in X} \frac{a(v, w)}{\|v\|_X} > 0 \quad \forall w \neq 0 \text{ in } Y. \tag{4.9}$$

We introduce the linear continuous operator $A : X \to Y'$ such that for all $(v, w) \in X \times Y$,

$$a(v, w) = \langle Av, w \rangle_{Y',Y}.$$

We denote by $A^* : Y \to X'$ the adjoint of A, defined by

$$\langle Av, w \rangle_{Y',Y} = \langle v, A^* w \rangle_{X,X'} \quad \forall (v, w) \in X \times Y.$$

Problem (4.6) is therefore equivalent to find $u \in X$ such that

$$Au = b. \tag{4.10}$$

Properties (4.7),(4.8) and (4.9) imply that A is a norm-isomorphism from X to Y' such that for all $v \in X$,

$$\alpha \|v\|_X \leq \|Av\|_{Y'} \leq \beta \|v\|_X \tag{4.11}$$

ensuring the well-posedness of problem (4.10) [52]. The norms of A and its inverse A^{-1} are such that $\|A\|_{X \to Y'} = \beta$ and $\|A^{-1}\|_{Y' \to X} = \alpha^{-1}$. Then, the condition number of the operator A is

$$\kappa(A) = \|A\|_{X \to Y'} \|A^{-1}\|_{Y' \to X} = \frac{\beta}{\alpha} \geq 1.$$

3 Approximation in low-rank tensor subsets

3.1 Hilbert tensor spaces

We here briefly recall basic definitions on Hilbert tensor spaces (see [76]). We consider Hilbert spaces X_μ, $1 \leq \mu \leq d$, equipped with norms $\|\cdot\|_{X_\mu}$ and associated inner products $\langle \cdot, \cdot \rangle_\mu$[1]. We denote by $\otimes_{\mu=1}^d v^\mu = v^1 \otimes \cdots \otimes v^d$, $v^\mu \in X_\mu$, an elementary tensor. We then define the algebraic tensor product space as the linear span of elementary tensors:

$$_a\bigotimes_{\mu=1}^d X_\mu = \text{span}\{\otimes_{\mu=1}^d v^\mu \ : \ v^\mu \in X_\mu, 1 \leq \mu \leq d\}.$$

A Hilbert tensor space X equipped with the norm $\|\cdot\|_X$ is then obtained by the completion with respect to $\|\cdot\|_X$ of the algebraic tensor space, i.e.

$$X = \overline{{}_a\bigotimes_{\mu=1}^d X_\mu}^{\|\cdot\|_X} = {}_{\|\cdot\|_X}\bigotimes_{\mu=1}^d X_\mu.$$

Note that for finite dimensional tensor spaces, the resulting space X is independent of the choice of norm and coincides with the normed algebraic tensor space.

[1] e.g. $X_\mu = \mathbb{R}^{n_\mu}$ equipped with the Euclidian norm, or $X_\mu = H_0^k(\Omega_\mu)$, $k \geq 0$, a Sobolev space of functions defined on a domain Ω_μ.

A natural inner product on X is induced by inner products $\langle \cdot, \cdot \rangle_\mu$ in X_μ, $1 \leq \mu \leq d$. It is defined for $v = \otimes_{\mu=1}^d v^\mu$ and $w = \otimes_{\mu=1}^d w^\mu$ by

$$\langle v, w \rangle_X = \prod_{\mu=1}^d \langle v^\mu, w^\mu \rangle_\mu$$

and extended by linearity on the whole algebraic tensor space. This inner product is called the *induced (or canonical) inner product* and the associated norm the *induced (or canonical) norm.*

3.2 Classical low-rank tensor subsets

Low-rank tensor subsets $\mathcal{M}_r(X)$ of a tensor space $X = {}_{\|\cdot\|}\bigotimes_{\mu=1}^d X_\mu$ are subsets of the algebraic tensor space ${}_a\bigotimes_{\mu=1}^d X_\mu$, which means that elements $v \in \mathcal{M}_r(X)$ can be written under the form

$$v = \sum_{i_1 \in I_1} \cdots \sum_{i_d \in I_d} \alpha_{i_1, \cdots, i_d} \bigotimes_{\mu=1}^d v_{i_\mu}^\mu, \tag{4.12}$$

where $\alpha = (\alpha_i)_{i \in I} \in \mathbb{R}^I$, with $I := I_1 \times \cdots \times I_d$, is a set of real coefficients that possibly satisfy some constraints, and $(v_{i_\mu}^\mu)_{i_\mu \in I_\mu} \in (X_\mu)^{I_\mu}$, for $1 \leq \mu \leq d$, is a set of vectors that also possibly satisfy some constraints (e.g. orthogonality).

Basic low-rank tensor subsets are the set of tensors with canonical rank bounded by r:

$$\mathcal{C}_r(X) = \left\{ v = \sum_{i=1}^r \otimes_{\mu=1}^d v_i^\mu \ : \ v_i^\mu \in X_\mu \right\},$$

and the set of Tucker tensors with multilinear rank bounded by $r = (r_1, \cdots, r_d)$:

$$\mathcal{T}_r(X) = \left\{ v = \sum_{i_1=1}^{r_1} \cdots \sum_{i_d=1}^{r_d} \alpha_{i_1, \cdots, i_d} \otimes_{\mu=1}^d v_{i_\mu}^\mu \ : \ v_{i_\mu}^\mu \in X_\mu, \ \alpha_{i_1, \cdots, i_d} \in \mathbb{R} \right\}$$

Other low-rank tensor subsets have been recently introduced, such as Tensor Train tensors [80,106] or more general tree-based Hierarchical Tucker tensors [57,78], these tensor subsets corresponding to a form (4.12) with a particular structure of tensor α. Note that for the case $d = 2$, all the above tensor subsets coincide.

Remark 3.1. *From a numerical point of view, the approximate solution of the variational problem (4.6) requires an additional discretization which consists in introducing an approximation space $\widetilde{X} = \otimes_{\mu=1}^d \widetilde{X}_\mu$, where the $\widetilde{X}_\mu \subset X_\mu$ are finite dimensional approximation spaces (e.g. finite element spaces). Then, approxima-*

tions are searched in low-rank tensor subsets $\mathcal{M}_r(\widetilde{X})$ of X (e.g. $\mathcal{C}_r(\widetilde{X})$ or $\mathcal{T}_r(\widetilde{X})$), thus introducing two levels of discretizations. In the following, we adopt a general point of view where X can either denote an infinite dimensional space, an approximation space obtained after the discretization of the variational problem, or even finite dimensional Euclidian spaces for problems written in an algebraic form.

3.3 Best approximation in tensor subsets

Low-rank tensor approximation methods consist in computing an approximation of a tensor $u \in X$ in a suitable low-rank subset $\mathcal{M}_r(X)$ of X. The best approximation of u in $\mathcal{M}_r(X)$ is defined by

$$\min_{v \in \mathcal{M}_r(X)} \|u - v\|_X. \tag{4.13}$$

The previously mentioned classical tensor subsets are neither linear subspaces nor convex sets. However, they usually satisfy properties that give sense to the above best approximation problem. We consider the case that $\mathcal{M}_r(X)$ satisfies the following properties:

$$\mathcal{M}_r(X) \text{ is weakly closed (or simply closed in finite dimension)}, \tag{4.14}$$

$$\mathcal{M}_r(X) \subset \gamma \mathcal{M}_r(X) \text{ for all } \gamma \in \mathbb{R}. \tag{4.15}$$

Property (4.15) is satisfied by all the classical tensor subsets mentioned above (canonical tensors, Tucker and tree-based Hierarchical Tucker tensors). Property (4.14) ensures the existence of solutions to the best approximation problem (4.13). This property, under some suitable conditions on the norm $\|\cdot\|_X$ (which is naturally satisfied in finite dimension), is verified by most tensor subsets used for approximation (e.g. the set of tensors with bounded canonical rank for $d = 2$, the set of elementary tensors $\mathcal{C}_1$ for $d \geq 2$ [58], the sets of Tucker or tree-based Hierarchical Tucker tensors [56]).

We then introduce the set-valued map $\Pi_{\mathcal{M}_r(X)} : X \to 2^{\mathcal{M}_r(X)}$ that associates to an element $u \in X$ the set of best approximations of u in $\mathcal{M}_r(X)$:

$$\Pi_{\mathcal{M}_r(X)}(u) = \arg\min_{v \in \mathcal{M}_r(X)} \|u - v\|_X. \tag{4.16}$$

Note that if $\mathcal{M}_r(X)$ were a closed linear subspace or a closed convex set of X, then $\Pi_{\mathcal{M}_r(X)}(u)$ would be a singleton and $\Pi_{\mathcal{M}_r(X)}$ would coincide with the classical definition of the metric projection on $\mathcal{M}_r(X)$. Property (4.15) still implies the

following property of projections: for all $v \in X$ and for all $w \in \Pi_{\mathcal{M}_r(X)}(v)$,

$$\|v - w\|_X^2 = \|v\|_X^2 - \|w\|_X^2 \quad \text{with} \quad \|w\|_X = \sigma(v; \mathcal{M}_r(X)) = \max_{z \in \mathcal{M}_r(X)} \frac{\langle v, z \rangle_X}{\|z\|_X}. \tag{4.17}$$

$\Pi_{\mathcal{M}_r(X)}(v)$ is therefore a subset of the sphere of radius $\sigma(v; \mathcal{M}_r(X))$ in X. In the following, we will use the following abuse of notation: for a subset $S \subset X$ and for $w \in X$, we define

$$\|S - w\|_X := \sup_{v \in S} \|v - w\|_X$$

With this convention, we have $\|\Pi_{\mathcal{M}_r(X)}(v)\|_X = \sigma(v; \mathcal{M}_r(X))$ and

$$\|\Pi_{\mathcal{M}_r(X)}(v) - v\|_X^2 = \|v\|_X^2 - \|\Pi_{\mathcal{M}_r(X)}(v)\|_X^2. \tag{4.18}$$

4 Minimal residual based approximation

We now consider that problem (4.10) is formulated in tensor Hilbert spaces $X = {}_{\|\cdot\|_X}\bigotimes_{\mu=1}^d X_\mu$ and $Y = {}_{\|\cdot\|_Y}\bigotimes_{\mu=1}^d Y_\mu$. The aim is here to find an approximation of the solution u of problem (4.10) in a given tensor subset $\mathcal{M}_r(X) \subset X$.

4.1 Best approximation with respect to residual norms

Since the solution u of problem (4.10) is not available, the best approximation problem (4.13) cannot be solved directly. However, tensor approximations can be defined using the residual of equation (4.10), which is a computable information. An approximation of u in $\mathcal{M}_r(X)$ is then defined by the minimization of a residual norm:

$$\min_{v \in \mathcal{M}_r(X)} \|Av - b\|_{Y'} = \min_{v \in \mathcal{M}_r(X)} \|A(v - u)\|_{Y'}. \tag{4.19}$$

Assuming that we can define a tangent space $T_v(\mathcal{M}_r(X))$ to $\mathcal{M}_r(X)$ at $v \in \mathcal{M}_r(X)$, the stationarity condition of functional $J : v \mapsto \|A(v - u)\|_{Y'}^2$ at $v \in \mathcal{M}_r(X)$ is

$$\langle J'(v), \delta v \rangle_{X',X} = 0 \quad \forall \delta v \in T_v(\mathcal{M}_r(X)),$$

or equivalently, noting that the gradient of J at v is $J'(v) = A^* R_Y^{-1}(Av - b) \in X'$,

$$\langle Av - b, A\delta v \rangle_{Y'} = 0 \quad \forall \delta v \in T_v(\mathcal{M}_r(X)).$$

4.2 Ideal choice of the residual norm

When approximating u in $\mathcal{M}_r(X)$ using (4.19), the obtained approximation depends on the choice of the residual norm. If we want to find a best approximation of u with respect to the norm $\|\cdot\|_X$, then the residual norm should be chosen [39, 41] such that

$$\|A(v-u)\|_{Y'} = \|v-u\|_X \quad \forall v \in X,$$

or equivalently such that the following relation between inner products holds:

$$\langle v, w\rangle_X = \langle Av, Aw\rangle_{Y'} \quad \forall v, w \in X. \tag{4.20}$$

This implies

$$\langle v, w\rangle_X = \langle Av, R_Y^{-1}Aw\rangle_{Y',Y} = \langle v, R_X^{-1}A^*R_Y^{-1}Aw\rangle_X,$$

for all $v, w \in X$, and therefore, by identification,

$$I_X = R_X^{-1}A^*R_Y^{-1}A \Leftrightarrow R_Y = AR_X^{-1}A^* \Leftrightarrow R_X = A^*R_Y^{-1}A. \tag{4.21}$$

Also, since

$$\begin{aligned}\langle v, w\rangle_Y &= \langle R_Y v, w\rangle_{Y',Y} = \langle AR_X^{-1}A^*v, w\rangle_{Y',Y} \\ &= \langle R_X^{-1}A^*v, A^*w\rangle_{X,X'} = \langle A^*v, A^*w\rangle_{X'}\end{aligned}$$

for all $v, w \in Y$, we also have that (4.20) is equivalent to the following relation:

$$\langle v, w\rangle_Y = \langle A^*v, A^*w\rangle_{X'} \quad \forall v, w \in Y. \tag{4.22}$$

Note that (4.20) and (4.22) respectively impose

$$\|v\|_X = \|Av\|_{Y'} \text{ and } \|w\|_Y = \|A^*w\|_{X'}. \tag{4.23}$$

This choice implies that the weak coercivity and continuity constants are such that $\alpha = \beta = 1$, and therefore

$$\kappa(A) = 1,$$

meaning that problem (4.10) is ideally conditioned.

In practice, we will first define the inner product $\langle\cdot,\cdot\rangle_X$ and the other inner product $\langle\cdot,\cdot\rangle_Y$ will be deduced from (4.22).

Example 4.1. *Consider that $X = Y$ and let $A = B + C$ with B a symmetric coercive and continuous operator and C a skew-symmetric operator. We equip X with inner product $\langle v, w\rangle_X = \langle Bv, w\rangle_{X',X}$, which corresponds to $R_X = B$. Therefore,*

$$\|v\|_Y^2 = \|A^*v\|_{X'}^2 = \|Bv\|_{X'}^2 + \|Cv\|_{X'}^2 = \|v\|_X^2 + \|Cv\|_{X'}^2.$$

$\|v\|_Y$ corresponds to the graph norm of the skew-symmetric part C of the operator A. When $C = 0$, we simply have $\|v\|_Y^2 = \|v\|_X^2$.

Example 4.2 (Finite dimensional problem). *Consider the case of finite dimensional tensor spaces $X = Y = \mathbb{R}^{n_1\times\cdots\times n_d}$, e.g. after a discretization step for the solution of a high-dimensional partial differential equation. The duality pairings are induced by the standard canonical inner product. We can choose for $\langle\cdot,\cdot\rangle_X$ the canonical inner product on $\mathbb{R}^{n_1\times\cdots\times n_d}$, which corresponds to $R_X = I_X$, the identity on X. Then, inner product on Y is defined by relation (4.22), which implies*

$$\langle v, w\rangle_Y = \langle A^*v, A^*w\rangle_X \quad \text{and} \quad R_Y = AA^*.$$

4.3 Gradient-type algorithm

For solving (4.19), we consider the following basic gradient-type algorithm: letting $u^0 = 0$, we construct a sequence $\{u^k\}_{k\geq 0}$ in $\mathcal{M}_r(X)$ and a sequence $\{y^k\}_{k\geq 0}$ in Y defined for $k \geq 0$ by

$$\begin{cases} y^k = R_Y^{-1}(Au^k - b) \\ u^{k+1} \in \Pi_{\mathcal{M}_r(X)}(u^k - \rho R_X^{-1}A^*y^k) \end{cases} \tag{4.24}$$

with $\rho > 0$. Equations (4.24) yield

$$u^{k+1} \in \Pi_{\mathcal{M}_r(X)}(u + B_\rho(u^k - u)),$$

with $B_\rho = I_X - \rho R_X^{-1}A^*R_Y^{-1}A$ a symmetric operator from X to X. For all $v \in X$,

$$\frac{\langle B_\rho v, v\rangle_X}{\|v\|_X^2} = 1 - \rho\frac{\|Av\|_{Y'}^2}{\|v\|_X^2}.$$

Here, we assume that $\|\cdot\|_X$ and $\|\cdot\|_Y$ do not necessarily satisfy the relation (4.23) (i.e. $\frac{\alpha}{\beta} \neq 1$). From (4.11), we deduce that the eigenvalues of B_ρ are in the interval $[1 - \rho\beta^2, 1 - \rho\alpha^2]$. The spectral radius of B_ρ is therefore bounded by

$$\gamma(\rho) = \max\{|1 - \rho\beta^2|, |1 - \rho\alpha^2|\}.$$

Proposition 4.3. *Assuming $\gamma(\rho) < 1/2$, the sequence $\{u^k\}_{k\geq 1}$ defined by (4.24) is such that*

$$\|u^k - u\|_X \leq (2\gamma)^k \|u^0 - u\|_X + \frac{1}{1-2\gamma}\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X \tag{4.25}$$

and

$$\limsup_{k\to\infty} \|u^k - u\|_X \leq \frac{1}{1-2\gamma}\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X \tag{4.26}$$

Proof: Denoting $v^k = u^k - u$, we have

$$\begin{aligned}\|u^{k+1} - u\|_X &\leq \|\Pi_{\mathcal{M}_r(X)}(u + B_\rho v^k) - u\|_X \\ &\leq \|\Pi_{\mathcal{M}_r(X)}(u + B_\rho v^k) - (u + B_\rho v^k)\|_X + \|B_\rho v^k\|_X \\ &\leq \|w - (u + B_\rho v^k)\|_X + \|B_\rho v^k\|_X\end{aligned}$$

for all $w \in \mathcal{M}_r(X)$. In particular, this inequality is true for all $w \in \Pi_{\mathcal{M}_r(X)}(u)$, and therefore, taking the supremum over all $w \in \Pi_{\mathcal{M}_r(X)}(u)$, we obtain

$$\begin{aligned}\|u^{k+1} - u\|_X &\leq \|\Pi_{\mathcal{M}_r(X)}(u) - (u + B_\rho v^k)\|_X + \|B_\rho v^k\|_X \\ &\leq \|\Pi_{\mathcal{M}_r(X)}(u) - u\|_X + 2\|B_\rho v^k\|_X\end{aligned}$$

Since $\|B_\rho v\|_X \leq \gamma\|v\|_X$ for all $v \in X$ and since $2\gamma < 1$, we have

$$\begin{aligned}\|u^{k+1} - u\|_X &\leq \|\Pi_{\mathcal{M}_r(X)}(u) - u\|_X + 2\gamma\|u - u^k\|_X \\ &\leq (2\gamma)^{k+1}\|u^0 - u\|_X + \frac{1-(2\gamma)^{k+1}}{1-2\gamma}\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X\end{aligned}$$

from which we deduce (4.25) and (4.26). ∎

The condition $\gamma(\rho) < 1/2$ imposes $\frac{\beta}{\alpha} < \sqrt{3}$ and $\rho \in (\frac{1}{2\alpha^2}, \frac{3}{2\beta^2})$. The condition $\frac{\beta}{\alpha} < \sqrt{3}$ is a very restrictive condition which is in general not satisfied without an excellent preconditioning of the operator A.

However, with the ideal choice of norms introduced in the previous section (equation (4.23)), we have $\alpha = \beta = 1$ and $B_\rho = (1-\rho)I_X$. That means that the problem is ideally conditioned and we have convergence for all $\rho \in [\frac{1}{2}, \frac{3}{2}]$ towards a neighborhood of $\Pi_{\mathcal{M}_r(X)}(u)$ of size $\frac{2\gamma}{1-2\gamma}\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X$ with $\gamma = |1-\rho|$.

Corollary 4.4. *Assume that (4.23) is satisfied. Then, if $\rho \in [\frac{1}{2}, \frac{3}{2}]$, the sequence $\{u^k\}_{k\geq 1}$ defined by (4.24) verifies (4.25) and (4.26) with $\gamma(\rho) = |1-\rho|$. Moreover, if $\rho = 1$, then $u^1 \in \Pi_{\mathcal{M}_r(X)}(u)$, which means that the algorithm converges in one iteration for any initialization u^0.*

5 Perturbation of the ideal approximation

We now consider that function spaces X and Y are equipped with norms satisfying the ideal condition

$$\|Av\|_{Y'} = \|v\|_X \quad \forall v \in X. \tag{4.27}$$

The solution of problem (4.19) using this ideal choice of norms is therefore equivalent to the best approximation problem (4.13), i.e.

$$\min_{v\in\mathcal{M}_r(X)} \|Av - b\|_{Y'} = \min_{v\in\mathcal{M}_r(X)} \|v - u\|_X. \tag{4.28}$$

Unfortunately, the computation of the solution of (4.28) would require the solution of the initial problem. We here propose to introduce a computable perturbation of this ideal approach.

5.1 Approximation of the ideal approach

Following the idea of [39], the problem (4.28) is replaced by the following problem:

$$\min_{v\in\mathcal{M}_r(X)} \|\Lambda^\delta(R_Y^{-1}(Av - b))\|_Y, \tag{4.29}$$

where $\Lambda^\delta : Y \to Y$ is a mapping that provides an approximation $\Lambda^\delta(r)$ of the residual $r = R_Y^{-1}(Av - b) \in Y$ with a controlled relative precision $\delta > 0$, i.e. $\|\Lambda^\delta(r) - r\|_Y \leq \delta\|r\|_Y$. We will then assume that the mapping Λ^δ is such that:

$$\|\Lambda^\delta(y) - y\|_Y \leq \delta\|y\|_Y, \quad \forall y \in \mathcal{D}_Y = \left\{R_Y^{-1}(Av - b); v \in \mathcal{M}_r(X)\right\}. \tag{4.30}$$

As we will see in the following algorithm, it is sufficient for Λ^δ to well approximate residuals that are in the subset $\mathcal{D}_Y$ whose content depends on the chosen subset $\mathcal{M}_r(X)$ and on the operator and right-hand side of the problem.

5.2 Quasi-optimal approximations in $\mathcal{M}_r(X)$

Here we consider the case where we are not able to solve the best approximation problem in $\mathcal{M}_r(X)$ exactly, because there is no available algorithm for computing a global optimum, or because the algorithm has been stopped at a finite precision (see section 6.1 for practical comments). We introduce a set of quasi-optimal approximations $\Pi^{\eta}_{\mathcal{M}_r(X)}(u) \subset \mathcal{M}_r(X)$ such that

$$\|u - \Pi^{\eta}_{\mathcal{M}_r(X)}(u)\|_X \leq \eta \|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X \quad (\eta \geq 1). \tag{4.31}$$

Remark 5.1. *Note that by introducing this new perturbation, we are able to remove the assumption that $\mathcal{M}_r(X)$ is closed and to handle the case where the problem (4.28) does not have a solution, i.e. $\Pi_{\mathcal{M}_r(X)}(u) = \emptyset$. In this case, we have to replace $\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X$ by $\inf_{w \in \mathcal{M}_r(X)} \|u - w\|_X$ in equation (4.31).*

Remark 5.2. *Note that if $\mathcal{M}_r(X)$ denotes a low-rank subset of an infinite dimensional space X, additional approximations have to be introduced from a numerical point of view (see remark 3.1). These additional approximations could be also considered as a perturbation leading to quasi-optimal approximations, where η takes into account the approximation errors. In numerical examples, we will not adopt this point of view and we will consider X as the approximation space and the approximate solution in X of the variational problem will serve as a reference solution.*

5.3 Perturbed gradient-type algorithm

For solving (4.29), we now introduce an algorithm which can be seen as a perturbation of the ideal gradient-type algorithm (4.24) introduced in section 4.3. Letting $u^0 = 0$, we construct a sequence $\{u^k\}_{k\geq 0} \subset \mathcal{M}_r(X)$ and a sequence $\{y^k\}_{k\geq 0} \subset Y$ defined for $k \geq 0$ by

$$\begin{cases} y^k = \Lambda^{\delta}(R_Y^{-1}(Au^k - b)) \\ u^{k+1} \in \Pi^{\eta}_{\mathcal{M}_r(X)}(u^k - R_X^{-1}A^* y^k) \end{cases} \tag{4.32}$$

Proposition 5.3. *Assume* (4.27), (4.30), *and* (4.31), *with* $\delta(1+\eta) < 1$. *Then, the sequence* $\{u^k\}_{k\geq 1}$ *defined by* (4.32) *is such that*

$$\|u^k - u\|_X \leq ((1+\eta)\delta)^k \|u^0 - u\|_X + \frac{\eta}{1-\delta(1+\eta)} \|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X. \tag{4.33}$$

Proof: Equation (4.32) can also be written

$$u^{k+1} \in \Pi^{\eta}_{\mathcal{M}_r(X)}(u + B^{\delta}(u^k - u))$$

with $B^{\delta}(v) = v - R_X^{-1} A^* \Lambda^{\delta}(R_Y^{-1} A(v))$. Denoting $v^k = u^k - u$, and following the proof of Proposition 4.3, we obtain

$$\begin{aligned}
\|u^{k+1} - u\|_X &\leq \|\Pi^{\eta}_{\mathcal{M}_r(X)}(u + B^{\delta} v^k) - (u + B^{\delta} v^k)\|_X + \|B^{\delta} v^k\|_X \\
&\leq \eta \|\Pi_{\mathcal{M}_r(X)}(u) - (u + B^{\delta} v^k)\|_X + \|B^{\delta} v^k\|_X \\
&\leq \eta \|\Pi_{\mathcal{M}_r(X)}(u) - u\|_X + (1+\eta)\|B^{\delta} v^k\|_X
\end{aligned}$$

Moreover, using (4.27) and (4.21), we have

$$\begin{aligned}
\|B^{\delta} v^k\|_X &= \|v^k - R_X^{-1} A^* \Lambda^{\delta}(R_Y^{-1} A v^k)\|_X \\
&= \|A v^k - A R_X^{-1} A^* \Lambda^{\delta}(R_Y^{-1} A v^k)\|_{Y'} \\
&= \|R_Y^{-1} A v^k - \Lambda^{\delta}(R_Y^{-1} A v^k)\|_Y.
\end{aligned}$$

Noting that $R_Y^{-1} A v^k = R_Y^{-1}(A u^k - b)$ belongs to the subset $\mathcal{D}_Y$, we deduce from assumption (4.30) and equation (4.27) that

$$\|B^{\delta} v^k\|_X \leq \delta \|R_Y^{-1} A v^k\|_Y = \delta \|v^k\|_X.$$

Denoting $\delta_\eta = \delta(1+\eta) < 1$, we finally have

$$\begin{aligned}
\|u^{k+1} - u\|_X &\leq \eta \|\Pi_{\mathcal{M}_r(X)}(u) - u\|_X + \delta_\eta \|u^k - u\|_X \\
&\leq \delta_\eta^{k+1} \|u^0 - u\|_X + \eta \frac{1 - \delta_\eta^{k+1}}{1 - \delta_\eta} \|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X,
\end{aligned}$$

from which we deduce (4.33). ∎

Comments We note the sequence converges towards a neighborhood of $\Pi_{\mathcal{M}_r(X)}(u)$ whose size is $\frac{\eta - 1 + (1+\eta)\delta}{1-(1+\eta)\delta} \|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X$. Indeed, (4.33) implies that

$$\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X \leq \|u - u^k\|_X \leq (1+\gamma_k)\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X, \tag{4.34}$$

with $\limsup_{k\to\infty}\gamma_k \leq \frac{\eta-1+(1+\eta)\delta}{1-(1+\eta)\delta}$. Therefore, the sequence tends to provide a good approximation of the best approximation of u in $\mathcal{M}_r(X)$, and the parameters δ and η control the quality of this approximation. Moreover, equation (4.33) indicates that the sequence converges quite rapidly to this neighborhood. Indeed, in the first iterations, when the error $\|u-u^k\|_X$ is dominated by the first term $((1+\eta)\delta)^k\|u-u^0\|_X$, the algorithm has at least a linear convergence with convergence rate $(1+\eta)\delta$ (note that for $\eta \approx 1$, the convergence rate is very high for small δ). Once both error terms are balanced, the error stagnates at the value $\frac{\eta}{1-(1+\eta)\delta}\|u-\Pi_{\mathcal{M}_r(X)}(u)\|_X$. Note that when $\delta \to 0$, we recover an ideal algorithm with a convergence in only one iteration to an element of the set $\Pi^{\eta}_{\mathcal{M}_r(X)}(u)$ of quasi-best approximations of u in $\mathcal{M}_r(X)$.

Remark 5.4. *Even if $\mathcal{M}_r(X)$ is chosen as a subset of low-rank tensors, the subset $\mathcal{D}_Y$ defined in (4.30) possibly contains tensors with high ranks (or even tensors with full rank) that are not easy to approximate with a small precision δ using low-rank tensor representations. However, the algorithm only requires to well approximate the sequence of residuals $\{R_Y^{-1}(Au^k-b)\}_{k\geq 0} \subset \mathcal{D}_Y$, which may be achievable in practical applications.*

5.4 Error indicator

Along the iterations of algorithm (4.32), an estimation of the true error $\|u-u^k\|_X$ can be simply obtained by evaluating the norm $\|y^k\|_Y$ of the iterate $y^k = \Lambda^\delta(r^k)$ with $r^k = R_Y^{-1}(Au^k-b)$. Indeed, from property (4.30), we have

$$(1-\delta)\|y\|_Y \leq \|\Lambda^\delta(y)\|_Y \leq (1+\delta)\|y\|_Y, \tag{4.35}$$

for all $y \in \mathcal{D}_Y$. Therefore, noting that $r^k \in \mathcal{D}_Y$ and $\|r^k\|_Y = \|A(u-u^k)\|_{Y'} = \|u-u^k\|_X$, we obtain

$$(1-\delta)\|u-u^k\|_X \leq \|y^k\|_Y \leq (1+\delta)\|u-u^k\|_X. \tag{4.36}$$

In other words,

$$\epsilon^k = \frac{1}{1-\delta}\|y^k\|_Y \tag{4.37}$$

provides an error indicator of the true error $\|u-u^k\|_X$ with an effectivity index $\tau^k = \frac{\epsilon^k}{\|u-u^k\|_X} \in (1, \frac{1+\delta}{1-\delta})$, which is very good for small δ.

Moreover, if Λ^δ is an orthogonal projection onto some subspace $Y^\delta \subset Y$, we easily obtain the following improved lower and upper bounds:

$$\sqrt{1-\delta^2}\|u-u^k\|_X \leq \|y^k\|_Y \leq \|u-u^k\|_X, \tag{4.38}$$

that means that the following improved error estimator can be chosen:

$$\hat{\epsilon}^k = \frac{1}{\sqrt{1-\delta^2}} \|y^k\|_Y, \tag{4.39}$$

with effectivity index $\hat{\tau}^k = \frac{\hat{\epsilon}^k}{\|u-u^k\|_X} \in (1, \frac{1}{\sqrt{1-\delta^2}})$.

6 Computational aspects

6.1 Best approximation in tensor subsets

We here discuss the available algorithms for computing an element in $\Pi_{\mathcal{M}_r(X)}(v)$, that means for solving

$$\min_{w \in \mathcal{M}_r(X)} \|v - w\|_X, \tag{4.40}$$

where v is a given tensor in the tensor space $X = {}_{\|\cdot\|_X}\bigotimes_{\mu=1}^d X_\mu$ equipped with norm $\|\cdot\|_X$, and where $\mathcal{M}_r(X)$ is a given tensor subset. Note that except for the case where $d = 2$ and $\|\cdot\|_X$ is the induced (canonical) norm, the computation of a global optimum is still an open problem.

Canonical norm, $d = 2$**.** For the case $d = 2$, we first note that all classical low-rank tensor formats coincide with the canonical format, that means $\mathcal{M}_r(X) = \mathcal{C}_r(X)$ for some rank r. When the norm $\|\cdot\|_X$ is the canonical norm, then $u_r \in \Pi_{\mathcal{M}_r(X)}(u)$ coincides with a rank-r singular value decomposition (SVD) of u (which is possibly not unique in the case of multiple singular values). Moreover, $\sigma(u; \mathcal{M}_r(X))^2 = \|\Pi_{\mathcal{M}_r(X)}(u)\|_X^2$ is the sum of the squares of the r dominant singular values of u (see e.g. [58]). Efficient algorithms for computing the SVD can therefore be applied to compute an element in $\Pi_{\mathcal{M}_r(X)}(v)$ (a best approximation). That means that the algorithm (4.32) can be applied with $\eta = 1$.

Canonical norm, $d > 2$**.** For $d > 2$ and when the norm $\|\cdot\|_X$ is the canonical norm, different algorithms based on optimization methods have been proposed for the different tensor formats (see e.g. [54,81] or [76] for a recent review). Very efficient algorithms based on higher order SVD have also been proposed in [44], [71] and [108], respectively for Tucker, Hierarchical Tucker and Tensor Train tensors. Note that these algorithms provide quasi-best approximations (but not necessarily best approximations) satisfying (4.31) with a η bounded by a function of the dimension d: $\eta \leq \sqrt{d}$, $\eta \leq \sqrt{2d-3}$ respectively for Tucker and Hierarchical Tucker formats

(see [76]). For a high dimension d, such bounds for η would suggest taking very small values for parameter δ in order to satisfy the assumption of Proposition 5.3. However, in practice, these a priori bounds are rather pessimistic. Moreover, quasi-best approximations obtained by higher order SVD can be used as initializations of optimization algorithms yielding better approximations, i.e. with small values of η.

General norms, $d \geq 2$**.** For a general norm $\|\cdot\|_X$, the computation of a global optimum to the best approximation problem is still an open problem for all tensor subsets, and methods based on SVD cannot be applied anymore. However, classical optimization methods can still be applied (such as Alternating Least Square (ALS)) in order to provide an approximation of the best approximation [54, 113, 122]. We do not detail further these computational aspects and we suppose that algorithms are available for providing an approximation of the best approximation in $\mathcal{M}_r(X)$ such that (4.31) holds with a controlled precision η, arbitrarily close to 1.

6.2 Construction of an approximation of $\Lambda^\delta(r)$

At each iteration of the algorithm (4.32), we have to compute $y^k = \Lambda^\delta(r^k)$, with $r^k = R_Y^{-1}(Au^k - b) \in Y$, such that it satisfies

$$\|y^k - r^k\|_Y \leq \delta \|r^k\|_Y. \tag{4.41}$$

First note that r^k is the unique solution of

$$\min_{r \in Y} \|r - R_Y^{-1}(Au^k - b)\|_Y^2. \tag{4.42}$$

Therefore, computing y^k is equivalent to solving the best approximation problem (4.42) with a relative precision δ. One can equivalently characterize $r^k \in Y$ by the variational equation

$$\langle r^k, \delta r\rangle_Y = \langle Au^k - b, \delta r\rangle_{Y',Y} \quad \forall \delta r \in Y,$$

or in an operator form:

$$R_Y r^k = Au^k - b, \tag{4.43}$$

where the Riesz map $R_Y = AR_X^{-1}A^*$ is a positive symmetric definite operator.

Remark 6.1. *For A symmetric and positive definite, it is possible to choose $R_X = R_Y = A$ (see example 4.2) that corresponds to the energy norm on X. For this*

choice, the auxiliary problem (4.42) has the same structure as the initial problem, with an operator A and a right-hand side $Au^k - b$.

6.2.1 Low-rank tensor methods

For solving (4.42), we can also use low-rank tensor approximation methods. Note that in general, $\|\cdot\|_Y$ is not an induced (canonical) norm in Y, so that classical tensor algorithms (e.g. based on SVD) cannot be applied for solving (4.42) (even approximatively). Different strategies have been proposed in the literature for constructing tensor approximations of the solution of optimization problems. We can either use iterative solvers using classical tensor approximations applied to equation (4.43) [8, 86, 89, 98], or directly compute an approximation y^k of r^k in low-rank tensor subsets using optimization algorithms applied to problem (4.42). Here, we adopt the latter strategy and rely on a greedy algorithm which consists in computing successive corrections of the approximation in a fixed low-rank subset.

6.2.2 A possible (heuristic) algorithm

We use the following algorithm for the construction of a sequence of approximations $\{y_m^k\}_{m\geq 0}$.

Let $y_0^k = 0$. Then, for each $m \geq 1$, we proceed as follows:

1. compute an optimal correction w_m^k of y_{m-1}^k in $\mathcal{M}_r(Y)$:
$$w_m^k \in \arg\min_{w\in\mathcal{M}_r(Y)} \|y_{m-1}^k + w - r^k\|_Y,$$
2. define a linear subspace Z_m^k such that $y_{m-1}^k + w_m^k \in Z_m^k$,
3. compute y_m^k as the best approximation of r^k in Z_m^k,
$$y_m^k = \arg\min_{y\in Z_m^k} \|y - r^k\|_Y,$$
4. return to step (2) or (1).

Remark 6.2. *The convergence proof for this algorithm can be found in [59]. The convergence ensures that the precision δ can be achieved after a certain number of iterations.*[a] *However, in practice, best approximation problems at step (1) can not be solved exactly except for particular situations (see section 6.1), so that the results of [59] do not guaranty anymore the convergence of the algorithm. If quasi-optimal solutions can be obtained, this algorithm is a modified version*

of weak greedy algorithms (see [119]) for which convergence proofs can also be obtained. Available algorithms for obtaining quasi-optimal solutions of best low-rank approximation problem appearing at step (1) are still heuristic but seem to be effective.

[a]Note however that a slow convergence of these algorithms may yield to high rank representations of iterates y_m^k, even for a low-rank subset $\mathcal{M}_r(Y)$.

In this chapter, we will only rely on the use of low-rank canonical formats for numerical illustrations. At step (1), we introduce rank-one corrections $w_m^k \in \mathcal{M}_r(Y) = \mathcal{C}_1(Y)$, where $Y = {}_{\|\cdot\|_Y}\bigotimes_{\mu=1}^d Y^\mu$. The auxiliary variable $y_m^k \in \mathcal{C}_m(Y)$ can be written in the form $y_m^k = \sum_{i=1}^m \otimes_{\mu=1}^d w_i^{k,\mu}$. At step (2), we select a particular dimension $\mu \in \{1, \cdots, d\}$ and define

$$Z_m^k = \left\{ \sum_{i=1}^m w_i^{k,1} \otimes \cdots \otimes v_i^\mu \otimes \cdots \otimes w_i^{k,d} \ , \ v_i^\mu \in Y^\mu \right\},$$

where $\dim(Z_m^k) = m \dim(Y^\mu)$. Step (3) therefore consists in updating functions $w_i^{k,\mu}$, $i = 1 \cdots d$, in the representation of y_m^k. Before returning to step (1), the updating steps (2)-(3) can be performed several times for a set of dimension $\mu \in I \subset \{1, \ldots, d\}$.

Remark 6.3. *Note that the solution of minimization problems at steps (1) and (3) do not require to know r^k explicitly. Indeed, the stationary conditions associated with these optimization problems only require the evaluation of $\langle r^k, \delta y\rangle_Y = \langle Au^k - b, \delta y\rangle_{Y',Y}$, for $\delta Y \in Y$. For step (1), the stationary equation reads $\langle R_Y w_m^k, \delta y\rangle_{Y',Y} = \langle R_Y y_{m-1}^k + Au^k - b, \delta y\rangle_{Y',Y}$ for all δy in the tangent space to $\mathcal{M}_r(Y)$, while the variational form of step (3) reads $\langle R_Y y_m^k, \delta y\rangle_{Y',Y} = \langle Au^k - b, \delta y\rangle_{Y',Y}$ for all δy in Z_m^k.*

Finally, as a stopping criterion, we use a heuristic error estimator based on stagnation. The algorithm is stopped at iteration m if

$$e_m^p = \frac{\|y_m^k - y_{m+p}^k\|_Y}{\|y_{m+p}^k\|_Y} \leq \delta, \tag{4.44}$$

for some chosen $p \geq 1$ (typically $p = 1$). Note that for p sufficiently large, y_{m+p}^k can be considered as a good estimation of the residual r^k and the criterion reads $\|r^k - y_m^k\|_Y \leq \delta \|r^k\|_Y$, which is the desired property. This stopping criterion is quite rudimentary and should be improved for a real control of the algorithm. Although numerical experiments illustrate that this heuristic error estimator provides

a rather good approximation of the true error, an upper bound of the true error should be used in order to guarantee that the precision δ is really achieved. However, a tight error bound should be used in order to avoid a pessimistic overestimation of the true error which may yield an (unnecessary) increase of the computational costs for the auxiliary problem. This key issue will be addressed in a future work.

Remark 6.4. *Other updating strategies could be introduced at steps (2)-(3). For example, we could choose $Z_m^k = \text{span}\{w_1^k, \cdots, w_m^k\}$, thus making the algorithm an orthogonal greedy algorithm with a dictionary $\mathcal{M}_r(Y)$ [119]. Nevertheless, numerical simulations demonstrate that when using rank-one corrections (i.e. $\mathcal{M}_r(Y) = \mathcal{C}_1(Y)$), this updating strategy do not significantly improve the convergence of pure greedy constructions. When it is used for obtaining an approximation y_m^k of r^k with a small relative error δ, it usually requires a very high rank m. A more efficient updating strategy consists in defining Z_m^k as the tensor space $\bigotimes_{\mu=1}^d Z_m^{k,\mu}$ with $Z_m^{k,\mu}$ span$\{w_1^{k,\mu}, \cdots, w_m^{k,\mu}\}$. Since $\dim(Z_m^k) = m^d$, the projection of r^k in Z_m^k can not be computed exactly for high dimensions d. However, approximations of this projection can be obtained using again low-rank formats (see [64]).*

6.2.3 Remark on the tensor structure of Riesz maps

We consider that operator A and right-hand side b admit low-rank representations

$$A = \sum_{i=1}^{r_A} \otimes_{\mu=1}^d A_i^\mu \quad \text{and} \quad b = \sum_{i=1}^{r_b} \otimes_{\mu=1}^d b_i^\mu.$$

We suppose that a norm $\|\cdot\|_X$ has been selected and corresponds to a Riesz map R_X with a low-rank representation:

$$R_X = \sum_{i=1}^{r_X} \otimes_{\mu=1}^d R_i^\mu.$$

The ideal choice of norm $\|\cdot\|_Y$ then corresponds to the following expression of the Riesz map R_Y:

$$R_Y = AR_X^{-1}A^* = (\sum_{i=1}^{r_A} \otimes_{\mu=1}^d A_i^\mu)(\sum_{i=1}^{r_X} \otimes_{\mu=1}^d R_i^\mu)^{-1}(\sum_{i=1}^{r_A} \otimes_{\mu=1}^d A_i^{\mu *}).$$

Note that the expression of R_Y cannot be computed explicitly (R_Y is generally a full rank tensor). Therefore, in the general case, algorithms for solving problem (4.43)

have to be able to handle an implicit formula for R_Y. However, in the particular case where the norm $\|\cdot\|_X$ is a canonical norm induced by norms $\|\cdot\|_\mu$ on X_μ, the mapping R_X is a rank one tensor $R_X = \otimes_{\mu=1}^d R_{X_\mu}$, where R_{X_μ} is the Riesz map associated with the norm $\|\cdot\|_\mu$ on X_μ. R_Y then admits the following explicit expression:

$$R_Y = AR_X^{-1}A^* = \sum_{i=1}^{r_A}\sum_{j=1}^{r_A} \otimes_{\mu=1}^d (A_i^\mu R_{X_\mu}^{-1} A_j^{\mu *}).$$

In the numerical examples, we only consider this simple particular case.

6.3 Summary of the algorithm

Algorithm 7 provides a step-by-step outline of the overall iterative method for the approximation of the solution of (4.28) in a fixed subset $\mathcal{M}_r(X)$ and with a chosen metric $\|\cdot\|_X$. Given a precision δ, an approximation of the residual is obtained with a greedy algorithm using a fixed subset $\mathcal{M}_r(Y)$ for computing successive corrections. We denote by $e(y_m^k, r^k)$ an estimation of the relative error $\|y_m^k - r^k\|_Y / \|r^k\|_Y$, where $r^k = R_Y^{-1}(Au^k - b)$.

Algorithm 7 Gradient-type algorithm

1: Set $u^0 = 0$;
2: **for** $k = 0$ to K **do**
3: Set $m = 0$;
4: **while** $e(y_m^k, r^k) \leq \delta$ **do**
5: $m = m + 1$;
6: Compute a correction $w_m^k \in \arg\min_{w \in \mathcal{M}_r(Y)} \|y_{m-1}^k + w - r^k\|_Y$;
7: Set $y_m^k = y_{m-1}^k + w_m^k$;
8: Define Z_m^k containing y_m^k ;
9: Compute the projection $y_m^k = \arg\min_{y \in Z_m^k} \|y - r^k\|_Y$;
10: Return to step 7 or continue ;
11: **end while**
12: Compute $u^{k+1} \in \Pi^\eta_{\mathcal{M}_r(X)}(u^k - R_X^{-1}A^* y_m^k)$;
13: **end for**

7 Greedy algorithm

In this section, we introduce and analyze a greedy algorithm for the progressive construction of a sequence $\{u_m\}_{m\geq 0}$, where u_m is obtained by computing a correction

of u_{m-1} in a given low-rank tensor subset $\mathcal{M}_r(X)$ (typically a small subset such as the set of rank-one tensors $\mathcal{C}_1(X)$). Here, we consider that approximations of optimal corrections are available with a certain precision. It results in an algorithm which can be considered as a modified version of weak greedy algorithms [119]. This weak greedy algorithm can be applied to solve the best approximation problem (4.19) where approximations of optimal corrections are obtained using Algorithm 7 with an updated right-hand side at each greedy step. The interest of such a global greedy strategy is twofold. First, an adaptive approximation strategy which would consist in solving approximation problems in an increasing sequence of low-rank subsets $\mathcal{M}_r(X)$ is often unpractical since for high dimensional problems and subspace based tensor formats, computational complexity drastically increases with the rank. Second, it simplifies the solution of auxiliary problems (i.e. the computation of the sequence of y^k) when solving best low-rank approximation problems using Algorithm 7. Indeed, if the sequence u^k in Algorithm 7 belongs to a low-rank tensor subset (typically a rank-one tensor subset), the residual r^k in Algorithm 7 admits a moderate rank or can be obtained by a low-rank correction of the residual of the previous greedy iteration.

Here, we assume that the subset $\mathcal{M}_r(X)$ verifies properties (4.14) and (4.15), and that $\mathrm{span}(\mathcal{M}_r(X))$ is dense in X (which is verified by all classical tensor subsets presented in section 3.2).

7.1 A weak greedy algorithm

We consider the following greedy algorithm. Given $u_0 = 0$, we construct a sequence $\{u_m\}_{m\geq 1}$ defined for $m \geq 1$ by

$$u_m = u_{m-1} + \widetilde{w}_m, \tag{4.45}$$

where $\widetilde{w}_m \in \mathcal{M}_r(X)$ is a correction of u_{m-1} satisfying

$$\|u - u_{m-1} - \widetilde{w}_m\|_X \leq (1+\gamma_m) \min_{w\in\mathcal{M}_r(X)} \|u - u_{m-1} - w\|_X, \tag{4.46}$$

with γ_m a sequence of small parameters.

Remark 7.1. *A $\widetilde{w}_m$ satisfying (4.46) can be obtained using the gradient type algorithm of section 5 that provides a sequence that satisfies (4.34). Given the parameter $\delta = \delta_m$ in (4.32), property (4.46) can be achieved with any $\gamma_m > \frac{2\delta_m}{1-2\delta_m}$.*

7.2 Convergence analysis

Here, we provide a convergence result for the above greedy algorithm whose proof follows the lines of [119] for the convergence proof of weak greedy algorithms[2].

In the following, we denote by $f_m = u - u_m$. For the sake of simplicity, we denote by $\|\cdot\| = \|\cdot\|_X$ and $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_X$ and we let $w_m \in \Pi_{\mathcal{M}_r(X)}(f_{m-1})$, for which we have the following useful relations coming from properties of best approximation problems in tensor subsets (see section 3.2):

$$\|f_{m-1} - w_m\|^2 = \|f_{m-1}\|^2 - \|w_m\|^2 \quad \text{and} \quad \|w_m\|^2 = \langle f_{m-1}, w_m\rangle. \tag{4.47}$$

We introduce the sequence $\{\alpha_m\}_{m\geq 1}$ defined by

$$\alpha_m = \frac{\|f_{m-1} - w_m\|}{\|f_{m-1}\|} \in [0, 1[. \tag{4.48}$$

It can be also useful to introduce the computable sequence $\{\widetilde{\alpha}_m\}_{m\geq 1}$ such that

$$\widetilde{\alpha}_m = \frac{\|f_{m-1} - \widetilde{w}_m\|}{\|f_{m-1}\|}. \tag{4.49}$$

that satisfies for all $m \leq 0$

$$\alpha_m \leq \widetilde{\alpha}_m \leq (1 + \gamma_m)\alpha_m. \tag{4.50}$$

Lemma 7.2. *Assuming that for all $m \geq 1$ we have*

$$(1 + \gamma_m)\alpha_m < 1, \tag{4.51}$$

the sequence $\{\|f_m\|\}_{m\geq 1}$ converges. Furthermore, it is possible to define a positive sequence $\{\kappa_m\}_{m\geq 1}$ as

$$\kappa_m^2 = 2\frac{\langle f_{m-1}, \widetilde{w}_m\rangle}{\|\widetilde{w}_m\|^2} - 1, \tag{4.52}$$

and we have $\{\kappa_m\|\widetilde{w}_m\|\}_{m\geq 1} \in \ell^2$.

[2]Note that the condition (4.46) on the successive corrections does not allow to directly apply the results on classical weak greedy algorithms.

Proof: From (4.45) and (4.46), we have

$$\|f_m\| = \|f_{m-1} - \widetilde{w}_m\| \leq (1+\gamma_m)\|f_{m-1} - w_m\| = (1+\gamma_m)\alpha_m\|f_{m-1}\|.$$

Under assumption (4.51), $\{\|f_m\|\}_{m\geq 1}$ is a strictly decreasing and positive sequence and therefore converges. Moreover, this implies that $\widetilde{w}_m \neq 0$ and since

$$\|f_{m-1} - \widetilde{w}_m\|^2 = \|f_{m-1}\|^2 - \left(2\langle f_{m-1}, \widetilde{w}_m\rangle - \|\widetilde{w}_m\|^2\right) \leq \|f_{m-1}\|^2,$$

it follows that $2\langle f_{m-1}, \widetilde{w}_m\rangle > \|\widetilde{w}_m\|^2$. Therefore, κ_m is positive and can be defined by (4.52) and we have

$$\|f_{m-1} - \widetilde{w}_m\|^2 = \|f_{m-1}\|^2 - \kappa_m^2\|\widetilde{w}_m\|^2 = \|f_0\|^2 - \sum_{i=1}^{m} \kappa_i^2\|\widetilde{w}_i\|^2,$$

that completes the proof. ∎

We now provide a result giving a relation between $\|w_m\|$ and $\|\widetilde{w}_m\|$.

Lemma 7.3. *Assume (4.51) holds and let* $\mu_m^2 = \dfrac{1-(1+\gamma_m)^2\alpha_m^2}{1-\alpha_m^2} \in [0,1]$. *Then, we have*

$$\mu_m\|w_m\| \leq \kappa_m\|\widetilde{w}_m\| \leq \|w_m\|, \tag{4.53}$$

and

$$\frac{\mu_m}{2} \leq \kappa_m. \tag{4.54}$$

Proof: From inequality (4.46) and from the optimality of w_m, it follows that

$$\begin{aligned}
&\|f_{m-1} - w_m\|^2 \leq \|f_{m-1} - \widetilde{w}_m\|^2 \leq (1+\gamma_m)^2\|f_{m-1} - w_m\|^2 \\
\Rightarrow\ &\|f_{m-1}\|^2 - \|w_m\|^2 \leq \|f_{m-1}\|^2 - \kappa_m^2\|\widetilde{w}_m\|^2 \leq (1+\gamma_m)^2\alpha_m^2\|f_{m-1}\|^2 \\
\Rightarrow\ &(1-(1+\gamma_m)^2\alpha_m^2)\|f_{m-1}\|^2 \leq \kappa_m^2\|\widetilde{w}_m\|^2 \leq \|w_m\|^2
\end{aligned}$$

Using $\|f_{m-1}\|^2 = \|f_{m-1} - w_m\|^2 + \|w_m\|^2 = \alpha_m^2\|f_{m-1}\|^2 + \|w_m\|^2$, and using the definition of μ_m, we obtain (4.53). In addition, from the optimality of w_m, we have $\langle \frac{\widetilde{w}_m}{\|\widetilde{w}_m\|}, f_{m-1}\rangle \leq \langle \frac{w_m}{\|w_m\|}, f_{m-1}\rangle = \|w_m\|$, or equivalently $\frac{\kappa_m^2+1}{2}\|\widetilde{w}_m\| \leq \|w_m\|$. Combined with (4.53), it gives $\frac{\kappa_m^2+1}{2} \leq \frac{\|w_m\|}{\|\widetilde{w}_m\|} \leq \frac{\kappa_m}{\mu_m}$, which implies (4.54). ∎

Proposition 7.4. *Assume* (4.51) *and that* $\{\mu_m^2\}_{m\geq 1}$ *is such that* $\sum_{m=1}^{\infty}\mu_m^2=\infty$. *Then, if* $\{f_m\}_{m\geq 1}$ *converges, it converges to zero.*

Proof: Let us use a proof by contradiction. Assume that $f_m \to f \neq 0$ as $m\to\infty$, with $f\in X$. As $\mathrm{span}(\mathcal{M}_r(X))$ is dense in X, there exists $\epsilon>0$ such that $\sup_{v\in\mathcal{M}_r(X)}|\langle f,\frac{v}{\|v\|}\rangle|\geq 2\varepsilon$. Using the definition of w_m and of f as a limit of f_m, we have that there exists $N>0$ such that

$$\|w_m\| = \sup_{v\in\mathcal{M}_r(X)}|\langle f_{m-1},\frac{v}{\|v\|}\rangle|\geq\varepsilon,\quad \forall m\geq N. \tag{4.55}$$

Thanks to (4.53), we have

$$\begin{aligned}\|f_m\|^2 &= \|f_{m-1}\|^2-\|\widetilde{w}_m\|^2\kappa_m^2\leq\|f_{m-1}\|^2-\|w_m\|^2\mu_m^2,\\ &\leq\|f_N\|^2-\sum_{i=N+1}^{m}\mu_i^2\|w_i\|^2\leq\|f_N\|^2-\varepsilon^2\sum_{i=N+1}^{m}\mu_i^2,\end{aligned}$$

which implies that $\{\mu_m\}_{m\geq 0}\in\ell^2$, a contradiction to the assumption.

■

Proposition 7.5. *Assume* (4.51). *Further assume that the sequence* μ_m *is non increasing and verifies*

$$\sum_{m=1}^{\infty}\frac{\mu_m^2}{m}=\infty. \tag{4.56}$$

Then the sequence $\{u_m\}_{m\geq 1}$ *converges to* u.

Proof: Let two integers $n<m$ and consider

$$\|f_n-f_m\|^2=\|f_n\|^2-\|f_m\|^2-2\langle f_n-f_m,f_m\rangle.$$

Defining $\theta_{n,m}=|\langle f_n-f_m,f_m\rangle|$ and using Lemma 7.3, we obtain

$$\theta_{n,m}\leq\sum_{i=n+1}^{m}|\langle\widetilde{w}_i,f_m\rangle|\leq\|w_{m+1}\|\sum_{i=1}^{m}\|\widetilde{w}_i\|\leq 2\frac{\kappa_{m+1}\|\widetilde{w}_{m+1}\|}{\mu_{m+1}^2}\sum_{i=1}^{m}\kappa_i\|\widetilde{w}_i\|.$$

Lemma 7.2 implies that $\kappa_m\|\widetilde{w}_m\|\in\ell^2$. Together with assumption (4.56), and using Lemma 2.7 in [119], we obtain that $\liminf_{m\to\infty}\max_{n<m}\theta_{n,m}=0$. Lemma

2.8 in [119] then proves that the sequence $\{f_m\}_{m\geq 1}$ converges. Noting that (4.56) implies that $\{\mu_m\}_{m=1}^{\infty} \notin \ell^2$, Lemma 7.4 allows to conclude the proof. ∎

In practice, condition (4.56) can be satisfied by the following sufficient condition on the sequence $\widetilde{\alpha}_m$, which is a computable sequence.

Corollary 7.6. *If there exists a constant $0 < \epsilon < 1$, independent of m, such that*

$$\widetilde{\alpha}_m^2 \leq \frac{1-\epsilon}{(1+\gamma_m)^2 - \epsilon}, \tag{4.57}$$

then the sequence $\{u_m\}_{m\geq 1}$ converges to u.

Proof: Under assumption (4.57) and using relation (4.50), it holds that for all $m \geq 0$

$$\alpha_m^2 \leq \frac{1-\epsilon}{(1+\gamma_m)^2 - \epsilon} \quad \Rightarrow \quad (1+\gamma_m)^2\alpha_m^2 \leq 1 - \epsilon(1-\alpha_m^2) < 1.$$

which implies condition (4.51). Moreover, we have

$$\epsilon(1-\alpha_m^2) \leq 1 - (1+\gamma_m)^2\alpha_m^2 \quad \Rightarrow \quad \epsilon \leq \frac{1-(1+\gamma_m)^2\alpha_m^2}{(1-\alpha_m^2)} = \mu_m^2,$$

which implies condition (4.56). Proposition 7.5 ends the proof. ∎

Remark 7.7. *From a practical point of view, condition (4.57) provides a sufficient criterion on γ_m (or equivalently on δ_m). Note that $\widetilde{\alpha}_m$ depends on $\widetilde{w}_m$ which depends on the choice of the precision γ_m. Therefore, (4.57) is an implicit condition on γ_m which suggests an iterative strategy for the control of the condition. A possible strategy would be to adapt the parameter γ_m during the iterations of the gradient type algorithm used to compute the $\widetilde{w}_m$.*

8 Numerical example

In this section, we apply the proposed method to the numerical solution of a stochastic steady reaction-advection-diffusion problem.

8.1 Stochastic reaction-advection-diffusion problem

We consider the following steady reaction-advection-diffusion problem on a two-dimensional unit square domain $\Omega = [0,1]^2$ (see Figure 4.1):

$$\begin{aligned} -\nabla \cdot (\kappa \nabla u) + c \cdot \nabla u + au &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.58}$$

First, we consider a constant diffusion $\kappa = 1$. The advection coefficient c and the reaction coefficient a are considered as random and are given by $c = \xi_1 c_0$ and $a = \exp(\xi_2)$, where $\xi_1 \sim U(-350, 350)$ and $\xi_2 \sim U(\log(0.1), \log(10))$ are independent uniform random variables, and $c_0(x) = (x_2 - 1/2, 1/2 - x_1)$, $x = (x_1, x_2) \in \Omega$. We denote by $\Xi_1 =]\text{-}350, 350[$ and $\Xi_2 =]\log(0.1), \log(10)[$, and we denote by $(\Xi, \mathcal{B}(\Xi), P_\xi)$ the probability space induced by $\xi = (\xi_1, \xi_2)$, with $\Xi = \Xi_1 \times \Xi_2$ and P_ξ the probability law of ξ. The external source term f is given by $f(x) = I_{\Omega_1}(x) - I_{\Omega_2}(x)$, where $\Omega_1 =]0.45, 0.55[\times]0.15, 0.25[$ and $\Omega_2 =]0.45, 0.55[\times]0.75, 0.85[$, and where I_{Ω_k} denotes the indicator function of Ω_k.

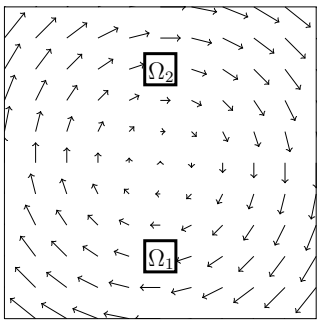


Figure 4.1: Example : reaction-advection-diffusion problem.

Let $V = \mathrm{H}_0^1(\Omega)$ and $S = \mathrm{L}^2(\Xi, dP_\xi)$. We introduce approximation spaces $V_N \subset V$ and $S_P \subset S$, with $N = \dim(V_N)$ and $P = \dim(S_P)$. V_N is a $\mathbb{Q}_1$ finite element space associated with a uniform mesh of 1600 elements such that $N = 1521$. We choose $S_P = S_{p_1}^{\xi_1} \otimes S_{p_2}^{\xi_2}$, where $S_{p_1}^{\xi_1}$ is the space of piecewise polynomials of degree 5 on Ξ_1 associated with the partition $\{]\text{-}350, 0[,]0, 350[\}$ of Ξ_1, and $S_{p_2}^{\xi_2}$ is the space

of polynomials of degree 5 on Ξ_2. This choice results in $P = 72$. The Galerkin approximation $u \in V_N \otimes S_P \subset V \otimes S$ of the solution of (4.58) is defined by the following equation[3]:

$$\int_\Xi \int_\Omega (\nabla u \cdot \nabla v + c \cdot \nabla u v + a u v)\, dx\, dP_\xi = \int_\Xi \int_\Omega f\, v\, dx\, dP_\xi, \tag{4.59}$$

for all $v \in V_N \otimes S_P$. Letting $V_N \otimes S_P = span\{\varphi_i \otimes \psi_j; 1 \leq i \leq N, 1 \leq j \leq P\}$, the Galerkin approximation $u = \sum_{i=1}^N \sum_{j=1}^P u_{ij} \varphi_i \otimes \psi_j$ can be identified with its set of coefficients on the chosen basis, still denoted u, which is a tensor

$$u \in X = \mathbb{R}^N \otimes \mathbb{R}^P \quad \text{such that} \quad Au = b, \tag{4.60}$$

where $b = b^x \otimes b^\xi$, with $b_i^x = \int_\Omega f \varphi_i$ and $b_j^\xi = \int_\Xi \psi_j dP_\xi$, and where A is a rank-3 operator such that $A = D^x \otimes M^\xi + C^x \otimes H^{\xi_1} + R^x \otimes H^{\xi_2}$, with $D_{ik}^x = \int_\Omega \nabla \varphi_i \cdot \nabla \varphi_k dx$, $C_{ik}^x = \int_\Omega \varphi_i c_0 \cdot \nabla \varphi_k dx$, $R_{ik}^x = \int_\Omega \varphi_i \varphi_k dx$, $M_{jl}^\xi = \int_\Xi \psi_j(y) \psi_l(y) dP_\xi(y)$, $H_{jl}^{\xi_n} = \int_\Xi y_n \psi_j(y) \psi_l(y) dP_\xi(y)$, $n = 1, 2$. Here, we use orthonormal basis functions $\{\psi_j\}$ in S_P, so that $M^\xi = I_P$, the identity matrix in $\mathbb{R}^P$.

8.2 Comparison of minimal residual methods

In this section, we present numerical results concerning the approximate ideal minimal residual method (A-IMR) applied to the algebraic system of equations (4.60) in tensor format. This method provides an approximation of the best approximation of u with respect to a norm $\|\cdot\|_X$ that can be freely chosen a priori. Here, we consider the application of the method for two different norms. We first consider the natural canonical norm on X, denoted $\|\cdot\|_2$ and defined by

$$\|v\|_2^2 = \sum_{i=1}^N \sum_{j=1}^P (v_{ij})^2. \tag{4.61}$$

This choice corresponds to an operator $R_X = I_X = I_N \otimes I_P$, where I_N (resp. I_P) is the identity in $\mathbb{R}^N$ (resp. $\mathbb{R}^P$). We also consider a weighted canonical norm, denoted $\|\cdot\|_w$ and defined by

$$\|v\|_w^2 = \sum_{i=1}^N \sum_{j=1}^P (w(x_i) v_{ij})^2, \tag{4.62}$$

where $w : \Omega \to \mathbb{R}$ is a weight function and the x_i are the nodes associated with finite element shape functions φ_i. This norm allows to give a more important weight to a

[3]The mesh Péclet number is sufficiently small so that an accurate Galerkin approximation can be obtained without introducing a stabilized formulation.

particular region $D \subset \Omega$, that may be relevant if one is interested in the prediction of a quantity of interest that requires a good precision of the numerical solution in this particular region (see section 8.2.3). This choice corresponds to an operator $R_X = D_w \otimes I_P$, with $D_w = \mathrm{diag}(w(x_1)^2, \cdots, w(x_N)^2)$.

The A-IMR provides an approximation $\widetilde{u} \in \mathcal{M}_r(X)$ of the $\|\cdot\|_X$-best approximation of u in $\mathcal{M}_r(X)$ (that means an approximation of an element in $\Pi_{\mathcal{M}_r(X)}(u)$), where $\|\cdot\|_X$ is either $\|\cdot\|_2$ or $\|\cdot\|_w$. The set $\mathcal{M}_r(X)$ is taken as the set $\mathcal{C}_r(X)$ of rank-r tensors in $X = \mathbb{R}^N \otimes \mathbb{R}^P$. The dimension of X is about 75,000 so that the exact solution u of (4.60) can be computed and used as a reference solution. We note that both norms are induced norms in $\mathbb{R}^N \otimes \mathbb{R}^P$ (associated with rank one operators R_X) so that the $\|\cdot\|_X$-best approximation of u in $\mathcal{M}_r(X)$ is a rank-r SVD that can be computed exactly using classical algorithms (see section 6.1).[4] For the construction of an approximation in $\mathcal{C}_r(X)$ using A-IMR, we consider two strategies: the direct approximation in $\mathcal{C}_r(X)$ using Algorithm 7 with $\mathcal{M}_r(X) = \mathcal{C}_r(X)$, and a greedy algorithm that consists in a series of r corrections in $\mathcal{C}_1(X)$ computed using Algorithm 7 with $\mathcal{M}_r(X) = \mathcal{C}_1(X)$ and with an updated residual b at each correction.

The A-IMR will be compared to a standard approach, denoted CMR, which consists in minimizing the canonical norm of the residual of equation (4.60), that means in solving

$$\min_{v \in \mathcal{M}_r(X)} \|Av - b\|_2. \tag{4.63}$$

This latter approach has been introduced and analyzed in different papers, using either direct minimization or greedy rank-one algorithms [2, 16, 51], and is known to suffer from ill-conditioning of the operator A. We note that this approach corresponds to choosing $R_X = A^*A$ and $R_Y = I_X = I_N \otimes I_P$.

8.2.1 Natural canonical norm $\|\cdot\|_2$

First, we compare both greedy and direct algorithms for $\|\cdot\|_X = \|\cdot\|_2$, using either CMR or A-IMR with different precisions δ. The convergence curves with respect to the rank are shown in Figure 4.2, where the error is measured in the $\|\cdot\|_2$ norm. Concerning the direct approach, we observe that the different algorithms have roughly the same rate of convergence. The A-IMR convergence curves are close to the optimal SVD (corresponding to $\widetilde{u}_2$) for a wide range of values of δ. One should note that A-IMR seems to provide good approximations also for the value

[4] Note that different truncated SVD are obtained when $\mathbb{R}^N$ is equipped with different norms.

$\delta = 0.9$ which is greater than the theoretical bound 0.5 ensuring the convergence of the gradient-type algorithm. Concerning the greedy approach, we observe a significant difference between A-IMR and CMR. We note that A-IMR is close to the optimal SVD up to a certain rank (depending on δ) after which the convergence rate decreases but remains better than the one of CMR. Finally, one should note that using a precision $\delta = 0.9$ for A-IMR yields less accurate approximations than CMR. However, A-IMR provides better results than CMR once the precision δ is lower than 0.5.

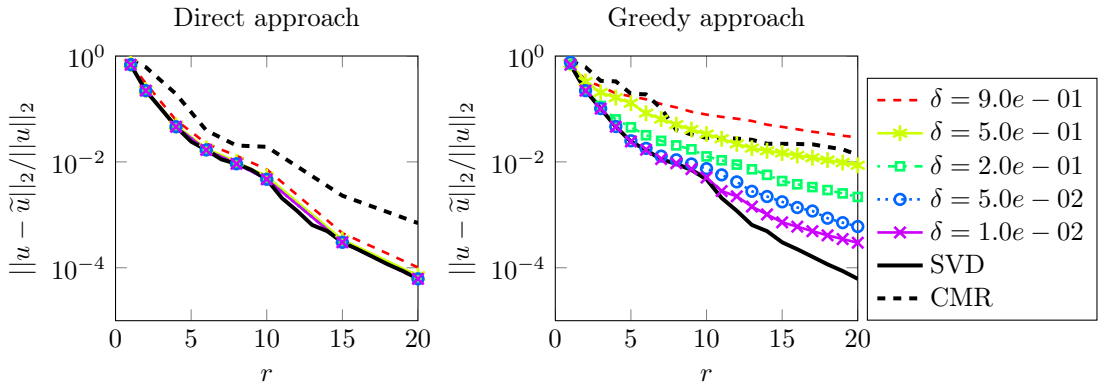


Figure 4.2: Comparison of minimal residual methods for $\mathcal{M}_r(X) = \mathcal{C}_r(X)$ and $\|\cdot\|_X = \|\cdot\|_2$. Convergence with the rank r of the approximations obtained with CMR or A-IMR with different precisions δ, and with direct (left) or greedy rank-one (right) approaches.

8.2.2 Weighted norm $\|\cdot\|_w$

Here, we perform the same numerical experiments as previously using the weighted norm $\|\cdot\|_X = \|\cdot\|_w$, with w equal to 10^3 on $D = [0.15, 0.25] \times [0.45, 0.55]$ and $w = 1$ on $\Omega \setminus D$. The convergence curves with respect to the rank are plotted on Figure 4.3. The conclusions are similar to the case $\|\cdot\|_X = \|\cdot\|_2$, although the use of the weighted norm seems to slightly deteriorate the convergence properties of A-IMR. However, the direct A-IMR still provides better approximations than the direct CMR, closer to the reference SVD (denoted by $\tilde{u}_w$) for different values of precision δ.

8.2.3 Interest of using a weighted norm

Here, we illustrate the interest of using the weighted norm rather than the natural canonical norm when one is interested in computing a quantity of interest. For the

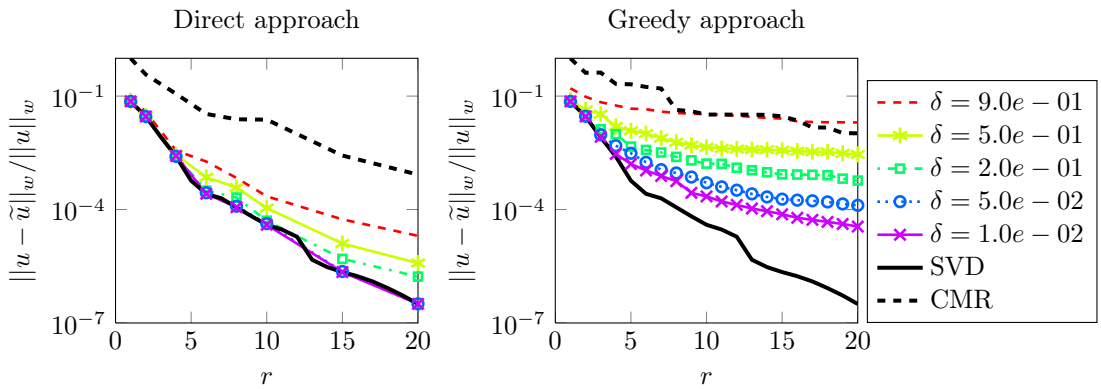


Figure 4.3: Comparison of minimal residual methods for $\mathcal{M}_r(X) = \mathcal{C}_r(X)$ and $\|\cdot\|_X = \|\cdot\|_w$. Convergence with the rank of the approximations obtained with CMR or A-IMR with different precisions δ, and with direct (left) or greedy rank-one (right) approaches.

sake of readability, we let $\widetilde{u}_w$ (resp. $\widetilde{u}_2$) denote the best approximation of u in $\mathcal{C}_r(X)$ with respect to the norm $\|\cdot\|_w$ (resp. $\|\cdot\|_2$). Figure 4.4 illustrates the convergence with r of these approximations. We observe that approximations $\widetilde{u}_w$ and $\widetilde{u}_2$ are of the same quality when the error is measured with the norm $\|\cdot\|_2$, while $\widetilde{u}_w$ is a far better approximation than $\widetilde{u}_2$ (almost two orders of magnitude) when the error is measured with the norm $\|\cdot\|_w$. We observe that $\widetilde{u}_w$ converges faster to u with $\|\cdot\|_w$ than $\widetilde{u}_2$ with $\|\cdot\|_2$. For example, with a rank $r = 9$, $\widetilde{u}_w$ has a $\|\cdot\|_w$-error of 10^4 while $\widetilde{u}_2$ has a $\|\cdot\|_2$-error of 10^2. On Figure 4.5, plotted are the spatial modes of the rank-r approximations $\widetilde{u}_2$ and $\widetilde{u}_w$. These spatial modes are significantly different and obviously capture different features of the solution.

Now, we introduce a quantity of interest Q which is the spatial average of u on subdomain D:

$$Q(u) = \frac{1}{|D|} \int_D u \, \mathrm{d}x. \tag{4.64}$$

Due to the choice of norm, $\widetilde{u}_w$ is supposed to be more accurate than $\widetilde{u}_2$ in the subdomain D, and therefore, $Q(\widetilde{u}_w)$ is supposed to provide a better estimation of $Q(u)$ than $Q(\widetilde{u}_2)$. This is confirmed by Figure 4.6, where we have plotted the convergence with the rank of the statistical mean and variance of $Q(\widetilde{u}_w)$ and $Q(\widetilde{u}_2)$. With only a rank $r = 5$, $\widetilde{u}_w$ gives a precision of 10^{-7} on the mean, whereas $\widetilde{u}_2$ gives only a precision of 10^{-2}. In conclusion, we observe that a approximation $\widetilde{u}_w$ with a very low rank is able to provide a very good approximation of the quantity of interest.

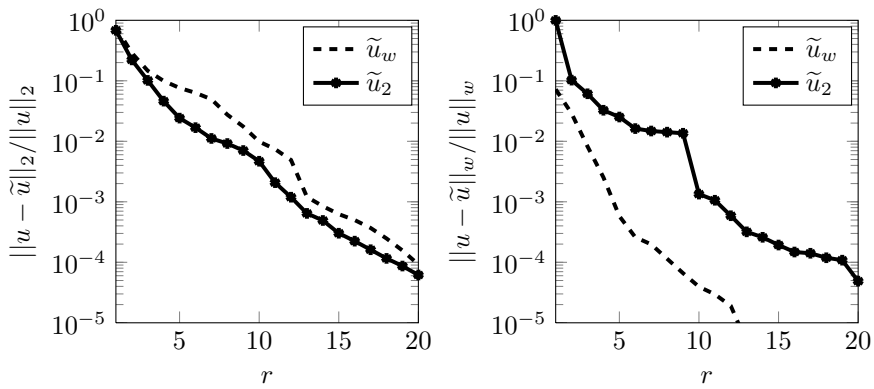


Figure 4.4: Convergence of best rank-r approximations $\widetilde{u}_2$ and $\widetilde{u}_w$ of the solution u measured with the natural canonical norm $\|\cdot\|_2$ or the weighted norm $\|\cdot\|_w$.

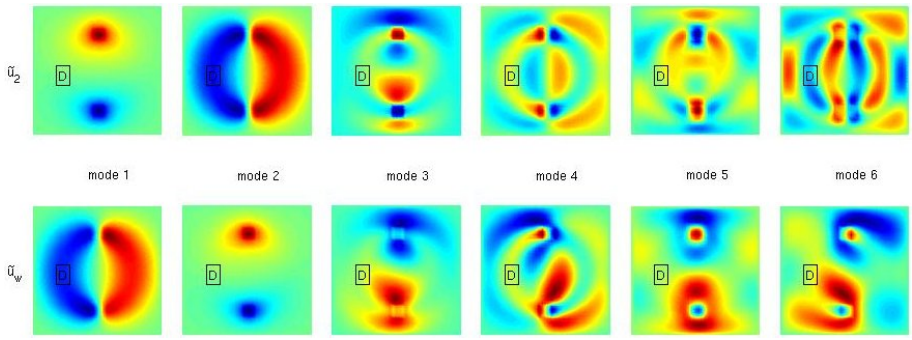
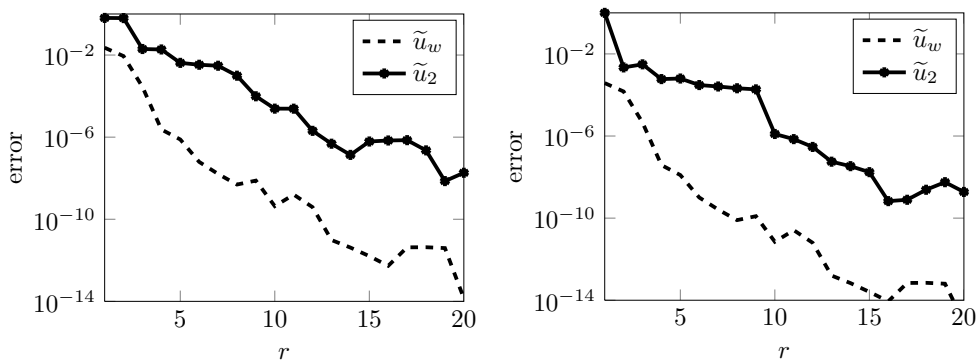


Figure 4.5: Comparison of the first spatial modes of the rank-r approximations $\widetilde{u}$ and $\widetilde{u}_w$.



(a) Error on the mean value of $Q(\widetilde{u})$ (b) Error on the variance of $Q(\widetilde{u})$

Figure 4.6: Convergence with the rank of the mean (left) and variance (right) of $Q(\widetilde{u}_2)$ and $Q(\widetilde{u}_w)$. Relative error with respect to the mean and variance of the reference solution $Q(u)$.

8.3 Properties of the algorithms

Now, we detail some numerical aspects of the proposed methodology. We first focus on the gradient-type algorithm, and then on evaluations of the map Λ^δ for the approximation of residuals.

8.3.1 Analysis of the gradient-type algorithm

The behavior of the gradient-type algorithm for different choices of norms $\|\cdot\|_X$ is very similar, so we only illustrate the case where $\|\cdot\|_X = \|\cdot\|_2$. The convergence of this algorithm is plotted in Figure 4.7 for the case $\mathcal{M}_r(X) = \mathcal{C}_{10}(X)$. It is in very good agreement with theoretical expectations (Proposition 5.3): we first observe a linear convergence with a convergence rate that depends on δ, and then a stagnation within a neighborhood of the solution with an error depending on δ.

The gradient-type algorithm is then applied for subsets $\mathcal{M}_r(X) = \mathcal{C}_r(X)$ with different ranks r. The estimate of the linear convergence rate ρ is given in Table 4.1. We observe that for all values of r, ρ takes values closer to δ than to the theoretical bound 2δ of Proposition 5.3. This means that the theoretical bound of the convergence rate overestimates the effective one, and the algorithm converges faster than expected.

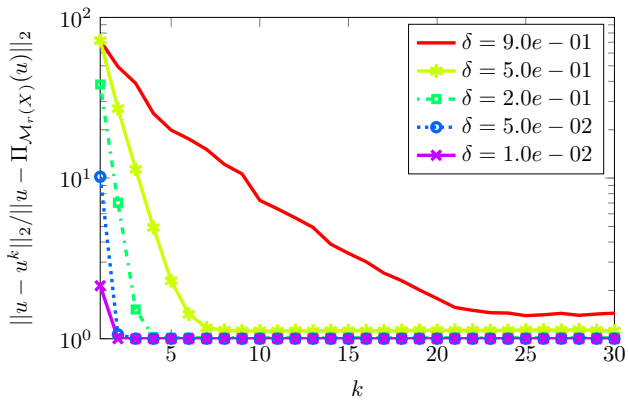


Figure 4.7: Convergence of the gradient-type algorithm for different values of the relative precision δ, for $\mathcal{M}_r(X) = \mathcal{C}_{10}(X)$ and $\|\cdot\|_X = \|\cdot\|_2$.

Now, in order to evaluate the quality of the resulting approximation, we compute the error after the stagnation phase has been reached. More precisely, we compute

δ	0.90	0.50	0.20	0.05	0.01
$r = 4$	0.78	0.36	≈ 0	≈ 0	≈ 0
$r = 6$	0.83	0.45	0.165	≈ 0	≈ 0
$r = 10$	0.82	0.42	0.183	≈ 0	≈ 0
$r = 15$	0.84	0.47	0.189	0.047	≈ 0
$r = 20$	0.86	0.48	0.197	0.051	0.011

Table 4.1: Estimation of the convergence rate ρ of the gradient-type algorithm (during the linear convergence phase) for different subsets $\mathcal{M}_r(X) = \mathcal{C}_r(X)$, and for $\|\cdot\|_X = \|\cdot\|_2$.

the value

$$\widetilde{\gamma}_k = \frac{\|u^k - u\|_X}{\|u - \Pi_{\mathcal{M}_r(X)}(u)\|_X} - 1,$$

for $k = 100$. Values of $\widetilde{\gamma}_{100}$ are summarized in Table 4.2 and are compared to the theoretical upper bound $\gamma = 2\delta/(1-2\delta)$ given by Proposition 5.3. Once again, one can observe that the effective error of the resulting approximation is lower than the predicted value regardless of the choice of $\mathcal{C}_r(X)$.

δ	0.90	0.50	0.20	0.05	0.01
$2\delta/(1-2\delta)$	-	-	6.6e-1	1.1e-1	2.1e-2
$r = 4$	3.3e-1	5.6e-2	4.9e-3	3.5e-4	3.0e-5
$r = 6$	3.0e-1	6.8e-2	1.1e-2	8.6e-4	8.0e-5
$r = 10$	5.2e-1	1.3e-1	1.7e-2	1.8e-3	3.3e-5
$r = 15$	4.9e-1	1.1e-1	1.5e-2	1.0e-3	7.5e-5
$r = 20$	6.4e-1	1.5e-1	1.9e-2	1.2e-3	7.3e-5

Table 4.2: Final approximation errors (estimated by $\widetilde{\gamma}_{100}$) for different subsets $\mathcal{M}_r(X) = \mathcal{C}_r(X)$ and different precisions δ. Comparison with the theoretical upper bound $2\delta/(1-2\delta)$.

Now, we focus on numerical estimations of the error $\|u - u^k\|_X$. It has been pointed out in Section 5.4 that $\hat{\epsilon}^k$, defined in Eq. (4.39), should provide a good error estimator with effectivity index $\hat{\tau}^k \in (1, (1-\delta^2)^{-1/2})$. For $\delta = 0.2$ and $\mathcal{M}_r(X) = \mathcal{C}_{10}(X)$, numerical values taken by $\hat{\tau}^k$ during the gradient-type algorithm are plotted on Figure 4.8 and are compared to the expected theoretical values of its lower and upper bounds 1 and $(1-\delta^2)^{-1/2}$ respectively. We observe that the theoretical upper bound is strictly satisfied, while the lower bound is almost but not exactly

satisfied. This violation of the theoretical lower bound is explained by the fact that the precision δ is not satisfied at each iteration of the gradient-type algorithm due to the use of a heuristic convergence criterion in the computation of residuals (see next section 8.3.2). However, although it does not provide a controlled error estimation, the error indicator based on the computed residuals is of very good quality.

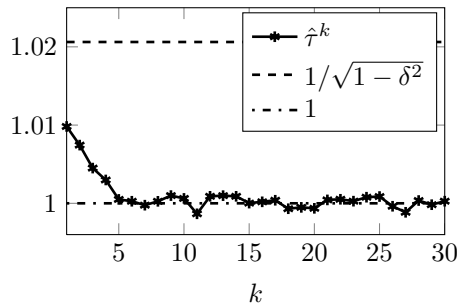


Figure 4.8: Effectivity index $\hat{\tau}^k$ of the error estimator $\hat{\epsilon}^k$ at different iterations k of the gradient-type algorithm, with $\mathcal{M}_r(X) = \mathcal{C}_{10}(X)$ and $\delta = 0.2$.

8.3.2 Application of Λ^δ for the approximation of residuals

We study the behavior of the updated greedy algorithm described in Section 6.2.2 for the computation of an approximation $y_m^k = \Lambda^\delta(r^k)$ of the residual r^k during the gradient-type algorithm. Here, we use the particular strategy which consists in updating functions associated to each dimension $\mu \in I = \{1, 2\}$ (steps (2)-(3) are performed two times per iteration). We first validate the ability of the heuristic stopping criterion (4.44) to ensure a prescribed relative precision δ. Let $M = M(\delta)$ denote the iteration for which the condition $e_M^p \leq \delta$ is satisfied. The exact relative error $e_M = \|y_M^k - r^k\|_Y / \|r^k\|_Y$ is computed using a reference computation of r^k, and we define the effectivity index $\lambda_M^p = e_M^p / e_M$. Figure 4.9 shows the convergence of this effectivity index with respect to p, when using the natural canonical norm $\|\cdot\|_2$ or the weighted norm $\|\cdot\|_w$. We observe that λ_M^p tends to 1 as $p \to \infty$, as it was expected since the sequence $\{y_m^k\}_{m\geq 1}$ converges to r^k. However, we clearly observe that the quality of the error indicator differs for the two different norms. When using the weighted norm, it appears that a large value of p (say $p \geq 20$) is necessary to ensure $\lambda_M^p \in [0.9, 1]$, while $p \leq 10$ seems sufficiently large when using the natural canonical norm. That simply reflects a slower convergence of the greedy algorithm when using the weighted norm.

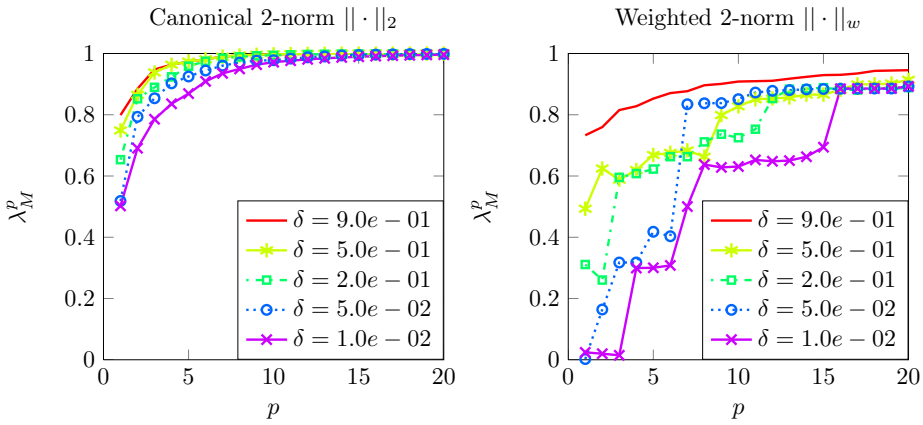


Figure 4.9: Evolution with p of the effectivity index λ_M^p for different δ at step $k = 1$ of the gradient-type algorithm with $\mathcal{M}_r(X) = \mathcal{C}_{10}(X)$ and for the natural canonical norm (left) or the weighted norm (right).

Remark 8.1. *One can prove that at step k of the gradient-type algorithm, when computing an approximation y_M^k of r^k with a greedy algorithm stopped using the heuristic stopping criterion (4.44), the effectivity index $\hat{\tau}^k$ of the computed error estimator $\hat{\epsilon}^k$ is such that*

$$\hat{\tau}^k \in \left(\sqrt{\frac{1-(\delta/\lambda_M^p)^2}{1-\delta^2}}, \sqrt{\frac{1}{1-\delta^2}} \right).$$

where λ_M^p is the effectivity index of error indicator e_M^p (supposed such that $\delta/\lambda_M^p < 1$). That provides an explanation for the observations made on Figure 4.8.

Now, we observe in Table 4.3 the number of iterations of the greedy algorithm for the approximation of the residual r^k with a relative precision δ, with a fixed value $p = 20$ for the evaluation of the stopping criterion. The number of iterations corresponds to the rank of the resulting approximation. We note that the required rank is higher when using the weighted norm. It reflects the fact that it is more difficult to reach precision δ when using the weighted norm rather than the natural canonical norm.

8.4 Higher dimensional case

Now, we consider a diffusion coefficient of the form $\kappa(x,\xi) = \kappa_0 + \sum_{i=1}^{8} \xi_i \kappa_i(x)$ where $\kappa_0 = 10$, $\xi_i \sim U(-1,1)$ are independent uniform random variables, and the functions

	Canonical 2-norm $\|\cdot\|_2$					Weighted 2-norm $\|\cdot\|_w$				
k\δ	0.9	0.5	0.2	0.05	0.01	0.9	0.5	0.2	0.05	0.01
1	1	1	3	7	11	8	21	31	35	51
2	1	3	7	16	27	5	22	14	24	42
3	1	5	11	19	24	4	15	24	23	43
4	1	3	11	14	24	8	11	19	37	42
5	1	6	7	15	24	6	19	23	14	38
6	1	8	8	16	24	3	12	47	25	63
7	1	5	7	17	24	7	14	16	29	47
8	1	4	8	16	24	5	12	22	21	40
9	1	4	8	16	24	7	13	18	36	45

Table 4.3: Computation of $\Lambda^\delta(r^k)$ for different precisions δ and at different steps k of the gradient-type algorithm, with $\mathcal{M}_r(X) = \mathcal{C}_{10}$ (direct approach). The table indicates the number of greedy corrections computed for reaching the precision δ using the heuristic stopping criterion (4.44) with $p = 20$.

$\kappa_i(x)$ are given by:

$$\begin{aligned}
\kappa_1(x) &= \cos(\pi x_1), & \kappa_5(x) &= \cos(\pi x_1)\cos(\pi x_2),\\
\kappa_2(x) &= \cos(\pi x_2), & \kappa_6(x) &= \sin(\pi x_1)\sin(\pi x_2),\\
\kappa_3(x) &= \sin(\pi x_1), & \kappa_7(x) &= \cos(\pi x_1)\sin(\pi x_2),\\
\kappa_4(x) &= \sin(\pi x_2), & \kappa_8(x) &= \sin(\pi x_1)\cos(\pi x_2)
\end{aligned}$$

In addition, the advection coefficient is given by $c = \xi_0 c_0$, where $\xi_0 \sim U(0, 4000)$ is a uniform random variable. We denote $V = \mathrm{H}^1_0(\Omega)$ and $S = \mathrm{L}^2(\Xi, dP_\xi)$ where $(\Xi, \mathcal{B}(\Xi), P_\xi)$ is a probability space with $\Xi =]-1, 1[^8 \times]0, 4000[$ and P_ξ the uniform measure. Here $V_N \subset V$ is a $\mathbb{Q}_1$ finite element space associated with a uniform mesh of 3600 elements, with a dimension $N = 3481$. We take $S_P = \otimes_{i=0}^8 S_P^{\xi_i} \subset S$, where $S_P^{\xi_i}$ are polynomial function spaces of degree 7 on Ξ_i with $P = \dim(S_P^{\xi_i}) = 8$. Then, the Galerkin approximation in $V_N \otimes S_P$ (solution of (4.59)) is searched under the form $u = \sum_{i=1}^N \sum_{j_0=1}^P \cdots \sum_{j_8=1}^P (u_{i,j_0,\cdots,j_9})\phi_j \otimes (\otimes_{\mu=0}^8 \psi_{j_\mu}^\mu)$. This Galerkin approximation can be identified with its set of coefficients, still denoted by u which is a tensor

$$u \in X = \mathbb{R}^N \otimes (\otimes_{\mu=0}^8 \mathbb{R}^P) \quad \text{such that} \quad Au = b, \tag{4.65}$$

where A and b are the algebraic representations on the chosen basis of $V_N \otimes S_P$ of the bilinear and linear forms in (4.59). The obtained algebraic system of equations has a dimension larger than 10^{11} and its solution clearly requires the use of model reduction methods.

Here, we compute low-rank approximations of the solution of (4.65) in the canonical tensor subset $\mathcal{C}_r(X)$ with $r \geq 1$. Since best approximation problems in $\mathcal{C}_r(X)$ are well posed for $r = 1$ but ill posed for $d > 2$ and $r > 1$, we rely on the greedy algorithm presented in section 7 with successive corrections in $\mathcal{M}_r(X) = \mathcal{C}_1(X)$ computed with Algorithm 7.

Remark 8.2. *Low-rank approximations could have been computed directly with Algorithm 7 by choosing for $\mathcal{M}_r(X)$ other stable low-rank formats adapted to high-dimensional problems, such as Hierarchical Tucker (or Tensor Train) low-rank formats.*

8.4.1 Convergence study

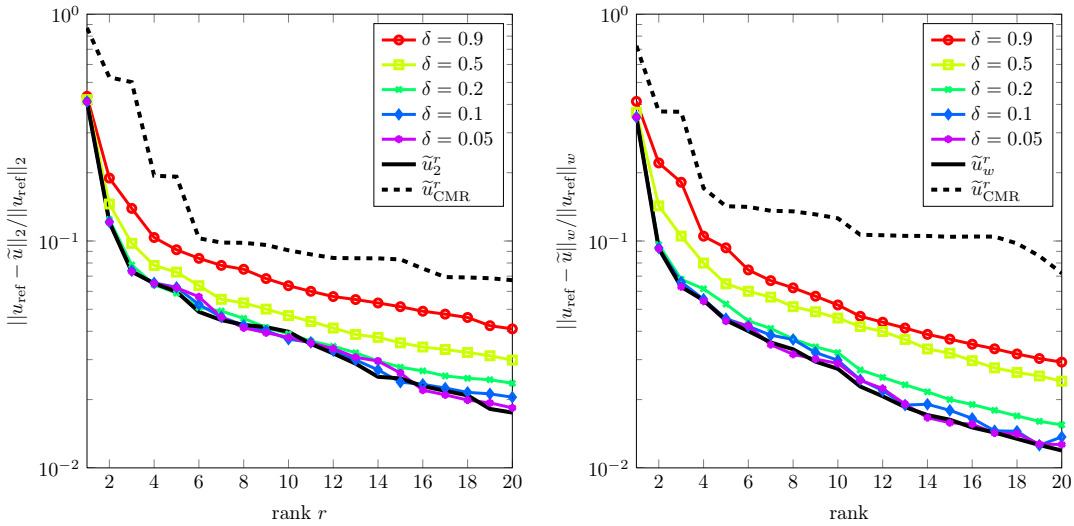
In this section, low-rank approximations of the solution u of (4.65) are computed for the two different norms $\|\cdot\|_2$ and $\|\cdot\|_w$ defined as in section 8.2. Here, we assume that the weighting function w is equal to 100 in the subdomain $D \subset \Omega$, and 1 elsewhere.

Since $\dim(X) \geq 10^{11}$, the exact Galerkin approximation u in X is no more computable. As a reference solution, we consider a low-rank approximation u_{ref} of u computed using a greedy rank-one algorithm based on a canonical minimal residual formulation. We introduce an estimation $\hat{E}_K$ of $\frac{\|u-u_{\text{ref}}\|_2}{\|u\|_2}$ based on Monte-Carlo integrations using K realizations $\{\xi_k\}_{k=1}^K$ of the random variable ξ, defined by

$$\hat{E}_K^2 = \frac{\frac{1}{K}\sum_{k=1}^K \|u(\xi_k) - u_{\text{ref}}(\xi_k)\|_V^2}{\frac{1}{K}\sum_{k=1}^K \|u(\xi_k)\|_V^2},$$

with a number of samples K such that the Monte-Carlo estimates has a relative standard deviation (estimated using the statistical variance of the sample) lower than 10^{-1}. The rank of u_{ref} is here selected such that $\hat{E}_K < 10^{-4}$, which gives a reference solution with a rank of 212.

On Figure 4.10, we plot the convergence with the rank r of the approximations computed by both A-IMR and CMR algorithms and of the greedy approximations $\widetilde{u}_2^r$ and $\widetilde{u}_w^r$ of the reference solution u_{ref} for both norms. We observe (as for the lower-dimensional example) that for both norms, with different values of the parameter δ (up to 0.9), the A-IMR method provides a better approximation of the solution in comparison to the CMR method. When decreasing δ, the proposed algorithm seems to provide approximations that tend to present the same convergence as the greedy approximations $\widetilde{u}_2^r$ and $\widetilde{u}_w^r$.



(a) Canonical norm (b) Weighted norm

Figure 4.10: Convergence with the rank of approximations obtained with the greedy CMR or A-IMR algorithms for different precisions δ. On the left (resp. right) plot, convergence is plotted with respect to the norm $\|\cdot\|_2$ (resp. $\|\cdot\|_w$) and A-IMR is used with the objective norm $\|\cdot\|_2$ (resp. $\|\cdot\|_w$).

8.4.2 Study of the greedy algorithm for Λ^δ

Now, we study the behavior of the updated greedy algorithm described in Section 6.2.2 for the computation of an approximation $y_m^k = \Lambda^\delta(r^k)$ of the residual r^k during the gradient-type algorithm. Here, we use the particular strategy which consists in updating functions associated to each dimension $\mu \in I = \{2, \cdots, 10\}$ (steps (2)-(3) are performed 9 times per iteration). The update of functions associated with the first dimension is not performed since it would involve the expensive computation of approximations in a space Z_m^k with a large dimension mN.

In table 4.4, we summarize the required number of greedy corrections needed at each iteration of the gradient type algorithm for reaching the precision δ with the heuristic stagnation criterion (4.44) with $p = 20$. As for the previous lower-dimensional test case, the number of corrections increases as δ decreases and is higher for the weighted norm than for the canonical norm. However, we observe that this number of corrections remains reasonable even for small δ.

	Canonical 2-norm $\|\cdot\|_2$					Weighted 2-norm $\|\cdot\|_w$				
k\δ	0.9	0.5	0.2	0.05	0.01	0.9	0.5	0.2	0.05	0.01
1	1	1	3	6	14	3	12	53	65	91
2	1	3	5	13	24	2	11	49	62	91
3	1	3	5	12	17	3	12	49	62	91
4	1	3	5	13	26	3	12	53	62	91
5	1	3	6	12	24	2	11	47	65	89
6	1	3	5	13	27	3	11	42	63	88
7	1	3	5	12	27	3	10	50	65	88
8	1	3	5	12	26	3	10	49	60	87
9	1	3	6	12	26	3	13	49	65	80

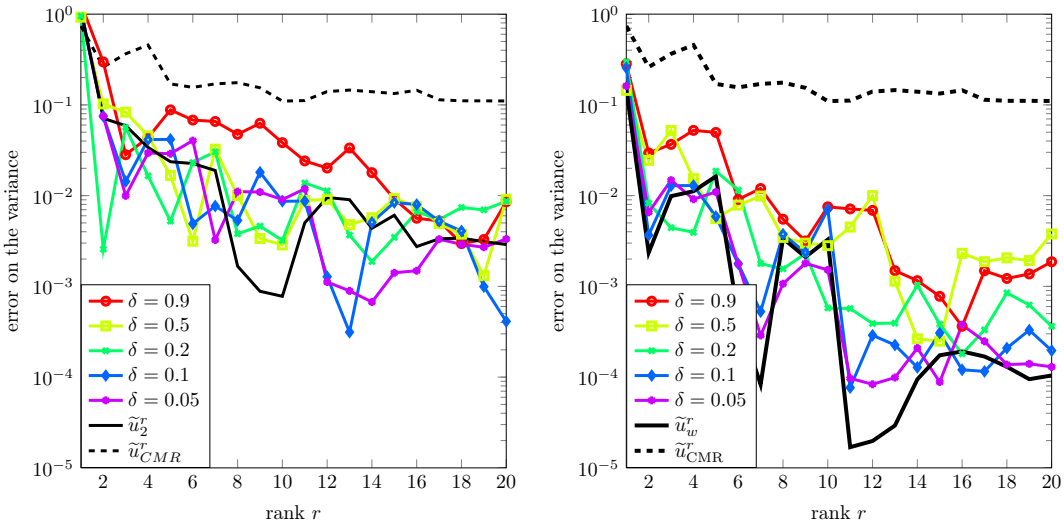
Table 4.4: Computation of $\Lambda^\delta(r^k)$ for different precisions δ and at different steps k of the gradient-type algorithm (first iteration $r = 1$ of the greedy approach with $\mathcal{M}_r(X) = \mathcal{C}_1$). The table indicates the number of greedy corrections computed for reaching the precision δ using the heuristic stopping criterion (4.44) with $p = 20$.

8.4.3 Estimation of a quantity of interest

Finally, we study the quality of the low-rank approximations $\widetilde{u}$ obtained with both CMR and A-IMR algorithms for the canonical and weighted norms. To this end, we compute the quantity of interest $Q(\widetilde{u})$ defined by (4.64). Figure 4.11 illustrates the convergence with the rank of the variance of the approximate quantities of interest. Note that the algorithm does not guarantee the monotone convergence of the quantity of interest with respect to the rank, that is confirmed by the numerical results. However, we observe that the approximations provided by the A-IMR algorithm are better than the ones given by the CMR, even for large δ. Also, when using the weighted norm in the A-IMR algorithm, the quantity of interest is estimated with a better precision. Similar behaviors are observed for the convergence of the mean.

9 Conclusion

In this chapter, we have proposed a new algorithm for the construction of low-rank approximations of the solutions of high-dimensional weakly coercive problems formulated in a tensor space X. This algorithm is based on the approximate minimization (with a certain precision δ) of a particular residual norm on given low-rank tensor subsets $\mathcal{M}_r(X)$, the residual norm coinciding with some measure of the error in solution. Therefore, the algorithm is able to provide a quasi-best low-rank approximation with respect to a norm $\|\cdot\|_X$ that can be designed for a certain



(a) Using the canonical norm.

(b) Using the weighted norm.

Figure 4.11: Relative error with respect to the variance of the reference solution $Q(u_{\mathrm{ref}})$ with the canonical (left) and weighted (right) norms.

objective. A weak greedy algorithm using this minimal residual approach has been introduced and its convergence has been proved under some conditions. A numerical example dealing with the solution of a stochastic partial differential equation has illustrated the effectivity of the method and the properties of the proposed algorithms. Some technical points have to be addressed in order to apply the method to a more general setting and to improve its efficiency and robustness: the development of efficient solution methods for the computation of residuals when using general norms $\|\cdot\|_X$ (that are not induced norms in the tensor space X), the introduction of robust error estimators during the computation of residuals (for the robust control of the precision δ, which is the key point for controlling the quality of the obtained approximations), the application of the method for using tensor formats adapted to high-dimensional problems (such as Hierarchical formats). Also, a challenging perspective consists in coupling low-rank approximation techniques with adaptive approximations in infinite-dimensional tensor spaces (as in [7]) in order to provide approximations of high-dimensional equations (PDEs or stochastic PDEs) with a complete control on the precision of quantities of interest.

Chapter 5

Goal-oriented low-rank approximation for the estimation of vector-valued quantities of interest

In this chapter, we address the problem of the goal-oriented low-rank approximation of the solution of a tensor-structured equation. We introduce a goal-oriented *norm that takes into account the error associated to some functional-valued (or vector-valued) quantity of interest we want to estimate. The advantage of controlling the approximation with this norm is that it can reduce the rank of the approximation while keeping an accurate estimation of the quantity of interest. However, the computation of such goal-oriented approximations requires dedicated algorithms. We first propose an iterative solver with low-rank truncations of the iterates, where the truncations are controlled with respect to the goal-oriented norm. We then propose an ideal minimal residual approach which is similar to the one introduced in Chapter 4. Finally, applications in uncertainty quantification are considered, where the quantity of interest is defined as the expectation or the conditional expectation of some variable of interest.*

Contents

1 Introduction . . . 153

2 Choice of norms . . . 154

2.1 Natural norms . . . 154

2.2 Goal-oriented norm . . . 156

3 Algorithms for goal-oriented low-rank approximations . . . 157

3.1 Iterative solver with goal-oriented truncations . . . 159

3.2 A method based on an ideal minimal residual formulation . . . 160

4 Application to uncertainty quantification . . . 162

4.1 Linear quantities of interest . . . 163

4.2 Properties of the norms . . . 164

4.3 Approximation of $u(\xi)$ by interpolation . . . 166

5 Numerical experiments . . . 168

5.1 Iterative solver (PCG) with truncations . . . 168

5.2 Ideal minimal residual formulation . . . 172

6 Conclusion . . . 177

7 Appendix: practical implementation of the approximation operators . . . 178

1 Introduction

We consider a linear equation formulated on a tensor product space:

$$Au = b, \quad u \in X = X_1 \otimes \ldots \otimes X_d. \tag{5.1}$$

When using low-rank tensor approximation techniques, we look for an approximation of the solution u in a subset of low-rank tensors $\mathcal{M}_r(X) \subset X$, such as a subset of Tucker or Hierarchical Tucker tensors. However, obtaining accurate low-rank approximations require in practice a huge computational effort. Besides, in many situations, one is not interested in the solution u itself, but rather in some partial information on u, called a quantity of interest. This quantity of interest can take the form

$$s = Lu,$$

where $L \in \mathcal{L}(X, Z)$ is a continuous linear application from X to some Hilbert space Z. *Primal-dual* approaches can be applied, where an approximation of s is obtained by solving a *dual problem*. This approach is used when the quantity of interest s is a scalar, which corresponds to $L \in X'$ or equivalently to $Z = \mathbb{R}$. The extension of this strategy to vector-valued or functional-valued quantities of interest is possible. In that case, the dual variable is an operator and its approximation requires appropriate techniques.

The main motivation of this chapter is to propose strategies for the goal-oriented low-rank approximation of u. The idea is the following. Since only a partial information is needed, it should be possible to build an approximation with a very small rank which still provides a good estimation of the quantity of interest. For this purpose, we propose to define a *goal-oriented* norm $\|\cdot\|_{X_\alpha}$ over the space X which is such that

$$\|v\|_{X_\alpha}^2 = \|v\|_X^2 + \alpha\|Lv\|_Z^2 \tag{5.2}$$

for all $v \in X$, where $\|\cdot\|_X$ and $\|\cdot\|_Z$ are the natural norms in the spaces X and Z respectively, and α is a positive scalar. The idea is that, for a sufficiently large value of α, an approximation of u with respect to the goal-oriented norm will provide a controlled approximation of $s = Lu$. The aim of this chapter is to propose algorithms for the *a priori* construction of such goal-oriented low-rank approximations.

The outline of this chapter is as follows. Section 2 introduces the natural norms over the spaces X and Z, and demonstrates the expected benefits of using the goal-oriented norm. In Section 3, we propose two strategies for the *a priori* construction

of a goal-oriented approximation. The first one consists in using iterative solvers with goal-oriented low-rank truncations of the iterates. The second one relies on an ideal minimal residual formulation. Examples of applications to parameter-dependent equations are presented in Section 4. In particular, we emphasize on the computation of expectations, or of conditional expectations, which are classical linear quantities of interest in uncertainty quantification. Finally, numerical results are presented in Section 5.

2 Choice of norms

In this section we define what we call the *natural norms* for the spaces X and Z. Then, we give some properties of the goal-oriented norm defined by (5.2) and we show the interest of using such a norm with a toy example.

2.1 Natural norms

When considering a tensor Hilbert space $X = X_1 \otimes \ldots \otimes X_d$, the natural norm $\|\cdot\|_X$ is the crossnorm, called the canonical norm. It is induced by the norms over the spaces $X_1, \ldots, X_d$ such that

$$\|v^{(1)} \otimes \ldots \otimes v^{(d)}\|_X = \|v^{(1)}\|_{X_1} \ldots \|v^{(d)}\|_{X_d} \tag{5.3}$$

holds for all $v^{(\nu)} \in X_\nu$, $\nu \in \{1, \ldots, d\}$, where $\|\cdot\|_{X_\nu}$ is the norm on the Hilbert space X_ν. We denote by $R_X : X \to X'$ the Riesz map associated to $\|\cdot\|_X$ such that for all $v \in X$,

$$\|v\|_X^2 = \langle R_X v, v\rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing. Then we have $R_X = R_{X_1} \otimes \ldots \otimes R_{X_d}$ where R_{X_ν} is the Riesz map associated to $\|\cdot\|_{X_\nu}$. Let us note that the inverse of R_X is given by $R_X^{-1} = R_{X_1}^{-1} \otimes \ldots \otimes R_{X_d}^{-1}$.

A natural norm for the space Z is the one induced by L and $\|\cdot\|_X$ which is defined by

$$\|z\|_Z = \inf_{v \in X, Lv = z} \|v\|_X \tag{5.4}$$

for all $z \in Z$. Here we need L to be surjective (*i.e.* $Z = \mathrm{range}(L)$).

Proposition 2.1. *The Riesz map associated to the norm* $\|\cdot\|_Z$ *defined by* (5.4) *is*

$$R_Z = (LR_X^{-1}L^*)^{-1}. \tag{5.5}$$

Furthermore we have

$$\|L\|_{X\to Z} = 1. \tag{5.6}$$

Proof: Since L is assumed to be surjective, the operator $LR_X^{-1}L^* : Z' \to Z$ is invertible, so that we can define $R_Z = (LR_X^{-1}L^*)^{-1}$. We show now that the norm $\|\cdot\|_Z$ defined by $\|\cdot\|_Z^2 = \langle R_Z\cdot, \cdot\rangle$ satisfies (5.4).

Let $z \in Z$. For any $v \in X$ such that $Lv = z$ we have

$$\|z\|_Z = \sup_{z'\in Z'} \frac{\langle z, z'\rangle}{\|z'\|_{Z'}} = \sup_{z'\in Z'} \frac{\langle Lv, z'\rangle}{\|L^*z'\|_{X'}} = \sup_{z'\in Z'} \frac{\langle v, L^*z'\rangle}{\|L^*z'\|_{X'}} \le \sup_{v'\in X'} \frac{\langle v, v'\rangle}{\|v'\|_{X'}} = \|v\|_X.$$

Taking the infimum over $v \in X$ subject to $Lv = z$, we obtain $\|z\|_Z \le \inf_{v\in X, Lv=z} \|v\|_X$. Moreover, let us consider $v_0 = R_X^{-1}L^*R_Z z \in X$. By construction, we have $Lv_0 = z$ so that

$$\inf_{v\in X, Lv=z} \|v\|_X \le \|v_0\|_X = \|L^*R_Z z\|_{X'} = \|z\|_Z.$$

Then we obtain (5.4).

To show (5.6), let us note that (5.4) yields

$$\|Lv\|_Z = \inf_{w\in X, Lw=Lv} \|w\|_X \overset{w=v}{\le} \|v\|_X$$

for all $v \in X$. Dividing by $\|v\|_X$ and taking the supremum over $v \in X$, we obtain $\|L\|_{Z\to X} \le 1$. Finally, we fix $v_0 \in X$ and we consider $w_0 = R_X^{-1}L^*R_Z Lv_0 \in X$. We have

$$\|L\|_{Z\to X} = \sup_{v\in X} \frac{\|Lv\|_Z}{\|v\|_X} \ge \frac{\|Lw_0\|_Z}{\|w_0\|_X} = \frac{\|Lv_0\|_Z}{\|Lv_0\|_Z} = 1,$$

which gives $\|L\|_{Z\to X} \ge 1$. Therefore we have (5.6). ∎

Remark 2.2. *The choice* (5.4) *for the norm* Z *is not mandatory, but it makes the forthcoming analysis easier. Also, this norm is natural for many applications. To illustrate this, suppose that* $X = H^1(\Omega)$*, where* Ω *is a Lipschitz domain, and let* L *be the trace operator associated to some part* Γ *of* $\partial\Omega$*. The space* $Z = H^{1/2}(\Gamma)$ *is the range of* L*, and the natural norm in* $H^{1/2}(\Gamma)$ *is defined by* (5.4).

2.2 Goal-oriented norm

The goal-oriented norm defined by equation (5.2) is associated to the Riesz map R_{X_α} given by

$$R_{X_\alpha} = R_X + \alpha L^* R_Z L. \tag{5.7}$$

Indeed, we can write $\langle R_{X_\alpha} v, v\rangle = \langle R_X v, v\rangle + \alpha\langle R_Z Lv, Lv\rangle = \|v\|^2_{X_\alpha}$ for $v \in X$. In the forthcoming sections, we will need the inverse of R_{X_α}. It is provided by the Woodbury formula

$$\begin{aligned} R_{X_\alpha}^{-1} &= (R_X + \alpha L^* R_Z L)^{-1} \\ &= R_X^{-1} - \alpha R_X^{-1} L^* (R_Z^{-1} + \alpha L R_X^{-1} L^*)^{-1} L R_X^{-1}, \\ &= R_X^{-1} - \alpha R_X^{-1} L^* ((1+\alpha) R_Z^{-1})^{-1} L R_X^{-1} \\ &= R_X^{-1} - \frac{\alpha}{1+\alpha} R_X^{-1} L^* R_Z L R_X^{-1}. \end{aligned} \tag{5.8}$$

Finally, let us note that

$$\|L\|^2_{X_\alpha \to Z} = \sup_{v\in X} \frac{\|Lv\|_Z^2}{\|v\|^2_{X_\alpha}} = \sup_{v\in X} \frac{\|Lv\|_Z^2}{\|v\|_X^2 + \alpha\|Lv\|_Z^2} = \frac{1}{\left(\sup\limits_{v\in X} \dfrac{\|Lv\|_Z^2}{\|v\|_X^2}\right)^{-1} + \alpha},$$

and since $\|L\|_{X\to Z} = 1$, we have

$$\|L\|_{X_\alpha \to Z} = \frac{1}{\sqrt{1+\alpha}}.$$

Let us assume that we dispose of an approximation v of u with a controlled relative precision $\varepsilon > 0$ measured with respect to the goal-oriented norm, *i.e.*

$$\|u - v\|_{X_\alpha} \le \varepsilon \|u\|_{X_\alpha}. \tag{5.9}$$

Provided $\|s\|_Z \neq 0$, we can write

$$\frac{\|s - Lv\|_Z^2}{\|s\|_Z^2} \le \|L\|^2_{X_\alpha \to Z} \frac{\|u - v\|^2_{X_\alpha}}{\|s\|_Z^2} \le \frac{\varepsilon^2}{1+\alpha} \frac{\|u\|^2_{X_\alpha}}{\|s\|_Z^2} = \varepsilon^2 \frac{\|u\|_X^2/\|s\|_Z^2 + \alpha}{1+\alpha},$$

so that the relation

$$\frac{\|s - Lv\|_Z}{\|s\|_Z} \le \varepsilon \sqrt{\frac{C^2 + \alpha}{1+\alpha}} \tag{5.10}$$

holds, where $C = \|u\|_X / \|s\|_Z \ge 1$.

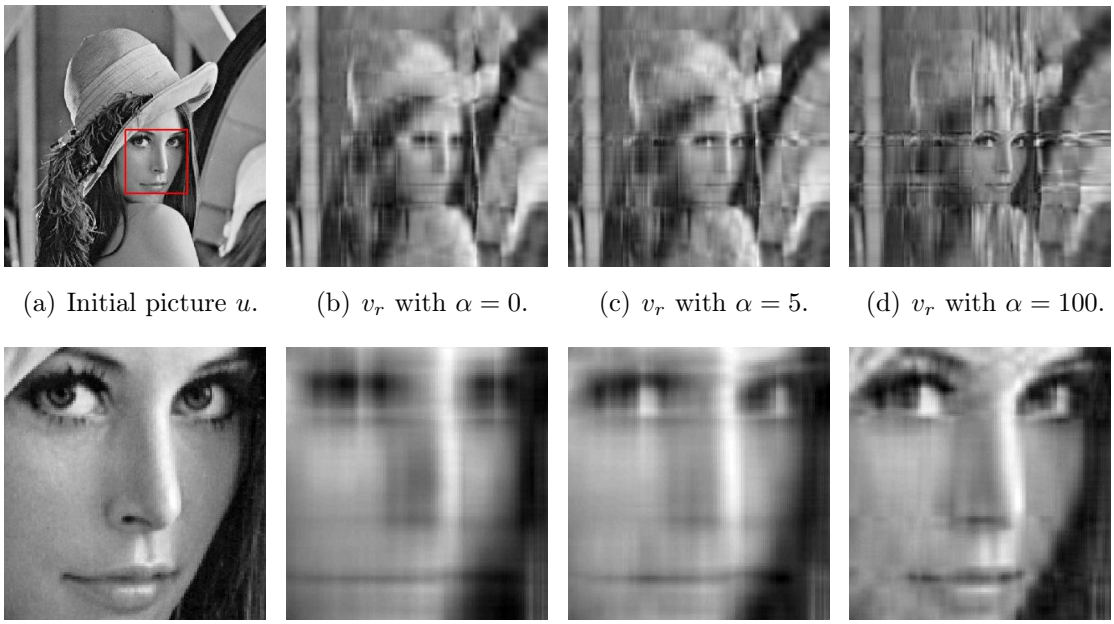
The advantage of using the goal-oriented norm is twofold. Firstly, for large values of α, the bound $\varepsilon\sqrt{(C^2+\alpha)/(1+\alpha)}$ in (5.10) is getting closer and closer to ε, so that the condition (5.9) tends to ensure a relative precision ε for the quantity of interest. Secondly, when considering low-rank approximations $v_r \in \mathcal{M}_r(X)$ of u, we expect the rank r which is necessary for obtaining (5.9) to decrease when α increases.

A simple illustration. To illustrate this second advantage, let us consider a matrix $u \in X = \mathbb{R}^{n\times n} = \mathbb{R}^n \otimes \mathbb{R}^n$ whose entries are the pixels of the picture given on Figure 5.1(a). When considering the canonical 2-norm for $\mathbb{R}^n$, the natural norm on X defined by (5.3) is the Frobenius norm such that $\|u\|_X^2 = \sum_{i,j=1}^n u_{i,j}^2$. We choose the quantity of interest $s \in Z$ as a part of the picture (see Figure 5.1(e)) represented by the red rectangle on Figure 5.1(a). We have $Z = \mathbb{R}^{l_1\times l_2}$ and $s_{i,j} = (Lu)_{i,j} = u_{\mathcal{I}_i,\mathcal{J}_j}$ where $\mathcal{I}$ and $\mathcal{J}$ are the vectors of the corresponding pixel indices. Here, the norm on Z defined by (5.4) is nothing but the Frobenius norm over $\mathbb{R}^{l_1\times l_2}$. We consider low-rank approximations of u in the subset of tensors $\mathcal{C}_r(X)$ with rank bounded by r. We solve the minimization problem $\min_{v_r\in\mathcal{C}_r(X)} \|u-v_r\|_{X_\alpha}$ for different values of r and α using an alternating minimization algorithm (see Chapter 1 Section 3.3.3). The approximations v_r and Lv_r are plotted on Figure 5.1 for $r = 10$. Qualitatively, the approximation of the quantity of interest is getting better when α increases. Table 5.1 also shows that for fixed ranks, the relative error $\|s - Lv_r\|_Z/\|s\|_Z$ is decreasing when α increases. However we don't observe significant improvements when $\alpha \geq 100$. In fact, since $L\mathcal{C}_r(X) = \mathcal{C}_r(Z)$, we have $\min_{v_r\in\mathcal{C}_r(X)} \|u - v_r\|_{X_\alpha}/\|u\|_{X_\alpha} \approx \min_{s_r\in\mathcal{C}_r(Z)} \|s - s_r\|_Z/\|u\|_Z$ for sufficiently large values of α, so that the approximation v_r yields the best possible approximation of the variable of interest $s_r \approx Lv_r$ in the low-rank tensor subset $\mathcal{C}_r(Z)$.

Of course, for this example, we could have computed directly a low-rank approximation $s_r \in \mathcal{C}_r(Z)$ of $s = Lu \in Z$ without using the goal-oriented norm. But for the intended applications where u is the solution of equation (5.1), s cannot be computed directly since u is unknown.

3 Algorithms for goal-oriented low-rank approximations

In this section, we propose algorithms to compute goal-oriented low-rank approximations, *i.e.* approximations of u in a set of low-rank tensors $\mathcal{M}_r(X)$ such that (5.9) holds for a given ε. We first present a method based on iterative solvers with



(a) Initial picture u. (b) v_r with $\alpha = 0$. (c) v_r with $\alpha = 5$. (d) v_r with $\alpha = 100$.

(e) Quantity of interest s. (f) Lv_r with $\alpha = 0$. (g) Lv_r with $\alpha = 5$. (h) Lv_r with $\alpha = 100$.

Figure 5.1: Goal-oriented low-rank approximation of a picture: representation of u and of the quantity of interest s (part of the picture corresponding to the red square in 5.1(a)). The rank for the approximation v_r is $r = 10$.

	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 50$
$\alpha = 0$	2.3×10^{-1}	1.2×10^{-1}	8.9×10^{-2}	3.9×10^{-2}	1.9×10^{-2}
$\alpha = 1$	1.8×10^{-1}	1.1×10^{-1}	8.7×10^{-2}	3.4×10^{-2}	1.4×10^{-2}
$\alpha = 5$	1.1×10^{-1}	6.8×10^{-2}	4.0×10^{-2}	2.1×10^{-2}	6.8×10^{-3}
$\alpha = 10$	1.1×10^{-1}	4.6×10^{-2}	3.1×10^{-2}	1.2×10^{-2}	3.5×10^{-3}
$\alpha = 100$	1.1×10^{-1}	5.2×10^{-2}	2.8×10^{-2}	1.1×10^{-2}	3.2×10^{-3}
$\alpha = 1000$	1.1×10^{-1}	5.2×10^{-2}	2.8×10^{-2}	1.0×10^{-2}	3.1×10^{-3}

Table 5.1: Goal-oriented low-rank approximation of a picture: relative error $\|s - Lv_r\|_Z / \|s\|_Z$ for different ranks r and weights α.

low-rank truncations using explicit preconditioners. Then, and similarly to [17], we propose a gradient-type algorithm which relies on an ideal minimal residual formulation. This algorithm can be interpreted as a gradient algorithm using an implicit preconditioner.

3.1 Iterative solver with goal-oriented truncations

We propose here a Preconditioned Conjugate Gradient (PCG) for the solution of equation (5.1). Such strategy as been already proposed in [89]. This algorithm requires the operator $A : X \to X'$ to be symmetric positive definite (SPD), and it also requires a preconditioner $P : X' \to X$ which is also SPD. When using representations of the iterates in a low-rank tensor format $\mathcal{M}_r(X)$, one needs an additional truncation step to avoid the representation rank to blow up. We propose here to define truncations by means of an approximation operator $\Pi_\alpha^\varepsilon : X \to \mathcal{M}_r(X)$ such that for any $x \in X$, $\Pi_\alpha^\varepsilon(x)$ is a low-rank approximation of x with a relative precision ε with respect to the goal-oriented norm $\|\cdot\|_{X_\alpha}$, *i.e.* such that

$$\|\Pi_\alpha^\varepsilon(x) - x\|_{X_\alpha} \leq \varepsilon \|x\|_{X_\alpha}. \tag{5.11}$$

The resulting algorithm is summarized in Algorithm 8. Note that Π_α^ε is applied in steps 4 and 8.

There are different possibilities for the definition of the truncation operator Π_α^ε. Note that in the general case, the goal-oriented norm $\|\cdot\|_{X_\alpha}$ is not a crossnorm, meaning that there do not exist norms on the spaces X_ν such that $\|\cdot\|_{X_\alpha}$ satisfies (5.3) (see the structure of R_{X_α} given by (5.7), which is the sum of two terms). As a consequence, one cannot use methods based on Singular Value Decompositions (SVD) for the low-rank approximation of a tensor with respect to the goal-oriented norm (see Chapter 1, Section 3.3.1). Then we propose to use a greedy algorithm based on rank-one corrections with additional update steps, as described in Chapter 1, Section 3.3.4 (see also Appendix 7).

In some particular situations, the goal-oriented norm can be a crossnorm. This is the case when considering a norm induced by goal-oriented norms $\|\cdot\|_{X_\nu}$ on the spaces X_ν, see Section 4.2. Then low-rank approximation based on SVD can be applied for the practical implementation of Π_α^ε.

Note that the principle of the PCG is to minimize the error measured with the norm $\|\cdot\|_A$ defined by $\|\cdot\|_A^2 = \langle A\cdot, \cdot\rangle$. In other words, the PCG algorithm does not really minimize the error associated to the goal-oriented norm. Only the truncation steps are goal-oriented. In the following section, we propose a method which is based on an ideal minimal residual formulation which aims at minimizing the error measured with the goal-oriented norm.

Algorithm 8 Preconditioned Conjugate Gradient with goal-oriented truncation

Require: $A : X \to X'$, $b \in X'$, a preconditioner $P : X' \to X$, and a truncation operator $\Pi_\alpha^\varepsilon : X \to X$.

Initialize $r^0 = b$, $z^0 = Pr^0$, $p^0 = z^0$ and $k = 0$.
while u^k not converged **do**
  $\alpha^k = \langle r^k, p^k\rangle / \langle p^k, Ap^k\rangle$;
  $u^{k+1} = \Pi_\alpha^\varepsilon(u^k + \alpha^k p^k)$;
  $r^{k+1} = b - Au^{k+1}$;
  $z^{k+1} = Pr^{k+1}$;
  $\beta^k = -\langle z^{k+1}, Ap^k\rangle / \langle p^k, Ap^k\rangle$;
  $p^{k+1} = \Pi_\alpha^\varepsilon(z^k + \beta^k p^k)$;
  $k = k + 1$;
end while
return u^k

3.2 A method based on an ideal minimal residual formulation

We propose now a method for the goal-oriented low-rank approximation of u which relies on an ideal minimal residual formulation. The principle is to use an *ideal residual norm* which allows us to reformulate the approximation problem (5.9) without involving the solution u itself. Then, and similarly to [17] (see Chapter 4), we propose a gradient-type algorithm for the solution of the resulting minimization problem. Contrarily to the PCG algorithm introduced in the previous section, this method does not require the operator A to be SPD.

We assume that A is an operator from X to Y', the dual space of a Hilbert space Y. This space is endowed with an ideal residual norm $\|\cdot\|_{Y_\alpha}$ such that

$$\|Av - b\|_{Y'_\alpha} = \|u - v\|_{X_\alpha} \tag{5.12}$$

for any $v \in X$. The relations

$$\begin{aligned} \|Av\|^2_{Y'_\alpha} &= \langle R_{Y_\alpha}^{-1} Av, Av\rangle = \langle A^* R_{Y_\alpha}^{-1} Av, v\rangle, \\ \|v\|^2_{X_\alpha} &= \langle R_{X_\alpha} v, v\rangle, \end{aligned}$$

hold for any $v \in X$. Then the Riesz map R_{Y_α} associated to the norm $\|\cdot\|_{Y_\alpha}$ satisfies

$$A^* R_{Y_\alpha}^{-1} A = R_{X_\alpha} \;\Leftrightarrow\; R_{Y_\alpha}^{-1} = A^{-*} R_{X_\alpha} A^{-1} \;\Leftrightarrow\; R_{Y_\alpha} = A R_{X_\alpha}^{-1} A^*. \tag{5.13}$$

Relation (5.8) provides an explicit expression for $R_{X_\alpha}^{-1}$, which also provides an explicit expression for R_{Y_α}. Thanks to relation (5.12), the equation (5.9) can be rewritten

as the following condition on the residual

$$\|Av_r - b\|_{Y'_\alpha} \leq \varepsilon \|b\|_{Y'_\alpha}, \tag{5.14}$$

which does no longer involve the solution u. However, the norm $\|\cdot\|_{Y'_\alpha}$ can not be directly computed since it requires the inverse of the Riesz map R_{Y_α}, which is not available.

We propose now a variant of the gradient-type algorithm introduced in [17] for computing a low-rank approximation of u verifying (5.14). We assume that Y possesses a tensor product structure $Y = Y_1 \otimes \ldots \otimes Y_d$, and we define the low-rank approximation operators $\Lambda_\alpha^\delta : Y \to Y$ such that for any $y \in Y$, $\Lambda_\alpha^\delta(y)$ is a low-rank approximation of y satisfying

$$\|\Lambda_\alpha^\delta(y) - y\|_{Y_\alpha} \leq \delta \|y\|_{Y_\alpha} \tag{5.15}$$

Let $\{u^k\}_{k\geq 0} \subset X$ and $\{y^k\}_{k\geq 1} \subset Y$ be two sequences defined by

$$y^{k+1} = \Lambda_\alpha^\delta(R_{Y_\alpha}^{-1}(Au^k - b)), \tag{5.16}$$

$$u^{k+1} = \Pi_\alpha^\varepsilon(u^k - R_{X_\alpha}^{-1} A^* y^{k+1}), \tag{5.17}$$

with $u^0 = 0$, and where Π_α^ε is the approximation operator defined in the previous section (see Equation (5.11)). The resulting algorithm (5.16)–(5.17) can be interpreted as a gradient-type algorithm. Indeed, the quantity $R_{X_\alpha}^{-1} A^* y^{k+1}$ in (5.17) can be interpreted as an approximation (with a relative precision δ) of the gradient $(u^k - u)$ of the function $v \mapsto \|(u - u^k) - v\|_{X_\alpha}^2$. The following proposition gives a convergence result for this algorithm.

Proposition 3.1. *Assuming that $\delta(1+\varepsilon) < 1$, the sequence $\{u^k\}_{k\geq 0}$ defined by (5.16)–(5.17) satisfies*

$$\frac{\|u - u^k\|_{X_\alpha}}{\|u\|_{X_\alpha}} \leq \frac{\varepsilon}{1 - \delta(1+\varepsilon)} + (\delta(1+\varepsilon))^k. \tag{5.18}$$

Proof: Let $w^{k+1} = u^k - R_{X_\alpha}^{-1} A^* y^{k+1}$ and $r^k = Au^k - b$. For any $k \geq 0$ we have

$$\begin{aligned}
\|u - u^{k+1}\|_{X_\alpha} &= \|(u - w^{k+1}) - (\Pi_\alpha^\varepsilon(w^{k+1}) - w^{k+1})\|_{X_\alpha} \\
&\leq \|u - w^{k+1}\|_{X_\alpha} + \|\Pi_\alpha^\varepsilon(w^{k+1}) - w^{k+1}\|_{X_\alpha} \\
&\overset{(5.11)}{\leq} \|u - w^{k+1}\|_{X_\alpha} + \varepsilon\|w^{k+1}\|_{X_\alpha} \\
&= \|u - w^{k+1}\|_{X_\alpha} + \varepsilon\|(u - w^{k+1}) - u\|_{X_\alpha} \\
&\leq (1+\varepsilon)\|u - w^{k+1}\|_{X_\alpha} + \varepsilon\|u\|_{X_\alpha} \\
&= (1+\varepsilon)\|R_{Y_\alpha}^{-1}(Au^k - b) - y^{k+1}\|_{Y_\alpha} + \varepsilon\|u\|_{X_\alpha} \\
&\overset{(5.15)}{\leq} \delta(1+\varepsilon)\|R_{Y_\alpha}^{-1}(Au^k - b)\|_{Y_\alpha} + \varepsilon\|u\|_{X_\alpha} \\
&= \delta(1+\varepsilon)\|u - u^k\|_{X_\alpha} + \varepsilon\|u\|_{X_\alpha}.
\end{aligned}$$

Provided $\delta(1+\varepsilon) < 1$, we obtain by recurrence that

$$\|u - u^k\|_{X_\alpha} \leq (\delta(1+\varepsilon))^k \|u - u^0\|_{X_\alpha} + \varepsilon \frac{1 - (\delta(1+\varepsilon))^k}{1 - \delta(1+\varepsilon)} \|u\|_{X_\alpha}$$

holds for $k \geq 1$. Since $u^0 = 0$ we obtain (5.18). ∎

In the same way as for Π_α^ε, we can use a greedy algorithm with additional update steps for the definition of the approximation operator Λ_α^δ. Denoting $r^k = Au^k - b$, the idea is to minimize the function $J : y \mapsto \|R_{Y_\alpha}^{-1} r^k - y\|_{Y_\alpha}^2$ over subsets of low-rank tensors. The solution of this minimization problem does not require the knowledge of $R_{Y_\alpha}^{-1} r^k$. Indeed the corresponding stationarity condition is $\langle r^k - R_{Y_\alpha} y, \widetilde{y}\rangle$ for any $\widetilde{y}$ belonging to the tangent space of a subset of low-rank tensors. However, evaluations of the function J cannot be obtained without knowing $R_{Y_\alpha}^{-1} r^k$. In practice, an error estimator based on stagnation is used. Therefore, the condition (5.15) may not be satisfied (indeed, the algorithm may stop before reaching the relative precision δ), so that the convergence result given by Proposition 3.1 is not ensured. Let us note that this algorithm is not the only possible algorithm for the definition of Λ_α^δ. One can for example use an iterative solver where the iteration process is stopped when the desired precision δ is reached.

4 Application to uncertainty quantification

We consider a parameter-dependent equation $A(\xi)u(\xi) = b(\xi)$, where $u(\xi)$ belongs to a Hilbert space V. The parameter $\xi = (\xi_1, \ldots, \xi_d)$ is a random vector that takes

values in the parameter set $\Xi = \Xi_1 \times \ldots \times \Xi_d$, with $\Xi_\nu \subset \mathbb{R}$. We assume that the components of ξ are independent, and that

$$u \in X = V \otimes S_1 \otimes \ldots \otimes S_d,$$

where for all $\nu \in \{1, \ldots, d\}$, $S_\nu = L^2_{\mu^{(\nu)}}(\Xi_\nu)$ and $\mu^{(\nu)}$ is the law of ξ_ν. In this section, we discuss different possibilities for quantities of interest of the form $s = Lu$, where $L : X \to Z$ is a linear function. Then we detail the structure of the natural norm in Z, and of the goal-oriented norm in X. Finally, approximation based on polynomial interpolation of the solution map is considered.

4.1 Linear quantities of interest

First we consider the case where we want to compute the expectation of a variable of interest defined by $\xi \mapsto \langle l, u(\xi)\rangle$ for some extractor $l \in V'$. For simplicity, we assume that l does not depend on ξ. The quantity of interest $s \in Z = \mathbb{R}$ is

$$s = \mathbb{E}(\langle l, u(\xi)\rangle), \tag{5.19}$$

which linearly depends on u. For any $\nu \in \{1, \ldots, d\}$, we define the *expectation function* $e^{(\nu)} \in S'_\nu$ such that for any $\lambda \in S_\nu$, $\langle e^{(\nu)}, \lambda\rangle = \int_{\Xi_\nu} \lambda(\xi_\nu) \mathrm{d}\mu^{(\nu)}(\xi_\nu)$ is the expectation of λ. We introduce the linear application L defined by

$$L = l \otimes e^{(1)} \otimes \ldots \otimes e^{(d)}, \tag{5.20}$$

such that for any elementary tensor $v = v^{(0)} \otimes v^{(1)} \otimes \ldots \otimes v^{(d)} \in X$ with $v^{(0)} \in V$ and $v^{(\nu)} \in S_\nu$, we have

$$\begin{aligned} Lv &= \langle l, v^{(0)}\rangle \langle e^{(1)}, v^{(1)}\rangle \ldots \langle e^{(d)}, v^{(d)}\rangle, \\ &= \langle l, v^{(0)}\rangle \; \mathbb{E}\big(v^{(1)}(\xi_1)\big) \ldots \mathbb{E}\big(v^{(d)}(\xi_d)\big), \\ &= \mathbb{E}(\langle l, v(\xi)\rangle). \end{aligned}$$

Then L is extended by linearity to X, so that s defined by (5.19) satisfies $s = Lu$. Note that L is a rank-one tensor. When considering a parameter-dependent extractor $l(\xi)$, the rank of L can be larger than one.

Remark 4.1 (Vector-valued or functional-valued variable of interest). *Assume that we are interested in the expectation of a vector-valued or functional-valued variable of interest $\xi \mapsto l(u(\xi)) \in W$, where $l \in \mathcal{L}(V, W)$ with W a Hilbert space. For example, $l(u(\xi))$ can be a vector containing several scalar-valued variables of interest (W is then an Euclidean space), or a function (W is then a function space). We can also choose l as the identity operator on V (with $W = V$), meaning*

that the variable of interest is the solution $u(\xi)$ itself. In any case, the quantity of interest $s = \mathbb{E}(l(u(\xi)))$ belongs to $Z = W$, and we can write $s = Lu$ with L defined as in (5.20).

In the case where we want to compute the variable of interest itself, the quantity of interest $s \in Z = S_1 \otimes \ldots \otimes S_d = L^2_\mu(\Xi)$ is defined by

$$s : \xi \mapsto \langle l, u(\xi) \rangle. \tag{5.21}$$

We can write $s = Lu$ with

$$L = l \otimes I^{(1)} \otimes \ldots \otimes I^{(d)}, \tag{5.22}$$

where for any $\nu \in \{1, \ldots, d\}$, $I^{(\nu)}$ denotes the identity operator on S_ν.

Now, assume that we want to compute the conditional expectation of a variable of interest $\xi \mapsto \langle l, u(\xi) \rangle$ with respect to the random variables $\xi_\tau = (\xi_\nu)_{\nu \in \tau}$, where $\tau \subset \{1, \ldots, d\}$. The quantity of interest $s \in Z = \otimes_{\nu \in \tau} S_\nu$ is defined by

$$s : \xi_\tau \mapsto \mathbb{E}(\langle l, u(\xi) \rangle | \xi_\tau). \tag{5.23}$$

We have $s = Lu$ where L is defined by

$$L = l \otimes l^{(1)} \otimes \ldots \otimes l^{(d)} \quad \text{with} \quad \begin{cases} l^{(\nu)} = I^{(\nu)} & \text{if } \nu \in \tau \\ l^{(\nu)} = e^{(\nu)} & \text{if } \nu \notin \tau \end{cases}. \tag{5.24}$$

The conditional expectation appears in the expression of the Sobol index[1] $S_\tau = \mathbb{V}(s(\xi_\tau))/\mathbb{V}(\langle l, u(\xi) \rangle)$, which represents the contribution of the random variables ξ_τ to the variance of the variable of interest, see [117].

Remark 4.2. *Note that if $\tau = \emptyset$, L defined by (5.24) yields the expectation of the variable of interest, see equation (5.20). If $\tau = \{1, \ldots, d\}$, L corresponds to the definition (5.22), meaning that the quantity of interest is the variable of interest itself.*

4.2 Properties of the norms

For any $\nu \in \{1, \ldots, d\}$, the natural norm $\| \cdot \|_{S_\nu}$ in S_ν is such that $\|\lambda\|^2_{S_\nu} = \langle R_{S_\nu} \lambda, \lambda \rangle = \int_{\Xi_\nu} \lambda(\xi_\nu)^2 \mathrm{d}\mu^{(\nu)}(\xi_\nu)$ for any $\lambda \in S_\nu$. The Riesz map associated to the natural norm in X is $R_X = R_V \otimes R_{S_1} \otimes \ldots \otimes R_{S_d}$, where R_V is the Riesz map

[1] $\mathbb{V}$ denotes the variance.

associated to $\|\cdot\|_V$.

We discuss now the structure and the properties of the natural norm of Z, where L is defined by (5.24) for some $\tau \subset \{1,\dots,d\}$. Thanks to relation (5.5), we can write

$$\begin{aligned} R_Z^{-1} &= L R_X^{-1} L^*, \\ &= \langle R_V^{-1} l, l\rangle \Big(\prod_{\nu\notin\tau} \langle R_{S_\nu}^{-1} e^{(\nu)}, e^{(\nu)}\rangle \Big) \Big(\bigotimes_{\nu\in\tau} I^{(\nu)} R_{S_\nu}^{-1} I^{(\nu)*} \Big), \\ &= \Big(\|l\|_{V'}^2 \prod_{\nu\notin\tau} \|e^{(\nu)}\|_{S'_\nu}^2 \Big) \bigotimes_{\nu\in\tau} R_{S_\nu}^{-1}. \end{aligned} \tag{5.25}$$

Note that for any $\nu \in \{1,\dots,d\}$ we have

$$\|e^{(\nu)}\|_{S'_\nu} = \sup_{\lambda\in S_\nu} \frac{\langle e^{(\nu)}, \lambda\rangle}{\|\lambda\|_{S_\nu}} = \sup_{\lambda\in S_\nu} \frac{\int_{\Xi_\nu} \lambda(\xi_\nu) \mathrm{d}\mu^{(\nu)}(\xi_\nu)}{\|\lambda\|_{L^2_{\mu^{(\nu)}}(\Xi_\nu)}} \leq \sup_{\lambda\in S_\nu} \frac{\|\lambda\|_{L^1_{\mu^{(\nu)}}(\Xi_\nu)}}{\|\lambda\|_{L^2_{\mu^{(\nu)}}(\Xi_\nu)}} \leq 1,$$

and

$$\|e^{(\nu)}\|_{S'_\nu} = \sup_{\lambda\in S_\nu} \frac{\int_{\Xi_\nu} \lambda(\xi_\nu) \mathrm{d}\mu^{(\nu)}(\xi_\nu)}{\|\lambda\|_{L^2_{\mu^{(\nu)}}(\Xi_\nu)}} \overset{\lambda(\xi_\nu)=1}{\geq} 1.$$

Then $\|e_\nu\|_{S'_\nu} = 1$, so that relation (5.25) provides the following simple expression for the Riesz map R_Z:

$$R_Z = \frac{1}{\|l\|_{V'}^2} \bigotimes_{\nu\in\tau} R_{S_\nu}. \tag{5.26}$$

As a consequence, the natural norm for Z is such that $\|Lv\|_Z = \|l\|_{V'}^{-1} \|Lv\|_{L^2_{\mu^{(\tau)}}(\Xi_\tau)}$ for any $v \in X$, where $\Xi_\tau = \times_{\nu\in\tau} X_\nu$ and $\mu^{(\tau)} = \otimes_{\nu\in\tau}\mu^{(\nu)}$. Up to the constant $\|l\|_{V'}^{-1}$, this is the usual norm of $L^2_{\mu^{(\tau)}}(\Xi_\tau)$.

Finally, let us consider the case where the quantity of interest is defined by (5.21), or equivalently by (5.23) with $\tau = \{1,\dots,d\}$. For any elementary tensor $v = v^{(0)} \otimes v^{(1)} \otimes \dots \otimes v^{(d)} \in X$, with $v^{(0)} \in V$ and $v^{(\nu)} \in S_\nu$, and according to (5.26), we have

$$\begin{aligned} \|v\|_{X_\alpha}^2 &= \|v\|_X^2 + \alpha \|Lv\|_Z^2, \\ &= \|v^{(0)}\|_V^2 \|v^{(1)}\|_{S_1}^2 \dots \|v^{(d)}\|_{S_d}^2 + \alpha \|l\|_{V'}^{-2} \langle l, v^{(0)}\rangle^2 \|v^{(1)}\|_{S_1}^2 \dots \|v^{(d)}\|_{S_d}^2, \\ &= \big(\underbrace{\|v^{(0)}\|_V^2 + \alpha \|l\|_{V'}^{-2} \langle l, v^{(0)}\rangle^2}_{:=\|v^{(0)}\|_{V_\alpha}^2} \big) \|v^{(1)}\|_{S_1}^2 \dots \|v^{(d)}\|_{S_d}^2, \end{aligned}$$

where $\|\cdot\|_{V_\alpha}$ can be interpreted as a goal-oriented norm on V. As a consequence, the norm $\|\cdot\|_{X_\alpha}$ is a crossnorm which is induced by the norms $\|\cdot\|_{V_\alpha}, \|\cdot\|_{S_1}, \dots, \|\cdot\|_{S_d}$.

4.3 Approximation of $u(\xi)$ by interpolation

In practice, the computation of the quantity of interest requires an approximation of the solution map. In this section, we consider a polynomial interpolation of $\xi \mapsto u(\xi)$. We interpret the collection of the solution evaluations at the interpolation points as a tensor which is the solution of a tensor-structured equation. Then we address the question of the natural norms for the approximation of this tensor.

Remark 4.3. *Interpolation is not the only possibility for the approximation of u. For example, we could have considered a* spectral stochastic Galerkin *approximation (using polynomial approximation space). The proposed goal-oriented low-rank approximation method also applies in this framework.*

For the sake of simplicity, we assume that V is the euclidian space $\mathbb{R}^n$ of dimension n ($A(\xi)$ is a n-by-n matrix). For any $\nu \in \{1, \dots, d\}$, we consider the Lagrange polynomials $\psi_1^{(\nu)}, \dots, \psi_{p_\nu}^{(\nu)}$ of degree $p_\nu - 1$ such that $\psi_i^{(\nu)}(\xi_\nu^j) = \delta_{ij}$ for any $1 \leq i, j \leq p_\nu$, where $\xi_\nu^1, \dots, \xi_\nu^{p_\nu}$ are the interpolation points. Here we choose the Gauss points, meaning that ξ_ν^j is the j-th root of a polynomial of degree p_ν which is orthogonal (with respect to the scalar product of S_ν) to any polynomial of degree strictly less than p_ν. We consider the multivariate interpolation $\widetilde{u}$ of u defined by

$$\widetilde{u}(\xi) = \sum_{i_1=1}^{p_1} \dots \sum_{i_d=1}^{p_d} \mathbf{u}_{i_1,\dots,i_d} \prod_{\nu=1}^{d} \psi_{i_\nu}^{(\nu)}(\xi_\nu). \tag{5.27}$$

where $\mathbf{u}_{i_1,\dots,i_d} = u\big(\xi_1^{i_1}, \dots, \xi_d^{i_d}\big)$. The tensor $\mathbf{u}$ satisfies

$$\mathbf{A}\mathbf{u} = \mathbf{b}, \quad \mathbf{u} \in \mathbf{X} := \mathbb{R}^n \otimes \mathbb{R}^{p_1} \otimes \dots \otimes \mathbb{R}^{p_d}, \tag{5.28}$$

where $\mathbf{A}$ is the block diagonal operator that contains the matrix $A(\xi_1^{i_1}, \dots, \xi_d^{i_d})$ on the $(i_1, \dots, i_d)$-th term of the "super" diagonal, and $\mathbf{b}$ the vector containing $b(\xi_1^{i_1}, \dots, \xi_d^{i_d})$ on its $(i_1, \dots, i_d)$-th component. Note that problem (5.28) is an algebraic version of problem (5.1).

In the following, for any $\mathbf{v} \in \mathbf{X}$, we denote by $\widetilde{v} \in X$ the polynomial function defined by (5.27) where we replace $\mathbf{u}$ by $\mathbf{v}$ and $\widetilde{u}$ by $\widetilde{v}$. We show how to compute $L\widetilde{v}$, starting with the case where L is defined by (5.20). For any elementary tensor $\mathbf{v} = \mathbf{v}^{(0)} \otimes \dots \otimes \mathbf{v}^{(d)}$ in $\mathbf{X}$, with $\mathbf{v}^{(0)} \in \mathbb{R}^n$ and $\mathbf{v}^{(\nu)} \in \mathbb{R}^{p_\nu}$, we have

$$\widetilde{v} = \mathbf{v}^{(0)} \otimes \left(\sum_{i_1=1}^{p_1} \mathbf{v}_{i_1}^{(1)} \psi_{i_1}^{(1)} \right) \otimes \dots \otimes \left(\sum_{i_d=1}^{p_d} \mathbf{v}_{i_d}^{(d)} \psi_{i_d}^{(d)} \right), \tag{5.29}$$

so that we can write

$$\begin{aligned} L\widetilde{v} &= \langle l, \mathbf{v}^{(0)} \rangle \left(\sum_{i_1=1}^{p_1} \mathbf{v}_{i_1}^{(1)} \langle e_1, \psi_{i_1}^{(1)} \rangle \right) \dots \left(\sum_{i_d=1}^{p_d} \mathbf{v}_{i_d}^{(d)} \langle e_d, \psi_{i_d}^{(d)} \rangle \right), \\ &= \langle l, \mathbf{v}^{(0)} \rangle \langle \omega^{(1)}, \mathbf{v}^{(1)} \rangle \dots \langle \omega^{(d)}, \mathbf{v}^{(d)} \rangle = \mathbf{L}\mathbf{v}, \end{aligned}$$

where $\mathbf{L} \in \mathbf{X}'$ is defined by $\mathbf{L} = l \otimes \omega^{(1)} \otimes \dots \otimes \omega^{(d)}$, with $\omega^{(\nu)} \in (\mathbb{R}^{p_\nu})'$. For the sake of simplicity, we consider now that $\omega^{(\nu)} \in \mathbb{R}^{p_\nu}$ is a vector, and that the duality paring $\langle \cdot, \cdot \rangle$ is the canonical scalar product of $\mathbb{R}^{p_\nu}$. Then $\omega_{i_\nu}^{(\nu)} = \int_{\Xi_\nu} \psi_{i_\nu}^{(\nu)}(\xi_\nu) \mathrm{d}\mu^{(\nu)}(\xi_\nu)$ is the *weight* associated to the Gauss point $\xi_\nu^{i_\nu}$. The generalization to the case where L is given by (5.24) is straightforward. The space $\mathbf{Z}$ is given by $\mathbf{Z} = \otimes_{\nu \in \tau} \mathbb{R}^{p_\nu}$ and the application $\mathbf{L} : \mathbf{X} \to \mathbf{Z}$ is

$$\mathbf{L} = l \otimes l^{(1)} \otimes \dots \otimes l^{(d)} \quad \text{with} \quad \begin{cases} l^{(\nu)} = I_{p_\nu} & \text{if } \nu \in \tau \\ l^{(\nu)} = \omega^{(\nu)} & \text{if } \nu \notin \tau \end{cases}, \tag{5.30}$$

where I_{p_ν} is the identity matrix of size p_ν. Then the quantity of interest $L\widetilde{v} \in Z$ is given by

$$L\widetilde{v} = \sum_{i \in \mathcal{I}_\tau} (\mathbf{L}\mathbf{u})_i \bigotimes_{\nu \in \tau} \psi_{i_\nu}^{(\nu)},$$

where $\mathcal{I}_\tau = \times_{\nu \in \tau} \{1, \dots, p_\nu\}$.

Now we address the question of the norm in $\mathbf{X}$. We propose to define $\| \cdot \|_{\mathbf{X}}$ such that the relation $\|\mathbf{v}\|_{\mathbf{X}} = \|\widetilde{v}\|_X$ holds for any $\mathbf{v} \in \mathbf{X}$. Given an elementary tensor $\mathbf{v} = \mathbf{v}^{(0)} \otimes \dots \otimes \mathbf{v}^{(d)}$, and thanks to (5.29), we can write

$$\begin{aligned} \|\widetilde{v}\|_X^2 &= \|\mathbf{v}^{(0)}\|_V^2 \prod_{\nu=1}^{d} \| \sum_{i_\nu=1}^{p_\nu} \mathbf{v}_{i_\nu}^{(\nu)} \psi_{i_\nu}^{(\nu)} \|_{S_\nu}^2, \\ &= \langle R_V \mathbf{v}^{(0)}, \mathbf{v}^{(0)} \rangle \prod_{\nu=1}^{d} \int_{\Xi_\nu} \Big(\sum_{i_\nu=1}^{p_\nu} \mathbf{v}_{i_\nu}^{(\nu)} \psi_{i_\nu}^{(\nu)}(\xi_\nu) \Big)^2 \mathrm{d}\mu^{(\nu)}(\xi_\nu), \\ &= \langle R_V \mathbf{v}^{(0)}, \mathbf{v}^{(0)} \rangle \prod_{\nu=1}^{d} \sum_{i_\nu, j_\nu=1}^{p_\nu} \mathbf{v}_{i_\nu}^{(\nu)} \mathbf{v}_{j_\nu}^{(\nu)} \underbrace{\int_{\Xi_\nu} \psi_{i_\nu}^{(\nu)}(\xi_\nu) \psi_{j_\nu}^{(\nu)}(\xi_\nu) \mathrm{d}\mu^{(\nu)}(\xi_\nu)}_{:=(R_\nu)_{i_\nu, j_\nu}}, \\ &= \langle R_V \mathbf{v}^{(0)}, \mathbf{v}^{(0)} \rangle \prod_{\nu=1}^{d} \langle R_\nu \mathbf{v}^{(\nu)}, \mathbf{v}^{(\nu)} \rangle = \langle R_{\mathbf{X}} \mathbf{v}, \mathbf{v} \rangle. \end{aligned}$$

Here, the Riesz map $R_{\mathbf{X}}$ associated to $\| \cdot \|_{\mathbf{X}}$ is given by $R_{\mathbf{X}} = R_V \otimes R_1 \otimes \dots \otimes R_d$, where for $\nu \in \{1, \dots, d\}$, R_ν is the Gram matrix associated to the polynomial basis

$\{\psi_1^{(\nu)}, \dots, \psi_{p_\nu}^{(\nu)}\}$. The advantage of using the Gauss points for the interpolation is that this basis is orthogonal. Indeed, the quadrature formula $\sum_{k=1}^{p_\nu} \omega_k^{(\nu)} f(\xi_\nu^k)$ for the approximation of $\int_{\Xi_\nu} f(\xi_\nu) \mathrm{d}\mu^{(\nu)}(\xi_\nu)$ is exact when f is any polynomial of degree less than $2p_\nu - 1$. Since for all $1 \leq i, j \leq p_\nu$, the polynomial $\psi_i^{(\nu)}\psi_j^{(\nu)}$ is of degree $2(p_\nu - 1)$, we can write

$$\int_{\Xi_\nu} \psi_i^{(\nu)}(\xi_\nu)\psi_j^{(\nu)}(\xi_\nu)\mathrm{d}\mu^{(\nu)}(\xi_\nu) = \sum_{k=1}^{p} \omega_k^{(\nu)}\psi_i^{(\nu)}(\xi_\nu^k)\psi_j^{(\nu)}(\xi_\nu^k) = \sum_{k=1}^{p} \omega_k^{(\nu)}\delta_{i,k}\delta_{j,k} = \omega_i^{(\nu)}\delta_{i,j}$$

As a consequence we have $R_\nu = \mathrm{diag}(\omega^{(\nu)})$. In particular, the computation of the inverse of R_{S_ν} is straightforward.

To conclude this section, we note that the natural norm in $\mathbf{Z}$ (defined by (5.4) when replacing $\|\cdot\|_X$ by $\|\cdot\|_{\mathbf{X}}$ and L by $\mathbf{L}$) satisfies the relation

$$\|\mathbf{Lv}\|_{\mathbf{Z}} = \inf_{\substack{\mathbf{w}\in\mathbf{X} \\ \mathbf{Lw}=\mathbf{Lv}}} \|\mathbf{w}\|_{\mathbf{X}} = \inf_{\substack{\mathbf{w}\in\mathbf{X} \\ L\widetilde{w}=L\widetilde{v}}} \|\widetilde{w}\|_X \geq \inf_{\substack{w\in X \\ Lw=L\widetilde{v}}} \|w\|_X = \|L\widetilde{v}\|_Z$$

for any $\mathbf{v} \in \mathbf{X}$. Then we can write

$$\|Lu - L\widetilde{v}\|_Z \leq \|Lu - L\widetilde{u}\|_Z + \|L\widetilde{u} - L\widetilde{v}\|_Z \leq \|u - \widetilde{u}\|_X + \|\mathbf{Lu} - \mathbf{Lv}\|_{\mathbf{Z}},$$

so that the error on the quantity of interest $\|Lu - L\widetilde{v}\|_Z$ is bounded by the sum of the interpolation error $\|u - \widetilde{u}\|_X$ and the approximation error $\|\mathbf{Lu} - \mathbf{Lv}\|_{\mathbf{Z}}$. This suggests that it is not necessary to compute an approximation $\mathbf{v}$ of $\mathbf{u}$ with a precision (on the quantity of interest $\mathbf{Lu}$) lower than the interpolation error. Also, this justifies the use of the goal-oriented norm for the space $\mathbf{X}$ since, up to the interpolation error, the error $\|\mathbf{Lu} - \mathbf{Lv}\|_{\mathbf{Z}}$ controls the precision on the quantity of interest.

5 Numerical experiments

5.1 Iterative solver (PCG) with truncations

5.1.1 The cookie problem

Consider the boundary value problem

$$\begin{aligned} -\nabla \cdot (\kappa(\xi)\nabla u_{\mathrm{ex}}(\xi)) &= f \text{ on } \Omega, \\ u_{\mathrm{ex}}(\xi) &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.31}$$

The diffusion coefficient $\kappa(\xi)$ is equal to 1 on $\Omega_0 \cup \Omega_9$, and to ξ_i on the domain Ω_i for $i \in \{1, \dots, 8\}$, see Figure 5.2. Here, $\xi = (\xi_1, \dots, \xi_8)$ is a random vector

with independent components which are log-uniformly distributed on $[10^{-1}, 10^{1}]$. The source term f is equal to 1 on Ω_9 and zero elsewhere. We consider a finite element approximation space $\text{span}\{\phi_1, \ldots, \phi_n\} \subset H_0^1(\Omega)$ of dimension $n = 2413$, where $\{\phi_1, \ldots, \phi_n\}$ are piecewise linear functions associated to the mesh given on Figure 5.2. The vector $u(\xi) \in V = \mathbb{R}^n$ containing the coefficients of the Galerkin approximation $u^h(\xi) = \sum_{i=1}^n u_i(\xi)\phi_i$ of $u_{\text{ex}}(\xi)$ is the solution of the linear system $A(\xi)u(\xi) = b$. The matrix $A(\xi)$ is given by $A(\xi) = (A_0 + A_9) + \sum_{i=1}^8 \xi_i A_i$, with $(A_i)_{p,q} = \int_{\Omega_i} \nabla\phi_p \nabla\phi_q$, and $b_p = \int_{\Omega_9} \phi_p$. We define the norm of V such that $\|v\|_V = \|v^h\|_{H_0^1(\Omega)}$ for any $v \in V$, where $v^h = \sum_{i=1}^n v_i\phi_i$. As a consequence, R_V is the matrix such that $(R_V)_{p,q} = \int_\Omega \nabla\phi_p \nabla\phi_q$. The quantity of interest is defined by $s : \xi \mapsto \langle l, u(\xi)\rangle$ which is the mean value of the solution over the domain Ω_1:

$$s : \xi \mapsto \langle l, u(\xi)\rangle = \frac{1}{|\Omega_1|} \int_{\Omega_1} u^h(\xi).$$

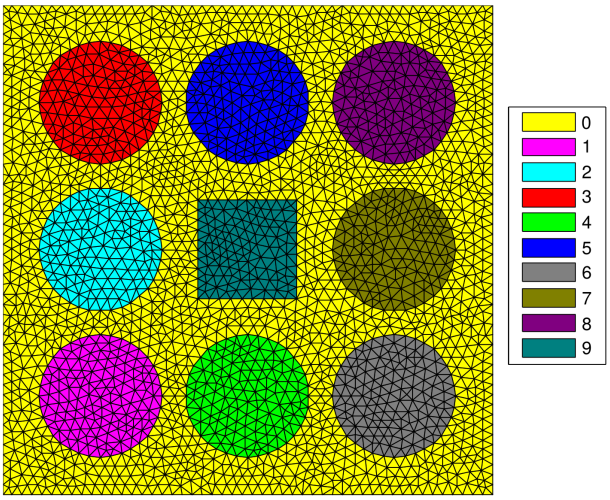


Figure 5.2: Geometry and mesh of the cookie problem.

Following the methodology presented in Section (4.3), we consider a polynomial interpolation of $u(\xi)$ on a tensor-structured interpolation grid using $p = 10$ interpolation points in each dimension, see (5.27). The tensor $\mathbf{u} \in \mathbf{X} = \mathbb{R}^n \otimes \mathbb{R}^p \otimes \ldots \otimes \mathbb{R}^p$ containing the evaluations of the solution at the interpolation points is the solution

of problem (5.28). The operator $\mathbf{A}$ is given by

$$\begin{aligned}\mathbf{A} = (A_0 + A_9) \otimes I \otimes \ldots \otimes I &+ A_1 \otimes D_1 \otimes \ldots \otimes I \\ &+ A_2 \otimes I \otimes D_2 \otimes \ldots \otimes I \\ &\ldots \\ &+ A_8 \otimes I \otimes \ldots \otimes I \otimes D_8,\end{aligned}$$

with $D_\nu = \mathrm{diag}(\xi_\nu^1, \ldots, \xi_\nu^p)$, where ξ_ν^i is the i-th interpolation point on the dimension ν. The vector $\mathbf{b}$ is given by $\mathbf{b} = b \otimes c \otimes \ldots \otimes c$, where $c \in \mathbb{R}^{10}$ is a vector with all components equal to one. The quantity of interest is then $\mathbf{Lu}$, with $\mathbf{L}$ defined by (5.30) with $\tau = \{1, \ldots, 8\}$.

We consider here low-rank approximations of $\xi \mapsto \mathbf{u}(\xi)$ in the Tensor Train format, meaning approximations with bounded t-ranks for $t \in \{\{1\}, \{1, 2\}, \ldots, \{1, \ldots, 8\}\}$, see Chapter 1, Section 3.2, Equation (27). As mentioned in Section 4.2, the goal-oriented norm is here a crossnorm, so that we can use a SVD algorithm for the practical implementation of Π_α^ε. The algorithms proposed in [107, 108] can be applied by taking into account the fact that we are using the goal-oriented norm.

5.1.2 *A posteriori* goal-oriented approximation

In order to assess the advantage of using the goal-oriented norm, we study here the *a posteriori* goal-oriented low-rank approximation of the tensor $\mathbf{u}$. We first compute an approximation $\mathbf{u}_{\mathrm{ref}}$ of $\mathbf{u}$ which is sufficiently accurate to serve as a reference solution. Such approximation is computed by Algorithm 8 (PCG) with $\varepsilon = 10^{-5}$ (using the natural norm $\|\cdot\|_X$ for the truncations), where the preconditioner $\mathbf{P}$ is defined by

$$\mathbf{P} = A(\overline{\xi})^{-1} \otimes I \otimes \ldots \otimes I, \tag{5.32}$$

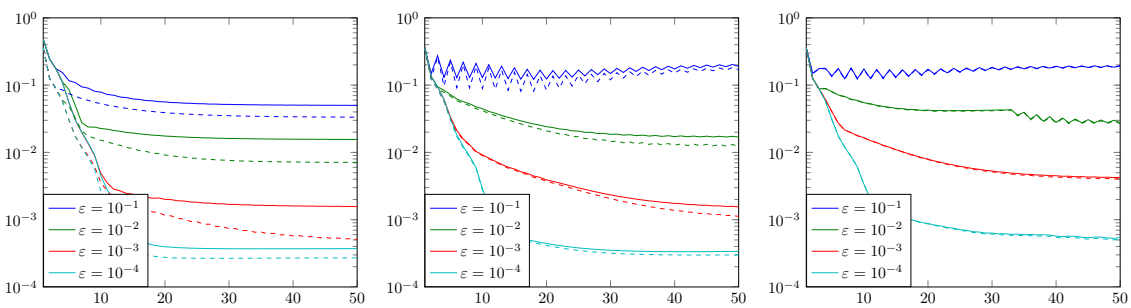
with $\overline{\xi} = \mathbb{E}(\xi) = (0.5, \ldots, 0.5)$, and I the identity matrix of size $p = 10$. Then we compute the goal-oriented low-rank approximation $\Pi_\alpha^\varepsilon(\mathbf{u}_{\mathrm{ref}})$ of $\mathbf{u}_{\mathrm{ref}}$. Table 5.2 shows the t-ranks of $\Pi_\alpha^\varepsilon(\mathbf{u}_{\mathrm{ref}})$ for different values of α and ε. Note that the ranks increase when ε decreases, which was expected since we demand more precision on the approximation. Also, we observe that for large values of α, the ranks are getting smaller, which confirms the interest of using the goal-oriented norm.

	$\varepsilon = 10^{-1}$				$\varepsilon = 10^{-2}$				$\varepsilon = 10^{-3}$				$\varepsilon = 10^{-4}$			
α	0	10	100	1000	0	10	100	1000	0	10	100	1000	0	10	100	1000
$t = \{1\}$	9	9	6	1	27	21	14	9	52	49	37	24	81	77	61	51
$t = \{1, 2\}$	8	8	5	1	25	20	14	8	49	46	34	22	73	70	60	50
$t = \{1, 2, 3\}$	7	7	4	1	19	15	11	7	39	37	28	18	64	60	54	43
$t = \{1, \ldots, 4\}$	6	6	4	1	15	12	9	6	31	29	21	15	56	52	43	36
$t = \{1, \ldots, 5\}$	5	5	3	1	11	8	6	5	23	21	15	11	49	43	33	28
$t = \{1, \ldots, 6\}$	4	4	2	1	7	6	4	4	14	14	10	8	33	27	20	17
$t = \{1, \ldots, 7\}$	3	3	2	1	5	4	3	3	8	8	6	5	15	13	10	9
$t = \{1, \ldots, 8\}$	2	2	1	1	2	2	2	2	3	3	3	3	4	4	4	3

Table 5.2: Cookie problem: t-rank of the *a posteriori* goal-oriented low-rank approximation $\Pi_\alpha^\varepsilon(\mathbf{u}_{\text{ref}})$ of $\mathbf{u}_{\text{ref}}$ for different values of α and ε.

5.1.3 Truncated iterative solver

Now, we consider the Algorithm 8 (PCG) with goal-oriented truncations. Figure 5.3 shows the convergence curves for different values of α and ε. In all cases we observe a first convergence phase, and then a plateau which corresponds approximately to the precision ε. The convergence curves are slightly deteriorated for large values of α, see for example the oscillations of the curve $\alpha = 100$ and $\varepsilon = 10^{-1}$, and the plateaus which are slightly higher when α increases. This indicates that the performances of the proposed algorithm deteriorate (moderately) for high values of α.

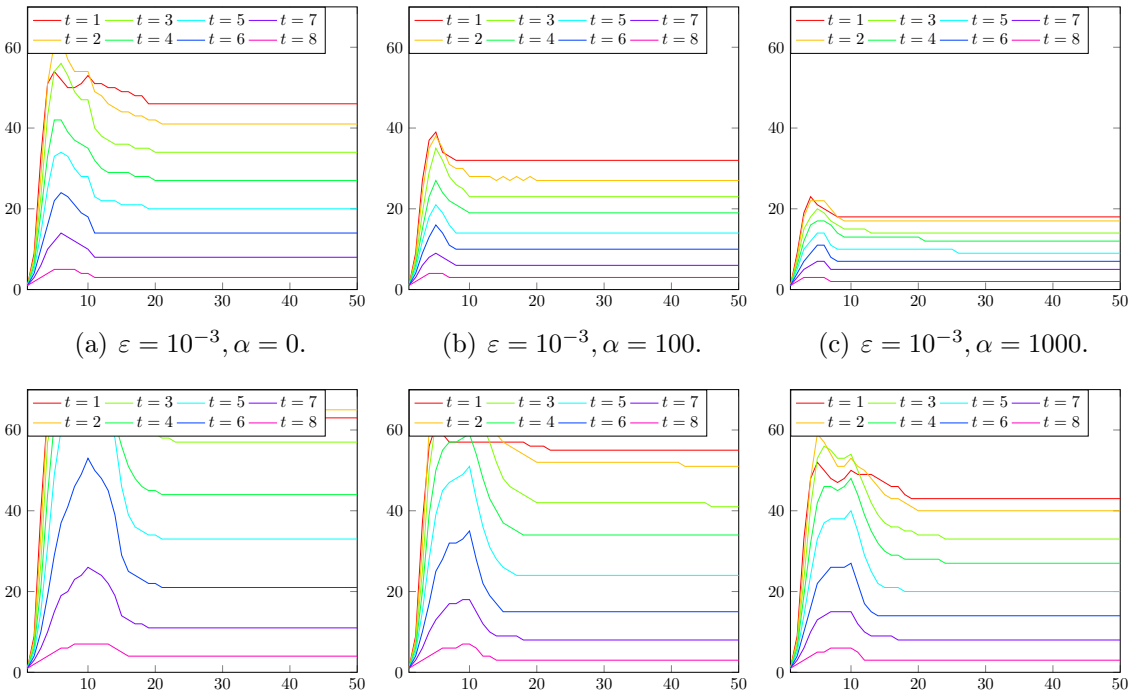


(a) $\alpha = 0$. (b) $\alpha = 100$. (c) $\alpha = 1000$.

Figure 5.3: Cookie problem: evolution of the relative errors on the solution $\|\mathbf{u}_{\text{ref}} - \mathbf{u}^k\|_{\mathbf{X}_\alpha} / \|\mathbf{u}_{\text{ref}}\|_{\mathbf{X}_\alpha}$ (continuous lines) and on the quantity of interest $\|\mathbf{L}\mathbf{u}_{\text{ref}} - \mathbf{L}\mathbf{u}^k\|_{\mathbf{Z}} / \|\mathbf{L}\mathbf{u}_{\text{ref}}\|_{\mathbf{Z}}$ (dashed lines) with respect to the iterations for different values of α and ε.

Figure 5.4 shows the t-ranks of the iterates $\mathbf{u}^k$ for different values of α with a precision $\varepsilon = 10^{-3}$ (Figures 5.4(a),5.4(b) and 5.4(c)) and $\varepsilon = 10^{-4}$ (Figures 5.4(d),5.4(e) and 5.4(f)). We observe that in the first iterations the ranks increase. Then the ranks decrease and converge. Note that for large α the ranks are significantly smaller,

which illustrates the benefits of using the goal-oriented norm.



(a) $\varepsilon = 10^{-3}, \alpha = 0$. (b) $\varepsilon = 10^{-3}, \alpha = 100$. (c) $\varepsilon = 10^{-3}, \alpha = 1000$.

(d) $\varepsilon = 10^{-4}, \alpha = 0$. (e) $\varepsilon = 10^{-4}, \alpha = 100$. (f) $\varepsilon = 10^{-4}, \alpha = 1000$.

Figure 5.4: Cookie problem: t-ranks of the iterate $\mathbf{u}^k$ during the PCG iteration process, for $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$.

5.2 Ideal minimal residual formulation

5.2.1 Benchmark Opus

We consider now the benchmark OPUS already presented in Chapter 2, Section 5.2. The parameter-dependent linear system $A(\xi)u(\xi) = b(\xi)$ of size $n = 2.8 \times 10^4$ results from the finite element discretization of an advection-diffusion equation which models the cooling of two electronic components. Four random parameters are considered: a geometrical parameter, a thermal conductance parameter, the diffusion coefficient of the components, and the amplitude of the advection field. The vector $\xi \in \mathbb{R}^4$ which contains these parameters is a random vector with independent log-uniform components. The variable of interest is here defined as the mean temperature of the two electronic components which correspond to the subdomain Ω_{IC}

(see Figure 7 of Chapter 2). We have

$$\xi \mapsto \langle l, u(\xi)\rangle = \frac{1}{\Omega_{IC}} \int_{\Omega_{IC}} u^h(\xi),$$

where $u^h(\xi) = \sum_{i=1}^n u_i(\xi)\phi_i$ is the Galerkin projection of the PDE solution on the finite element approximation space $\mathrm{span}\{\phi_1, \ldots, \phi_n\}$. We consider the conditional expectations of $\langle l, u(\xi)\rangle$ with respect to the variables $\xi_\tau = (\xi_\nu)_{\nu\in\tau}$ for $\tau \in \big\{\{1\}, \{1,2\}, \{1,2,3\}\big\}$, see (5.23). Here, we will compute these quantities of interest separately (*i.e.* one computation for each τ). Note that it is also possible to consider a unique quantity of interest which is a vector containing these three quantities of interest. However, we don't investigate this possibility here.

Similarly to the cookie problem, we consider the polynomial interpolation of $u(\xi)$ defined by (5.27) with $p = 20$ interpolation points in each dimension. The tensor $\mathbf{u}$ belongs to $\mathbf{X} = \mathbb{R}^n \otimes \mathbb{R}^p \otimes \ldots \otimes \mathbb{R}^p$, and is the solution of the algebraic equation $\mathbf{Au} = \mathbf{b}$, see (5.28). For each value of τ, the quantity of interest is then $\mathbf{Lu}$, with $\mathbf{L}$ defined by (5.30). The Riesz map $R_{\mathbf{X}}$ is defined as $R_{\mathbf{X}} = R_V \otimes R_1 \otimes \ldots \otimes R_4$, where for $\nu \in \{1, \ldots, 4\}$, R_ν is the diagonal matrix containing the weights of the interpolation polynomials associated to the dimension ν, see Section 4.3. For the sake of simplicity, R_V is the identity matrix of size n (meaning that the norm of $\mathbb{R}^n$ is the canonical norm). Note that the inverse of $R_{\mathbf{X}}$ can be easily computed, so that, thanks to (5.8), the operator $R_{\mathbf{Y}_\alpha}$ defined by (5.13) can be explicitly computed.

We consider here low-rank approximations in the following tree-based tensor subset

$$\mathcal{H}_r^T(\mathbf{X}) = \{\mathbf{x} \in \mathbf{X} : \mathrm{rank}_t(\mathbf{x}) \leq r_t,\ t \in T\}, \tag{5.33}$$

where T is the unbalanced tree given on Figure 5.5. We refer to Chapter 1, Section 3.2, for more information about this tensor format. For the approximation operators Π_α^ε and Λ_α^δ, we use a greedy rank-one algorithm with update of the core tensor. We refer to Appendix 7 for a detailed presentation of the algorithm. When using such algorithm, the t-rank of the resulting approximations is bounded by the number of greedy iterations. In the following and for simplicity, the "rank" of a tensor resulting from the approximation operators Π_α^ε or Λ_α^δ refers to its largest t-rank for $t \in T$ (which is in practice the number of greedy iterations). Finally, as mentioned in Appendix 7, the rank of the approximation is bounded by 20 for computational reasons (this is due to the update phase of the core tensor, which involves the solution of linear systems whose sizes scale in $\mathcal{O}(m^3)$ where m is the number of greedy iterations).

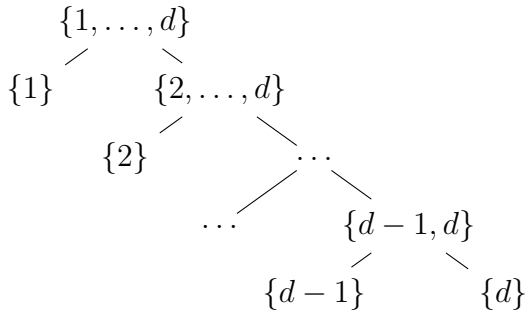


Figure 5.5: Unbalanced tree.

5.2.2 *A posteriori* goal-oriented approximation

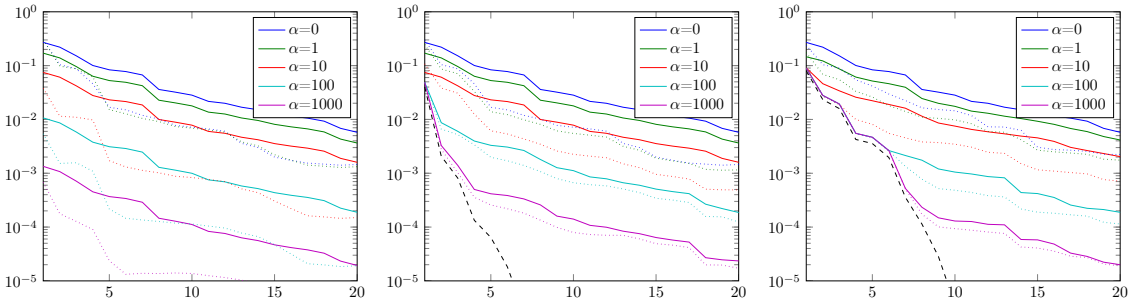
We study here the *a posteriori* goal-oriented approximation of the tensor $\mathbf{u}$. Here again, we first compute a reference solution $\mathbf{u}_{\text{ref}}$. To do that, we consider the Galerkin projection of $u(\xi)$ on a reduced space which is the span of 300 snapshots of the solution randomly selected. The error in the norm $\|\cdot\|_{\mathbf{X}}$ (estimated by a Monte Carlo method on the set of interpolation points) of this approximation is lower than 10^{-6}, which is considered sufficiently small in the present context. This model order reduction makes possible the numerical computation of this Galerkin approximation at each interpolation point with a reasonable computational time. The tensor $\mathbf{u}_{\text{ref}}$ is defined as the set of these evaluations.

Then we compute goal-oriented low-rank approximations $\Pi_\alpha^\varepsilon(\mathbf{u}_{\text{ref}})$ for different values of α and ε. The relative errors associated to these approximations are plotted on Figure 5.6. We see that the curves are shifted down when α increases. Also, the relative errors for the quantities of interest (measured with the norm $\|\cdot\|_{\mathbf{Z}}$) are comparable to relative errors for $\mathbf{u}_{\text{ref}}$ with respect to the goal-oriented norm. Besides, we observe on Figures 5.6(b) and 5.6(c) that the errors cannot be lower than a certain bound (see the dashed black curves), even for very large values of α. To explain this, let us note that when $\alpha \gg 1$, we can write

$$\begin{aligned}\min_{\mathbf{u}_r\in\mathcal{H}_r^T(\mathbf{X})} \frac{\|\mathbf{u}_{\text{ref}}-\mathbf{u}_r\|_{\mathbf{X}_\alpha}}{\|\mathbf{u}_{\text{ref}}\|_{\mathbf{X}_\alpha}} &\approx \min_{\mathbf{u}_r\in\mathcal{H}_r^T(\mathbf{X})} \frac{\|\mathbf{L}\mathbf{u}_{\text{ref}}-\mathbf{L}\mathbf{u}_r\|_{\mathbf{Z}}}{\|\mathbf{L}\mathbf{u}_{\text{ref}}\|_{\mathbf{Z}}},\\ &= \min_{\mathbf{s}_r\in\mathbf{L}\mathcal{H}_r^T(\mathbf{X})} \frac{\|\mathbf{L}\mathbf{u}_{\text{ref}}-\mathbf{s}_r\|_{\mathbf{Z}}}{\|\mathbf{L}\mathbf{u}_{\text{ref}}\|_{\mathbf{Z}}}.\end{aligned} \tag{5.34}$$

We understand that the quantity (5.34) (which does not depend on α) corresponds to the bound that we observe. To show that, we compute the low-rank approximation of $\mathbf{L}\mathbf{u} \in \mathbf{Z} = \otimes_{\nu\in\tau}\mathbb{R}^p$ in $\mathbf{L}\mathcal{H}_r^T(\mathbf{X}) = \mathcal{H}_r^T(\mathbf{Z})$ using the same greedy algorithm.

The dashed black curves of Figure 5.6 are the relative errors (measured with the norm $\|\cdot\|_{\mathbf{Z}}$) associated to these approximations. Note that for $\tau = \{1\}$, we have $\mathbf{L}\mathcal{H}_r^T(\mathbf{X}) = \mathbf{Z}$, so that the error (5.34) is zero. In other words, for any r, there exists a tensor $\mathbf{u}_r \in \mathcal{H}_r^T(\mathbf{X})$ which exactly reproduce the quantity of interest. This is why the erros given on Figure 5.6(a) are decreasing for large values of α, even for a rank-one approximation.



(a) $\tau = \{1\}$. (b) $\tau = \{1, 2\}$. (c) $\tau = \{1, 2, 3\}$.

Figure 5.6: Benchmark OPUS: *a posteriori* goal-oriented low-rank approximation of $\mathbf{u}_{\text{ref}}$. Plot of relative errors with respect to the ranks of $\Pi_\alpha^\varepsilon(\mathbf{u}_{\text{ref}})$. The Figures are associated to the quantities of interest given by (5.23) for different set τ. Continuous lines: relative error on $\mathbf{u}_{\text{ref}}$ measured with the goal-oriented norm $\|\cdot\|_{\mathbf{X}_\alpha}$. Dotted lines: relative error on $\mathbf{L}\mathbf{u}_{\text{ref}}$ measured with the norm $\|\cdot\|_{\mathbf{Z}}$. Black-dashed lines: relative error (with the norm $\|\cdot\|_{\mathbf{Z}}$) for the low-rank approximation of $\mathbf{L}\mathbf{u}_{\text{ref}}$ computed in $\mathbf{L}\mathcal{H}_r^T(\mathbf{X})$.

5.2.3 Gradient-type algorithm

We consider now the gradient-type algorithm proposed in Section 3. Here, we present the results obtained for the quantity of interest (5.23) associated to $\tau = \{1, 2\}$. Similar results have been obtained for the other sets τ.

We run the gradient-type algorithm for different values of α and ε, and the parameter δ is fixed to $\delta = 10^{-1}$. Figures 5.7(a), 5.7(b) and 5.7(c) give the relative errors for the quantity of interest. We observe a first phase of linear convergence and then a stagnation phase, as predicted by the convergence results (5.18). We note that for large values of α, the levels of stagnation correspond to the prescribed precision ε. The ranks[2] of the iterates $\{\mathbf{u}^k\}_{k\geq 1}$ are given on Figures 5.7(d), 5.7(e)

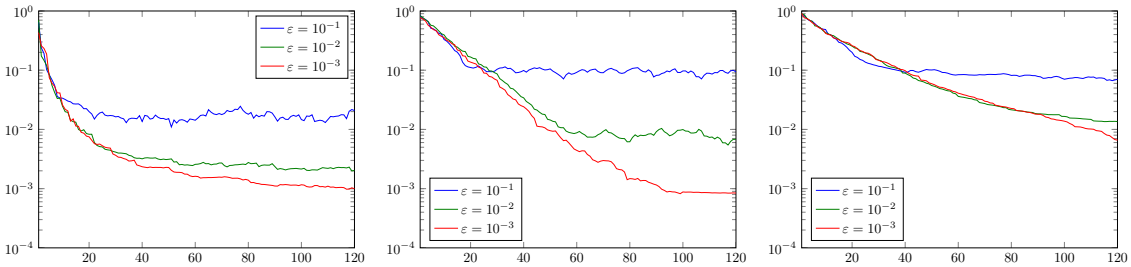
[2]We recall that the rank of $\mathbf{u}^k$ refers to $\max_{t\in T} \text{rank}_t(\mathbf{u}^k)$.

and 5.7(f). We observe that in the first iterations, the ranks are constant and equal to 20. This indicates that the relative precision ε is not reached (we recall that for computational reasons, we do not authorize ranks larger that 20). But after some iterations, the ranks are decreasing and then stabilizing around a value which (roughly) matches the one given on Figure 5.6(b). For example, with $\alpha = 10$ and $\varepsilon = 10^{-2}$, the rank of the *a priori* approximation is 10 (see Figure 5.7(e)), and the rank for which the *a posteriori* approximation achieves a relative error lower than ε is around 9 (see Figure 5.6(b)). Furthermore we observe that the ranks of the approximation (at convergence of the gradient algorithm) decreases for large values of α, see for example the curves $\varepsilon = 10^{-2}$ of Figures 5.7(d), 5.7(e) and 5.7(f).

But it is important to emphasize on the bad behavior of the gradient-type algorithm for large values of α. Indeed, we observe that the rate of convergence seriously deteriorates when α increases. This can be explained by the fact that the iterate $\mathbf{y}^k$ never reaches the relative precision δ, as it should (see Section 3.2). According to Table (5.3), the ranks of $\mathbf{y}^k$ remain constant and equal to the maximal value 20. This reflects the fact that the low-rank approximation of the residual $R_{\mathbf{Y}_\alpha}^{-1}(\mathbf{A}\mathbf{u}^k - \mathbf{b})$ (see equation (5.16)) with respect to the norm $\|\cdot\|_{\mathbf{Y}_\alpha}$ is a difficult problem for large values of α. However, even if the relative precision δ cannot be reached, we still observe a linear convergence of the gradient-type algorithm. But the corresponding rate of convergence is so small that we can not obtain the goal-oriented approximation in a reasonable computational time.

		$k=1$	$k=5$	$k=10$	$k=20$	$k=40$	$k=60$	$k=90$	$k=120$
$\alpha = 0$	$\varepsilon = 10^{-1}$	12	15	13	15	14	16	15	14
	$\varepsilon = 10^{-2}$	13	13	15	13	16	14	15	16
	$\varepsilon = 10^{-3}$	12	14	16	14	17	13	14	16
$\alpha = 10$	$\varepsilon = 10^{-1}$	20	20	20	20	18	20	20	18
	$\varepsilon = 10^{-2}$	20	20	20	19	20	20	19	20
	$\varepsilon = 10^{-3}$	20	19	20	20	19	20	20	20
$\alpha = 100$	$\varepsilon = 10^{-1}$	20	20	20	20	20	20	20	20
	$\varepsilon = 10^{-2}$	20	20	20	20	20	20	20	20
	$\varepsilon = 10^{-3}$	20	20	20	20	20	20	20	20

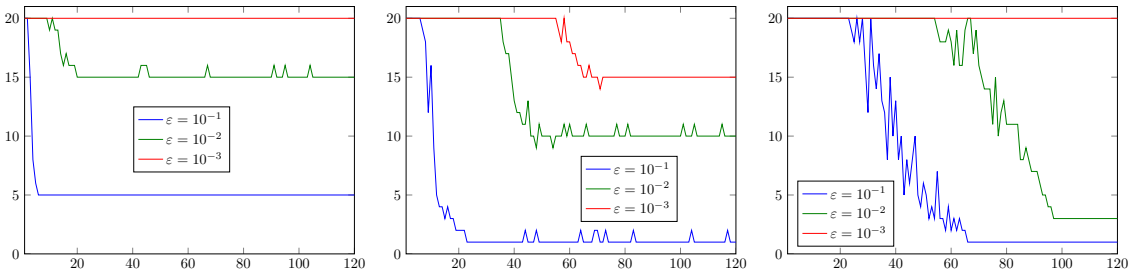
Table 5.3: Rank of the iterate $\mathbf{y}^k$ with the iterations, for $\tau = \{1, 2\}$.



(a) Error for the QoI, with $\alpha = 0$.

(b) Error for the QoI, with $\alpha = 10$.

(c) Error for the QoI, with $\alpha = 100$.



(d) Rank of $\mathbf{u}^k$, with $\alpha = 0$.

(e) Rank of $\mathbf{u}^k$, with $\alpha = 10$.

(f) Rank of $\mathbf{u}^k$, with $\alpha = 100$.

Figure 5.7: *A priori* goal-oriented low-rank approximation of $\mathbf{u}$ using the gradient-type algorithm, see Section 3. Evolution with the iterations of the relative error $\|\mathbf{Lu} - \mathbf{Lu}^k\|_{\mathbf{Z}}/\|\mathbf{Lu}\|_{\mathbf{Z}}$ (Figures 5.7(a), 5.7(b) and 5.7(c)), and of the rank of the iterate $\mathbf{u}^k$ (Figures 5.7(d), 5.7(e) and 5.7(f)), for $\tau = \{1, 2\}$.

6 Conclusion

In this chapter, we have proposed different strategies for the goal-oriented low-rank approximation of the solution of a tensor-structured equation. The basic idea is to use a goal-oriented norm for the definition of the approximation.

We first proposed an iterative solver with goal-oriented low-rank truncations of the iterates. The numerical results showed a significant reduction of the ranks of the approximation. However, this algorithm is not really designed to minimize the error with respect to the goal-oriented norm.

For the ideal minimal residual formulation, the numerical results showed that the proposed gradient-type algorithm is able to compute *a priori* the goal-oriented low-rank approximation. However, for large values of α, this algorithm becomes impractical due to its slow convergence. A possible explanation is that the problem of the low-rank approximation of the residual (see equation (5.16)) becomes ill-

conditioned when α increases. In order to understand that, let us note that thanks to relations (5.13) and (5.7), the Riesz map $R_{Y_\alpha}^{-1}$ can be written as:

$$R_{Y_\alpha}^{-1} = R_{Y_0}^{-1} + \alpha(A^{-*}L^*)R_Z(LA^{-1}).$$

Using the notation $r^k = Au^k - b$, we have

$$R_{Y_\alpha}^{-1} r^k \in R_{Y_0}^{-1} r^k + \mathrm{range}(Q),$$

where Q denotes the so-called *dual variable* defined by $Q = A^{-*}L^* \Leftrightarrow A^*Q = L^*$. Note that Q is a linear operator from Y to Z'. Therefore, for large values of α, the approximation problem $y^{k+1} \approx R_{Y_\alpha}^{-1} r^k$ involved in (5.16) corresponds somehow to the approximation of the dual variable, which can be a difficult problem.

To conclude, we showed that the use of goal-oriented norms is promising for the low-rank approximation when linear quantities of interest are considered. However, dedicated preconditioning strategies are still needed for the efficient computation of the approximation.

7 Appendix: practical implementation of the approximation operators

In this appendix, we give a possible algorithm for the practical implementation of the approximation operators Π_α^ε and Λ_α^δ. The tensor product space $H = H_1 \otimes \ldots \otimes H_d$ denotes either $\mathbf{X}$ or $\mathbf{Y}$, the norm $\|\cdot\|_H$ is the corresponding norm $\|\cdot\|_{\mathbf{X}_\alpha}$ or $\|\cdot\|_{\mathbf{Y}_\alpha}$, and e refers either to ε or to δ. We consider here approximation in the Hierarchical format $\mathcal{H}_r^T(H)$ defined by (5.33). Given $x \in H$, the goal is to compute $x_r \in \mathcal{H}_r^T(H)$ such that $\|x - x_r\|_H \leq e\|x\|_H$.

An element $x_r \in \mathcal{H}_r^T(H)$ such that $\mathrm{rank}_{\{\nu\}}(x_r) \leq r_\nu$ for all $\nu \in \{1, \ldots, d\}$, and $\mathrm{rank}_{\{1,\ldots,\nu\}}(x_r) \leq r_\nu^a$ for all $\nu \in \{1, \ldots, d-1\}$ can be written as

$$x_r = \sum_{i_1=1}^{r_1} \cdots \sum_{i_d=1}^{r_d} a_{i_1,\ldots,i_d}\, x_{i_1}^{(1)} \otimes \ldots \otimes x_{i_d}^{(d)}, \tag{5.35}$$

$$\text{with} \quad a_{i_1,\ldots,i_d} = \sum_{k_1=1}^{r_1^a} \cdots \sum_{k_{d-1}=1}^{r_{d-1}^a} a_{i_1,k_1}^{(1)} a_{k_1,i_2,k_2}^{(2)} \ldots a_{k_{d-2},i_{d-1},k_{d-1}}^{(d-1)} a_{k_{d-1},i_d}^{(d)}, \tag{5.36}$$

where $x_1^{(\nu)}, \ldots, x_{r_\nu}^{(\nu)}$ are r_ν elements of the space H_ν, and the tensor $a^{(\nu)}$ belongs to the algebraic space $\mathcal{S}_\nu = \mathbb{R}^{r_{\nu-1}^a \times r_\nu \times r_\nu^a}$ (with $r_0^a = r_d^a = 1$ by convention). The tensor

a is called the core tensor in the Tucker representation (5.35). Expression (5.36) corresponds to a representation of a in the Tensor Train format.

We propose a greedy algorithm with update of the core tensor a for the low-rank approximation in the $\mathcal{H}_r^T(H)$-format of an element $x \in H$. The idea is to build a sequence $\{x^m\}_{m\geq 0}$ of low-rank tensors such that $\|x - x^m\|_H \underset{m\to\infty}{\to} 0$, and to stop the iteration process when a relative precision $e > 0$ is achieved, meaning when $\|x - x^m\|_H \leq e\|x\|_H$. The initialization is $x^0 = 0$. At each iteration, we compute the best rank-one approximation of $x - x^{m-1}$:

$$w^m \in \arg\min_{w\in\mathcal{C}_1(H)} \|x - x^{m-1} - w\|_H,$$

where $\mathcal{C}_1(H) = \{w^{(1)} \otimes \ldots \otimes w^{(d)} : w^{(\nu)} \in H_\nu\}$. This minimization problem is solved using the Alternated Least Square (ALS) algorithm. Then we set $x^m = x^{m-1} + w^m$, which is stored in the $\mathcal{H}_r^T(H)$-format. Note that x^m linearly depends on the tensors $a^{(\nu)}$. We adopt the notation $x^m = x^m(a^{(1)}, \ldots, a^{(d)})$ for simplicity. At each iteration, the ALS Algorithm 9 is then used to improve the current approximation by optimizing the parameters $a^{(\nu)}$. The resulting algorithm is summarized in Algorithm 10. As a consequence, after m iterations, $\mathrm{rank}_t(x^m)$ is bounded by m for any $t \in T$. Then, we will say that x^m is of rank m, although it is not the tree-based rank of x^m.

Algorithm 9 Update of the core of x^m

Require: $x \in H$, $x^m = x^m(a^{(1)}, \ldots, a^{(d)}) \in \mathcal{H}_r^T(H)$, tol (in practice $tol = 10^{-3}$)

1: Initialize $stag = 1$, $x^m_{old} = x^m$, and $a^{(\nu)}_{new} = a^{(\nu)}_{old} = a^{(\nu)}$ for all $\nu \in \{1, \ldots, d\}$
2: **while** $stag \geq tol$ **do**
3: **for** $\nu = 1$ to d **do**
4: Solve

$$a^{(\nu)}_{new} \leftarrow \underset{a^{(\nu)}\in\mathcal{S}_\nu}{\mathrm{argmin}} \|x - x^m(a^{(1)}_{new}, \ldots, a^{(\nu-1)}_{new}, a^{(\nu)}, a^{(\nu+1)}_{old}, \ldots, a^{(d)}_{old})\|_H \tag{5.37}$$

5: **end for**
6: Set $x^m_{new} \leftarrow x^m(a^{(1)}_{new}, \ldots, a^{(d)}_{new})$
7: Compute $stag = \|x^m_{old} - x^m_{new}\|_H / \|x^m_{old}\|_H$
8: Set $x^m_{old} \leftarrow x^m_{new}$ and $a^{(\nu)}_{old} \leftarrow a^{(\nu)}_{new}$ for all $\nu \in \{1, \ldots, d\}$
9: **end while**
10: **return** x^m_{new}

Algorithm 10 Greedy low-rank approximation

Require: $x \in H$, e, M

Initialize $x^0 = 0$, $err = 1$, $m = 1$
while $err \geq e$ and $m \leq M$ **do**
  Compute $w^m \in \underset{w \in \mathcal{C}_1(H)}{\operatorname{argmin}} \|x - x^{m-1} - w\|_H$
  Set $x^m = x^{m-1} + w^m$
  Update the core of x^m using Algorithm 9
  Compute the error $err = \|x - x^m\|_H / \|x\|_H$
  Set $m = m + 1$
end while
return x^m

Let us note that the update of the core tensor given by Algorithm 9 requires the solution of minimization problem (5.37), which is to find $a_{new}^{(\nu)} \in \mathcal{S}_\nu$ such that

$$\langle R_H\big(x - x^m(a_{new}^{(1)}, \ldots, a_{new}^{(\nu-1)}, a_{new}^{(\nu)}, a_{old}^{(\nu+1)}, \ldots, a_{old}^{(d)})\big), \widetilde{x}\rangle = 0 \tag{5.38}$$

holds for all $\widetilde{x}$ in the vector space $\{x^m(a_{new}^{(1)}, \ldots, a_{new}^{(\nu-1)}, \widetilde{a}^{(\nu)}, a_{old}^{(\nu+1)}, \ldots, a_{old}^{(d)}),\ \widetilde{a}^{(\nu)} \in \mathcal{S}_\nu\} \subset H$. When $\nu \in \{2, \ldots, d-1\}$, this corresponds to the solution of a linear system of size $\dim(S_\nu) = m^3$, which increases rapidly with m. Then in practice, we limit the number of iterations to $m \leq M = 20$. As a consequence, the algorithm may stop before achieving the requested relative precision e.

Bibliography

[1] N. Ailon and B. Chazelle, *The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors*, SIAM J. Comput., 39 (2009), pp. 302–322. 37

[2] A. Ammar, F. Chinesta, and A. Falcó, *On the Convergence of a Greedy Rank-One Update Algorithm for a Class of Linear Systems*, Arch. Comput. Methods Eng., 17 (2010), pp. 473–486. 110, 137

[3] A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings, *A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modelling of complex fluids. Part II: Transient simulation using space-time separated representations*, J. Nonnewton. Fluid Mech., 144 (2007), pp. 98–121. 25, 110

[4] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), pp. 1–34. 36

[5] I. Babuška, F. Nobile, and R. Tempone, *A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data*, SIAM J. Numer. Anal., 45 (2007), pp. 1005–1034. 4

[6] I. Babuska, R. Tempone, and G. E. Zouraris, *Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825. 4

[7] M. Bachmayr and W. Dahmen, *Adaptive Near-Optimal Rank Tensor Approximation for High-Dimensional Operator Equations*, Found. Comput. Math., (2014). 149

[8] J. Ballani and L. Grasedyck, *A projection method to solve linear systems in tensor format*, Numer. Linear Algebr. with Appl., 20 (2013), pp. 27–43. 110, 126

[9] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, *An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations*, Comptes Rendus Math., 339 (2004), pp. 667–672. 11

[10] V. Barthelmann, E. Novak, and K. Ritter, *High dimensional polynomial interpolation on sparse grids*, Adv. Comput. Math., 12 (2000), pp. 273–288. 4

[11] C. Bekas, E. Kokiopoulou, and Y. Saad, *An estimator for the diagonal of a matrix*, Appl. Numer. Math., 57 (2007), pp. 1214–1229. 35, 36, 41

[12] R. Bellman, *Dynamic Programming*, Princeton University Press, 1957. 4

[13] R. Bellman and R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961. 4

[14] G. Berkooz, P. Holmes, and J. Lumley, *The proper orthogonal decomposition in the analysis of turbulent flows*, Annu. Rev. Fluid Mech., 25 (1993), pp. 539–575. 11

[15] M. Berveiller, B. Sudret, and M. Lemaire, *Stochastic finite element: a non intrusive approach by regression*, Eur. J. Comput. Mech. Eur. Mécanique Numérique, 15 (2006), pp. 81–92. 4

[16] G. Beylkin and M. J. Mohlenkamp, *Algorithms for numerical analysis in high dimensions*, Siam J. Sci. Comput., 26 (2005), pp. 2133–2159. 110, 137

[17] M. Billaud-Friess, A. Nouy, and O. Zahm, *A tensor approximation method based on ideal minimal residual formulations for the solution of high-dimensional problems*, ESAIM Math. Model. Numer. Anal., 48 (2014), pp. 1777–1806. 107, 158, 160, 161

[18] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Convergence Rates for Greedy Algorithms in Reduced Basis Methods*, SIAM J. Math. Anal., 43 (2011), pp. 1457–1472. 14, 50

[19] G. Blatman and B. Sudret, *Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach*, Comptes Rendus Mécanique, 336 (2008), pp. 518–523. 4, 5

[20] ——, *Adaptive sparse polynomial chaos expansion based on least angle regression*, J. Comput. Phys., 230 (2011), pp. 2345–2367. 5

[21] J. Bourgain, J. Lindenstrauss, and V. Milman, *Approximation of zonoids by zonotopes*, Acta Math., 162 (1989), pp. 73–141. 38

[22] C. Boutsidis and A. Gittens, *Improved Matrix Algorithms via the Subsampled Randomized Hadamard Transform*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1301–1340. 37

[23] S. Boyaval, C. Le Bris, Y. Maday, N. C. Nguyen, and A. T. Patera, *A reduced basis approach for variational problems with stochastic parameters: Application to heat conduction with variable Robin coefficient*, Comput. Methods Appl. Mech. Eng., 198 (2009), pp. 3187–3206. 7

[24] A. N. Brooks and T. J. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Eng., 32 (1982), pp. 199–259. 63

[25] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici, *A priori convergence of the Greedy algorithm for the parametrized reduced basis method*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 595–603. 14

[26] E. Cances, V. Ehrlacher, and T. Lelievre, *Convergence of a greedy algorithm for high-dimensional convex nonlinear problems*, Math. Model. Methods Appl. Sci., 21 (2010), p. 36. 110

[27] É. Cancès, V. Ehrlacher, and T. Lelièvre, *Greedy algorithms for high-dimensional non-symmetric linear problems*, ArXiv e-prints, (2012), p. 57. 110

[28] K. Carlberg and C. Farhat, *A low-cost, goal-oriented 'compact proper orthogonal decomposition' basis for model reduction of static systems*, Int. J. Numer. Methods Eng., 86 (2011), pp. 381–402. 12

[29] F. Casenave, A. Ern, and T. Lelièvre, *A nonintrusive reduced basis method applied to aeroacoustic simulations*, Adv. Comput. Math., (2014), p. 28. 11, 45

[30] P. Chen, *Model Order Reduction Techniques for Uncertainty Quantification Problems*, PhD thesis, 2014. 15

[31] P. Chen, A. Quarteroni, and G. Rozza, *Comparison Between Reduced Basis and Stochastic Collocation Methods for Elliptic Problems*, J. Sci. Comput., 59 (2013), pp. 187–216. 5, 54

[32] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61. 5

[33] Y. Chen, S. Gottlieb, and Y. Maday, *Parametric analytical preconditioning and its applications to the reduced collocation methods*, Comptes Rendus Math., 352 (2014), pp. 661–666. 31

[34] F. Chinesta, R. Keunings, and A. Leygue, *The proper generalized decomposition for advanced numerical simulations: a primer*, Springer Science & Business Media, 2013. 7

[35] F. Chinesta, P. Ladeveze, and E. Cueto, *A Short Review on Model Order Reduction Based on Proper Generalized Decomposition*, Arch. Comput. Methods Eng., 18 (2011), pp. 395–404. 25, 109

[36] A. Chkifa, A. Cohen, and C. Schwab, *Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs*, J. Math. Pures Appl., 1 (2014), pp. 1–29. 5

[37] ——, *High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs*, Found. Comput. Math., 14 (2014), pp. 601–633. 5

[38] A. Cohen, A. Chkifa, R. De Vore, and C. Schwab, *Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs*, ESAIM Math. Model. Numer. Anal., 47 (2013), pp. 253–280. 5

[39] A. Cohen, W. Dahmen, and G. Welper, *Adaptivity and variational stabilization for convection-diffusion equations*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 1247–1273. 111, 117, 120

[40] A. Cohen and R. Devore, *Approximation of high-dimensional parametric PDEs*, arXiv Prepr. arXiv1502.06797, (2015). 5

[41] W. Dahmen, C. Huang, C. Schwab, and G. Welper, *Adaptive Petrov–Galerkin Methods for First Order Transport Equations*, SIAM J. Numer. Anal., 50 (2012), pp. 2420–2445. 47, 111, 117

[42] W. Dahmen, C. Plesken, and G. Welper, *Double greedy algorithms: Reduced basis methods for transport dominated problems*, ESAIM Math. Model. Numer. Anal., 48 (2013), pp. 623–663. 10, 49, 55, 82, 87, 90

[43] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney, *Sampling Algorithms and Coresets for lp Regression*, SIAM J. Comput., 38 (2009), pp. 2060–2078. 38

[44] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A Multilinear Singular Value Decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278. 124

[45] V. de Silva and L.-H. Lim, *Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem*, SIAM J. Matrix Anal. Appl., 30 (2008), p. 1084. 18

[46] M. K. Deb, I. M. Babuška, and J. Oden, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Eng., 190 (2001), pp. 6359–6372. 4, 33

[47] C. Desceliers, R. Ghanem, and C. Soize, *Polynomial chaos representation of a stochastic preconditioner*, Int. J. Numer. Methods Eng., 64 (2005), pp. 618–634. 31

[48] R. DeVore and A. Kunoth, eds., *Multiscale, Nonlinear and Adaptive Approximation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 5

[49] R. DeVore, G. Petrova, and P. Wojtaszczyk, *Greedy Algorithms for Reduced Bases in Banach Spaces*, Constr. Approx., 37 (2013), pp. 455–466. 14, 50, 54

[50] R. A. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150. 5

[51] A. Doostan and G. Iaccarino, *A least-squares approximation of partial differential equations with high-dimensional random inputs*, J. Comput. Phys., 228 (2009), pp. 4332–4345. 110, 137

[52] A. Ern and J.-L. J.-L. Guermond, *Theory and Practice of Finite Elements*, vol. 159 of Applied Mathematical Sciences, Springer New York, New York, NY, 2004. 3, 113

[53] O. Ernst, C. Powell, D. Silvester, and E. Ullmann, *Efficient Solvers for a Linear Stochastic Galerkin Mixed Formulation of Diffusion Problems with Random Data*, SIAM J. Sci. Comput., 31 (2009), pp. 1424–1447. 31

[54] M. Espig and W. Hackbusch, *A regularized Newton method for the efficient approximation of tensors represented in the canonical tensor format*, Numer. Math., 122 (2012), pp. 489–525. 24, 109, 124, 125

[55] M. Espig, W. Hackbusch, and A. Khachatryan, *On the convergence of alternating least squares optimisation in tensor format representations*, Preprint, (2014). 24

[56] A. Falcó and W. Hackbusch, *On Minimal Subspaces in Tensor Representations*, Found. Comput. Math., 12 (2012), pp. 765–803. 115

[57] A. Falcó, W. Hackbusch, and A. Nouy, *Geometric structures in tensor representations*, Prepr. 9/2013, MPI MIS, (2013). 109, 114

[58] a. Falcó and A. Nouy, *A Proper Generalized Decomposition for the solution of elliptic problems in abstract form by using a functional Eckart-Young approach*, J. Math. Anal. Appl., 376 (2011), pp. 469–480. 110, 115, 124

[59] A. Falcó and A. Nouy, *Proper generalized decomposition for nonlinear convex problems in tensor Banach spaces*, Numer. Math., 121 (2012), pp. 503–530. 110, 126

[60] L. E. Figueroa and E. Süli, *Greedy Approximation of High-Dimensional Ornstein-Uhlenbeck Operators*, Found. Comput. Math., 12 (2012), pp. 573–623. 110

[61] G. Blatman, *Chaos polynomial creux et adaptatif pour la propagation d'incertitudes et l'analyse de sensibilité*, PhD thesis, Université Blaise Pascal, 2009. 5

[62] R. Ghanem and P. Spanos, *Stochastic finite elements : a spectral approach*, Springer, 1991. 4

[63] R. G. Ghanem and R. M. Kruger, *Numerical solution of spectral stochastic finite element systems*, Comput. Methods Appl. Mech. Eng., 129 (1996), pp. 289–303. 31

[64] L. Giraldi, *Contributions aux méthodes de calcul basées sur l'approximation de tenseurs et applications en mécanique numérique*, PhD thesis, 2012. 128

[65] L. Giraldi, A. Litvinenko, D. Liu, H. G. Matthies, and A. Nouy, *To Be or Not to Be Intrusive? The Solution of Parametric and Stochastic Equations—the "Plain Vanilla" Galerkin Case*, SIAM J. Sci. Comput., 36 (2014), pp. A2720–A2744. 33

[66] L. Giraldi, A. Nouy, and G. Legrain, *Low-Rank Approximate Inverse for Preconditioning Tensor-Structured Linear Systems*, SIAM J. Sci. Comput., 36 (2014), pp. A1850–A1870. 23, 33

[67] L. Giraldi, A. Nouy, G. Legrain, and P. Cartraud, *Tensor-based methods for numerical homogenization from high-resolution images*, Comput. Methods Appl. Mech. Eng., 254 (2013), pp. 154–169. 110

[68] G. H. Golub, F. T. Luk, and M. L. Overton, *A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix*, ACM Trans. Math. Softw., 7 (1981), pp. 149–169. 12

[69] G. H. Golub and C. Reinsch, *Singular value decomposition and least squares solutions*, Numer. Math., 14 (1970), pp. 403–420. 12

[70] L. González, *Orthogonal Projections of the Identity: Spectral Analysis and Applications to Approximate Inverse Preconditioning*, SIAM Rev., 48 (2006), pp. 66–75. 34, 41, 45

[71] L. Grasedyck, *Hierarchical Singular Value Decomposition of Tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054. 18, 19, 22, 124

[72] L. Grasedyck, D. Kressner, and C. Tobler, *A literature survey of low-rank tensor approximation techniques*, GAMM Mitteilungen, 36 (2013), pp. 53–78. 22, 109

[73] M. A. Grepl and A. T. Patera, *A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations*, ESAIM Math. Model. Numer. Anal., 39 (2005), pp. 157–181. 90

[74] M. J. Grote and T. Huckle, *Parallel Preconditioning with Sparse Approximate Inverses*, SIAM J. Sci. Comput., 18 (1997), pp. 838–853. 34

[75] B. Haasdonk, *Reduced Basis Methods for Parametrized PDEs – A Tutorial Introduction for Stationary and Instationary Problems*, Reduc. Order Model. Luminy B. Ser., (2014). 75, 90, 95

[76] W. Hackbusch, *Tensor spaces and numerical tensor calculus*, Springer Science and Business Media, 2012. 18, 109, 113, 124, 125

[77] ——, *l-infinity estimation of tensor truncations*, Numer. Math., 125 (2013), pp. 419–440. 7

[78] W. Hackbusch and S. Kühn, *A New Scheme for the Tensor Representation*, J. Fourier Anal. Appl., 15 (2009), pp. 706–722. 18, 19, 109, 114

[79] N. Halko, P.-G. Martinsson, and J. a. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288. 8

[80] S. Holtz, T. Rohwedder, and R. Schneider, *On manifolds of tensors of fixed TT-rank*, Numer. Math., 120 (2012), pp. 701–731. 109, 114

[81] ——, *The Alternating Linear Scheme for Tensor Optimization in the Tensor Train Format*, SIAM J. Sci. Comput., 34 (2012), pp. A683–A713. 124

[82] M. Hutchinson, *A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines*, Commun. Stat. - Simul. Comput., 19 (1990), pp. 433–450. 36

[83] D. B. P. Huynh, G. Rozza, S. Sen, and A. T. Patera, *A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants*, Comptes Rendus Math., 345 (2007), pp. 473–478. 15, 52, 89

[84] K. Kahlbacher and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems*, Discuss. Math. Differ. Inclusions, Control Optim., 27 (2007), pp. 95–117. 11

[85] B. N. Khoromskij, *Tensors-structured numerical methods in scientific computing: Survey on recent advances*, Chemom. Intell. Lab. Syst., 110 (2012), pp. 1–19. 109

[86] B. N. Khoromskij and C. Schwab, *Tensor-Structured Galerkin Approximation of Parametric and Stochastic Elliptic PDEs*, SIAM J. Sci. Comput., 33 (2011), pp. 364–385. 33, 110, 126

[87] T. G. Kolda and B. W. Bader, *Tensor Decompositions and Applications*, SIAM Rev., 51 (2009), pp. 455–500. 109

[88] A. Kolmogoroff, *Uber die beste Annaherung von Funktionen einer gegebenen Funktionenklasse*, Ann. Math., (1936), pp. 107–110. 13

[89] D. Kressner and C. Tobler, *Low-Rank Tensor Krylov Subspace Methods for Parametrized Linear Systems*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1288–1316. 110, 126, 159

[90] P. Ladevèze, *Nonlinear Conputational Structural Mechanics - New Approaches and non-Incremental Methods of Calculation*, (1999). 110

[91] P. Ladevèze, J. C. Passieux, and D. Néron, *The LATIN multiscale computational method and the Proper Generalized Decomposition*, Comput. Methods Appl. Mech. Eng., 199 (2010), pp. 1287–1296. 110

[92] L. D. E. Lathauwer, B. D. E. Moor, and J. Vandewalle, *A multilinear singular value decomposition*, 21 (2000), pp. 1253–1278. 21, 22

[93] O. P. Le Maître, O. M. Knio, H. N. Najm, and R. G. Ghanem, *Uncertainty propagation using Wiener-Haar expansions*, J. Comput. Phys., 197 (2004), pp. 28–57. 4

[94] Y. Maday, N. C. Nguyen, A. T. Patera, and G. S. H. Pau, *A general multipurpose interpolation procedure: The magic points*, Commun. Pure Appl. Anal., 8 (2009), pp. 383–404. 46

[95] Y. Maday, A. T. Patera, and G. Turinici, *A Priori Convergence Theory for Reduced-Basis Approximations of Single-Parameter Elliptic Partial Differential Equations*, J. Sci. Comput., 17 (2002), pp. 437–446. 14

[96] P. Maréchal and J. J. Ye, *Optimizing Condition Numbers*, SIAM J. Optim., 20 (2009), pp. 935–947. 31

[97] H. G. Matthies and A. Keese, *Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Eng., 194 (2005), pp. 1295–1331. 4, 33

[98] H. G. Matthies and E. Zander, *Solving stochastic systems with low-rank tensor compression*, Linear Algebra Appl., 436 (2012), pp. 3819–3838. 33, 110, 126

[99] F. Nobile, R. Tempone, and C. G. Webster, *An Anisotropic Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data*, SIAM J. Numer. Anal., 46 (2008), pp. 2411–2442. 5

[100] A. Nouy, *A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations*, Comput. Methods Appl. Mech. Eng., 196 (2007), pp. 4521–4537. 110

[101] ——, *Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations*, Arch. Comput. Methods Eng., 16 (2009), pp. 251–285. 25, 33

[102] ——, *A priori model reduction through Proper Generalized Decomposition for solving time-dependent partial differential equations*, Comput. Methods Appl. Mech. Eng., 199 (2010), pp. 1603–1626. 25, 110

[103] ——, *Proper Generalized Decompositions and Separated Representations for the Numerical Solution of High Dimensional Stochastic Problems*, Arch. Comput. Methods Eng., 17 (2010), pp. 403–434. 110

[104] E. Novak, H. Woźniakowski, and P. C. Newport, *Tractability of Multivariate Problems: Standard information for functionals*, vol. 12, European Mathematical Society, 2010. 4

[105] I. Oseledets and E. Tyrtyshnikov, *Tensor tree decomposition does not need a tree*, Submitt. to Linear Algebr. Appl, (2009). 20

[106] I. V. Oseledets, *Tensor-Train Decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317. 20, 109, 114

[107] I. V. Oseledets and E. E. Tyrtyshnikov, *Breaking the Curse of Dimensionality, Or How to Use SVD in Many Dimensions*, SIAM J. Sci. Comput., 31 (2009), pp. 3744–3759. 22, 170

[108] ——, *Breaking the Curse of Dimensionality, Or How to Use SVD in Many Dimensions*, SIAM J. Sci. Comput., 31 (2009), pp. 3744–3759. 124, 170

[109] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition*, in Signals, Syst. Comput. 1993. 1993 Conf. Rec. Twenty-Seventh Asilomar Conf., IEEE, 1993, pp. 40–44. 5

[110] N. a. Pierce and M. B. Giles, *Adjoint Recovery of Superconvergent Functionals from PDE Approximations*, SIAM Rev., 42 (2000), pp. 247–264. 75, 79

[111] C. Prud'homme, D. V. Rovas, K. Veroy, L. Machiels, Y. Maday, a. T. Patera, and G. Turinici, *Reliable Real-Time Solution of Parametrized Partial Differential Equations: Reduced-Basis Output Bound Methods*, J. Fluids Eng., 124 (2002), p. 70. 87

[112] A. Quarteroni, G. Rozza, and A. Manzoni, *Certified reduced basis approximation for parametrized partial differential equations and applications*, J. Math. Ind., 1 (2011), p. 3. 90

[113] T. Rohwedder and A. Uschmajew, *On Local Convergence of Alternating Schemes for Optimization of Convex Problems in the Tensor Train Format*, SIAM J. Numer. Anal., 51 (2013), pp. 1134–1162. 24, 109, 125

[114] G. Rozza, D. B. P. Huynh, and A. T. Patera, *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 Eng., 15 (2008), pp. 229–275. 14, 49, 62, 78, 90

[115] G. Rozza and K. Veroy, *On the stability of the reduced basis method for Stokes equations in parametrized domains*, Comput. Methods Appl. Mech. Eng., 196 (2007), pp. 1244–1260. 87

[116] Y. Saad and M. H. Schultz, *GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869. 66

[117] I. M. Sobol', *On sensitivity estimation for nonlinear mathematical models*, Mat. Model., 2 (1990), pp. 112–118. 164

[118] P. Sonneveld, *CGS, A Fast Lanczos-Type Solver for Nonsymmetric Linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52. 66

[119] V. N. Temlyakov, *Greedy approximation*, Acta Numer., 17 (2008), pp. 235–409. 25, 127, 128, 130, 131, 133, 134

[120] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal., 03 (2010), p. 8. 37

[121] A. Uschmajew, *Local Convergence of the Alternating Least Squares Algorithm for Canonical Tensor Approximation*, 2012. 24

[122] A. Uschmajew and B. Vandereycken, *The geometry of algorithms using hierarchical tensors*, Linear Algebra Appl., 439 (2013), pp. 133–166. 109, 125

[123] K. Veroy, C. Prud'homme, D. V. Rovas, A. T. Patera, and C. Prud'homme, *A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations*, Proc. 16th AIAA Comput. fluid Dyn. Conf., 3847 (2003), pp. 1–18. 13, 14, 49

[124] X. Wan and G. E. Karniadakis, *An adaptive multi-element generalized polynomial chaos method for stochastic differential equations*, J. Comput. Phys., 209 (2005), pp. 617–642. 4

[125] L. Welch, *Lower bounds on the maximum cross correlation of signals (Corresp.)*, IEEE Trans. Inf. Theory, 20 (1974), pp. 397–399. 36

[126] D. Xiu and J. S. Hesthaven, *High-Order Collocation Methods for Differential Equations with Random Inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139. 4

[127] O. Zahm and A. Nouy, *Interpolation of inverse operators for preconditioning parameter-dependent equations*, arXiv Prepr. arXiv1504.07903, (2015), pp. 1–37. 29, 75, 86, 90, 100, 101