Criticality calculations in neutronics : model order reduction and a posteriori error estimators

Yonah Conjungo Taumhas

École des Ponts ParisTech

THÈSE DE DOCTORAT
de l'École des Ponts ParisTech

Criticality calculations in neutronics: model order reduction and a posteriori error estimators

École doctorale N°532, Mathématiques et STIC (MSTIC)

Spécialité: Mathématiques appliquées

Thèse préparée au Service d'Études des Réacteurs et de Mathématiques Appliquées, CEA Paris-Saclay

Thèse soutenue le 08 décembre 2023, par
Yonah CONJUNGO TAUMHAS

Composition du jury:

Yvon MADAY Professeur des universités, Sorbonne Université	*Président du jury*
Bruno DESPRÉS Professeur des universités, Sorbonne Université	*Rapporteur*
Agnieszka MIĘDLAR Associate Professor, Virginia Tech	*Rapporteure*
Olga MULA Associate Professor, Eindhoven University of Technology	*Examinatrice*
Grégoire ALLAIRE Professeur, École Polytechnique	*Examinateur*
Virginie EHRLACHER Chargée de recherche, École des Ponts ParisTech	*Examinatrice*
Tony LELIÈVRE Professeur, École des Ponts ParisTech	*Directeur de thèse*
François MADIOT Ingénieur de recherche, CEA Paris-Saclay	*Encadrant de thèse*

Abstract

For some applications involving loading optimization for nuclear reactor cores, such as irradiation experiments on Materials Testing Reactors (MTRs), or optimizing fuel assembly loading patterns, the challenge is to reduce the computational time of the simulation of the reactor state, while controlling calculation biases and errors. A reduced-basis approach is a natural candidate to meet this constraint.

In the context of reduced bases, we build an approximation space associated with a partial differential equation that depends on parameters encoding the loading pattern. The construction of this approximation space involves a phase of exploration of the parameter space, in which it is important to quantify the error between the solution obtained from the approximation space (under construction) and the solution obtained with a standard high-fidelity calculation (fine discretization). This crucial step enables the certification of the reduced basis construction via *a posteriori* error estimates. Recent works have been carried out to obtain a computable error estimate for symmetric eigenvalue problems. In the context of neutronics, we are interested in generalized non-symmetric eigenvalue problems (criticality calculations).

Thus, in this work, we extend the *a posteriori* estimation for eigenvalue problems to the non-symmetric case. Then, a reduced basis method, based on the greedy algorithm and using the *a posteriori* error estimates that were developed is first implemented in a mock-up code, in order to validate and certify the method through simple and illustrative test cases. Then, the implementation of such a reduced-order model is carried out in the APOLLO3®code, developed at CEA.

Keywords: Model Order Reduction (ROM), reduced basis method, non-symmetric eigenvalue problems, criticality, *a posteriori* error estimation

Résumé

Titre de la thèse en français: Calculs critiques en neutronique: réductions de modèles et estimateurs d'erreur *a posteriori*.

Pour certaines applications concernant l'optimisation du chargement du coeur d'un réacteur nucléaire, telles que les expériences sur les réacteurs d'irradiation technologique (*Materials Testing Reactors*), ou l'optimisation du placement des assemblages, une difficulté majeure est de pouvoir réduire le temps de calcul de l'état du réacteur, tout en maîtrisant les biais et les erreurs de calcul. Une approche de type "bases réduites" est un candidat naturel pour répondre à cette contrainte.

Dans le cadre des bases réduites, nous construisons un espace d'approximation associé à une équation aux dérivées partielles dépendant de paramètres qui encodent le chargement du coeur. La construction de cet espace d'approximation comporte une phase d'exploration de l'espace des paramètres dans laquelle il est important de quantifier l'erreur entre la solution obtenue à partir de l'espace d'approximation (en construction) et la solution obtenue avec un calcul standard (discrétisation fine). Cette étape cruciale permet de certifier la construction de la base réduite via des estimateurs d'erreur *a posteriori*. Récemment, des travaux ont été menés pour obtenir un estimateur d'erreur calculable sur des problèmes aux valeurs propres symétriques. Dans le contexte de la neutronique, on s'intéresse à des problèmes aux valeurs propres généralisés non symétriques (problèmes de criticité).

Par conséquent, dans cette étude, nous généralisons les estimateurs d'erreur *a posteriori* pour les problèmes aux valeurs propres au cas non symétrique. Ainsi, une méthode des bases réduites, reposant sur l'algorithme *greedy* et utilisant les estimateurs d'erreur *a posteriori* ayant été développés, est d'abord implémentée dans une maquette, et permet de valider et certifier le modèle réduit obtenu sur cas tests simples et illustratifs. Puis, nous présentons l'implémentation d'un tel modèle réduit dans le code APOLLO3®, développé au CEA.

Mots-clés: Réduction de modèles, méthode des bases réduites, problèmes aux valeurs propres non symétriques, criticité, estimation *a posteriori*

Acknowledgments

First of all, I would like to express my gratitude to all of you who have helped me, supported me, and contributed, directly or indirectly, to my thesis, through this very specific doctoral process. We have come a long way together, and none of this work would have been achieved without you all.

To begin with, I would like to thank warmly my thesis director, Tony Lelièvre, and my thesis supervisor, François Madiot, for trusting me, for their unwavering guidance, for their admirable resilience, and for all our fruitful discussions. It was a pleasure working with you. Many thanks to Virginie Ehrlacher and Geneviève Dusson, as well, for their precious advice, uninterrupted commitment and dedication to this work, from beginning to end.

I thank Agnieszka Międ lar and Bruno Després for accepting to be reviewers and for showing great interest in my work. Let me also thank the thesis examiners, Grégoire Allaire, Olga Mula and Yvon Maday, for accepting to be part of my thesis jury.

I cannot help thinking about all the teachers, professors and tutors that I have admired and looked up to along the path, particularly Guillaume Leturgez, Jean-Yves Boyer and Andrés Medaglia. I also dedicate this thesis to you.

Special thanks go to all the people that I met at CEMRACS 2021 summer school, during the first year of my PhD. I thank Tommaso Taddei and Olga Mula, once again, for initiating me into co-working and co-publishing for this very interesting project we worked on together. Thank you Sarah and Elham, I am so grateful for meeting you at this time, and I am proud to call you now my friends.

I would like to thank CEA and SERMA for trusting me, for providing funding for my research project, and also for giving me the opportunity to meet wonderful people and colleagues I will never forget. I think particularly of Francesco, Andrea, David and Sabah. You have helped me grow over the past three years as a professional, but also as a human being.

Last but not least, I would like to thank all my close friends and family. I will never forget these summer weeks spent with my mom while I was writing my thesis, who did the best she could so that this process would feel much less overwhelming and stressful for me, even though my work was all Greek to her. Also, to all my friends, from Paris to Southwest France, that I have met throughout my academic life and that I have known for a while now, thank you so much, especially to my dear friend Ruben, who always has been there to cheer me up through difficult times of doubt and loss of confidence.

I will finish here with expressing my deepest gratitude to my travel and life partner, Santiago, without whom none of this would have been possible. Thank you for giving me the strength to keep growing every day by your side.

Contents

Introduction

In the field of nuclear reactors, Materials Testing Reactors (MTRs) are research reactors which aim at carrying out experiments on materials or fuel elements of power reactors, as it is is notably the case for reactors exploited at EDF (*Électricité de France*), at the end of nuclear fuel cycles. At CEA, it is the case of the OSIRIS reactor at CEA/Saclay (shut down in 2015), as well as the Jules Horowitz reactor (RJH), still under construction at CEA/Cadarache. These nuclear reactors have the ability to run several irradiation experiments inside the nuclear core, or at the scale of the neutron reflector, and ensure, as well, the radioisotope production for medical purposes, especially the Technetium-99m (99mTc). Nevertheless, such reactors present numerous heterogeneities. The acute management of fuel elements using enriched uranium (20%), as well as the respect of the requirements for the different irradiation experiments are major issues in the optimal use of fuel elements in the reactor. The main challenge of these experiments is to guarantee the expected performance by minimizing the fuel consumption while meeting all safety requirements.

More generally, the operation of both research reactors and EPRs consists of similar experiments, which notably introduce the loading pattern optimization problem, and consists of studying criticality inside the core, which amounts to solving a non-symmetric eigenvalue problem. The resolution of this high-fidelity problem comes with a certain computational cost, which is high when it comes to optimization problems, like the loading pattern optimization problem. Such a multi-query problem indeed requires to solve in many high-fidelity computations associated with different core configurations. To perform these computations, there exists various codes at CEA, such as APOLLO3®. This deterministic neutron transport code provides functionalities and advanced methods which enable the bias reduction without penalizing the computational times, in comparison with the two-step calculation scheme provided by second-generation codes, such as the APOLLO2 transport code with the CRONOS2 core calculation code. Undergoing works at CEA aim at developing a "best-estimate" calculation scheme using the APOLLO3® functionalities, in order to support neutron studies of first loading cores for RJH.

However, such an advanced calculation scheme in APOLLO3® will not meet the need of an operating tool for the fuel and the core irradiation, in the context of real-time monitoring or irradiation campaigns based on the different fuel irradiation states. For a given evaluation scheme, the main challenge is to run inexpensive calculations while preserving or monitoring bias and calculation errors. Indeed, the key element is to run inexpensive loading core calculations, while being able to update calculations in the case of hazards during exploitation. For example, an unexpected removal of an experiment may occur during the cycle. In the case of the OSIRIS reactor, one fuel cycle calculation takes a few minutes.

In this context, a reduced-basis approach is proposed. It consists of the development and implementation of a Reduced-Order Model (ROM) for criticality calculations in neutronics, via the development and use of error estimates, based on an *a posteriori* analysis for non-symmetric generalized eigenvalue problems, which allow to quantify the approximation error.

The outline of this manuscript reads as follows. In Chapter 1, we recall and describe the standard high-fidelity discretization techniques for core calculation and criticality

problems in neutronics. Chapter 2 then aims at establishing computable and inexpensive *a posteriori* error estimates for non-symmetric generalized eigenvalue problems. It notably reminds the state-of-the-art error estimates that were developed in the symmetric case and the link between error estimation and the spectral gap for eigenvalue problems. An *a priori* error analysis apprehends the construction of a reduced-order model in the non-symmetric case via a Galerkin projection method and indicates how important it is to consider both left and right eigenvectors in our approach. Then, the key difficulty is to propose reliable, efficient and computable error estimates. To do so, we develop residual-based error estimates which all exhibit multiplicative parameter-dependent *prefactors* in the error upper bounds. The deployment of a heuristic approach enables the estimation of the prefactors, as they are not computable in practice, but hold key information in the error behavior. Afterwards, in Chapter 3, an efficient implementation of a reduced-basis method, using the *a posteriori* error estimates developed in the previous chapter, is detailed, in the case of an affine decomposition of the high-fidelity matrices with respect to their parameter dependency. Based on a greedy algorithm, it consists of a two-step *offline/online* procedure and a Galerkin projection of the high-fidelity problem on a well-chosen reduced space. Chapter 4 illustrates the achievements of such a reduced-basis method on two-group neutron diffusion mock-up codes through several numerical tests. A first test case on a non-physical small nuclear core highlights the necessity of considering the whole upper bound in the error estimation for the certification of the reduced-order model. The second test case, namely the *Minicore*, shows to what extent the model order reduction provides a reliable model in very small computational times. At last, the third test case numerically explains the rationale behind the choice of including both direct and adjoint eigenvectors in the error estimation. Then, Chapter 5 introduces the preliminary implementation of the reduced-basis method in the APOLLO3® code. While, at this stage, the complexity of the Galerkin projection step dominates the computational cost of a greedy procedure, a Proper Orthogonal Decomposition (POD) approach is proposed in this context, and provides promising results, notably with the use of the error estimates that were developed and used in the previous chapters. Finally, in Chapter 6, we give a direct industrial application of a POD-type reduced-order model in the context of state estimation and data assimilation for criticality calculations. Note that this last chapter is a published proceeding of the CEMRACS 2021 research session.

In summary, our main contributions are as follows:

- the development of *a posteriori* error estimates for non-symmetric eigenvalue problems;
- the analysis of the parameter-dependency of the so-called *prefactors* in the error bounds with respect to the spectral norm, and the development of numerical methods taking these prefactors into account in the implementation of the residual-based error estimates;
- the efficient construction of a reduced-order model for non-symmetric eigenvalue problems, via a greedy algorithm and using *a posteriori* error estimates in order to select specific basis functions to add in the reduced basis, under the assumption of parametric affine decomposition of the high-fidelity matrices;
- the pioneering implementation of a reduced-basis method and associated error estimates in the APOLLO3® neutron code for industrial purposes.

Chapter 1

An overview of classical discretization techniques for the steady-state Boltzmann equation

This chapter was written based on the following references:

- [40] M. Coste-Delclaux, C. Diop, A. Nicolas, and B. Bonin, Neutronique, CEA Saclay; Groupe Moniteur, 2013.
- [97] O. Mula, Some contributions towards the parallel simulation of time dependent neutron transport and the integration of observed data in real time, PhD thesis, Paris VI, 2014.
- [62] L. Giret, Numerical analysis of a non-conforming domain decomposition for the multigroup SPN equations, PhD thesis, Université Paris-Saclay (ComUE), 2018.
- [78] D. Labeurthre, Development and comparison of high-order finite element bases for solving the transport equation on hexagonal meshes, PhD thesis, Université Grenoble Alpes, 2022.

We start with a general overview of discretization techniques for the Boltzmann equation. In Section 1.1, we first recall the time-dependent neutron transport equation, which gives a general model for neutronics dynamics in a nuclear reactor core. In Section 1.2, we introduce the criticality problem, which can be derived from the stationary neutron transport equation, and defines the generalized eigenvalue problem of interest in this work. In Section 1.3, some standard discretization techniques of the continuous problem are recalled. Finally, in Section 1.4, we motivate the idea of approximating the neutron transport model, in particular by the neutron diffusion model.

1.1 The neutron transport equation

1.1.1 The phase space

In neutronics, the Boltzmann equation is used to describe the neutron population dynamics in a nuclear reactor core $\mathcal{R}$. The solution to the Boltzmann equation is described over the *phase space*, namely the position and the velocity. In neutronics, it is common to represent the velocity $\vec{v}$ by the pair $(\vec{\omega}, E)$, where $\vec{\omega}$ is the direction and E is the energy.

Note that the velocity $\vec{v}$ of a neutron of mass m is totally determined by its direction $\vec{\omega}$ and its energy E, as $\vec{\omega} = \vec{v}/|\vec{v}|$ and $E = m|\vec{v}|^2/2$. Hence, we consider the following four variables:

- the time variable $t \in [0, T]$, where $T > 0$ is some characteristic time;
- the space variable $r \in \mathcal{R}$; we assume that $\mathcal{R}$ is a bounded, connected and open subset of $\mathbb{R}^3$, with a piecewise regular Lipschitz boundary $\partial\mathcal{R}$;
- the angular variable, or direction, $\vec{\omega} \in \mathbb{S}_2$, where $\mathbb{S}_2$ stands for the unit sphere, which indicates the direction of the neutron;
- the energy $E \in [E_{\min}, E_{\max}]$ of the neutron, with $0 < E_{\min} < E_{\max}$.

The generic observed data is the total neutron density $n\,(t, r, \vec{\omega}, E)$ of velocity distribution $\vec{v} \in \mathcal{V}$, inside a reactor core, in the phase space $\mathcal{D} = \mathcal{R} \times \mathcal{V}$, with $\mathcal{V} = \mathbb{S}_2 \times [E_{\min}, E_{\max}]$ at time $t \in [0, T]$. In fact, the neutron population inside the studied system is totally described by the neutron angular flux defined as

$$\psi\,(t, r, \vec{\omega}, E) = n\,(t, r, \vec{\omega}, E)\,|\vec{v}|, \tag{1.1}$$

associated with a **transport** problem. At the core scale, the transport problem may be approximated by a **diffusion** problem. We further discuss this problem in Section 1.4.

1.1.2 The time-dependent neutron transport equation

The evolution of the neutron population inside a reactor core is described by the Boltzmann equation, i.e., a balance between the neutrons that disappear and the neutrons that are created in the nuclear core $\mathcal{R}$. The neutron transport equation over $[0, T] \times \mathcal{D}$ can be written as

$$\begin{cases} \dfrac{1}{|\vec{v}|}\partial_t\psi\,(t, r, \vec{\omega}, E) + \mathbb{L}\psi\,(t, r, \vec{\omega}, E) = \mathbb{H}\psi\,(t, r, \vec{\omega}, E) + \mathbb{F}_p\psi\,(t, r, \vec{\omega}, E) + \displaystyle\sum_{l=1}^{L} \mathbb{F}_{d,l} C_l(t, r) \\ \partial_t C_l(t, r) + \lambda_l C_l(t, r) = \displaystyle\int_{E_{\min}}^{E_{\max}} \beta_l(t, r, E')(\nu\Sigma_f)(t, r, E')\phi\,(t, r, E')\,dE', \quad \forall l \in [\![1, L]\!], \\ \psi\,(t = 0, r, \vec{\omega}, E) = \psi^0(r, \vec{\omega}, E), \end{cases} \tag{1.2}$$

introducing the operators of

- **advection**: it describes the pure transport inside the system, as well as the loss of neutrons from interaction with any nucleus. It is then linked to the probability of interaction between neutrons, modeled by the macroscopic total cross section Σ_t. The advection operator is defined by

$$\mathbb{L}\psi\,(t, r, \vec{\omega}, E) = \vec{\omega}.\nabla\psi\,(t, r, \vec{\omega}, E) + \Sigma_t(t, r, E)\psi\,(t, r, \vec{\omega}, E)\,; \tag{1.3}$$

- **scattering**: it describes the collisions that neutrons undergo, implying changes of direction and/or energy. The probability of a neutron to transition from a direction $\vec{\omega'}$ and an energy E' to a direction $\vec{\omega}$ and an energy E by collision is modeled by the macroscopic scattering cross section Σ_s. The scattering operator is defined by

$$\mathbb{H}\psi\,(t, r, \vec{\omega}, E) = \int_{E_{\min}}^{E_{\max}} \int_{\mathbb{S}_2} \Sigma_s\left(t, r, (\vec{\omega'}, E') \to (\vec{\omega}, E)\right) \psi\left(t, r, \vec{\omega'}, E'\right) d\vec{\omega'} dE'; \tag{1.4}$$

- **prompt fission**: it describes the creation of new neutrons among the population by fission of a heavier particle, which is likely to take place under some probability of fission modeled by the macroscopic fission cross section Σ_f. The average number of neutrons created by this reaction is given by $\nu(t,r,E)$, and we denote by $\chi_p(t,r,E)$ the prompt spectrum. The prompt fission operator is defined by

$$\mathbb{F}_p\psi\left(t,r,\vec{\omega},E\right) = \frac{\chi_p(t,r,E)}{4\pi}\int_{E_{\min}}^{E_{\max}}(1-\beta(t,r,E'))(\nu\Sigma_f)(t,r,E')\int_{\mathbb{S}_2}\psi\left(t,r,\vec{\omega},E'\right)d\vec{\omega}dE' \tag{1.5}$$

$$= \frac{\chi_p(t,r,E)}{4\pi}\int_{E_{\min}}^{E_{\max}}(1-\beta(t,r,E'))(\nu\Sigma_f)(t,r,E')\phi\left(t,r,E'\right)dE',$$

  where $\beta(t,r,E) = \sum_{l=1}^{L}\beta_l(t,r,E)$ is the total delayed neutron fraction and

$$\phi\left(t,r,E\right) = \int_{\mathbb{S}_2}\psi\left(t,r,\vec{\omega},E\right)d\vec{\omega}, \tag{1.6}$$

  is the neutron scalar flux;

- **delayed fission**: it describes the creation by fission of radioactive isotopes, namely the precursors. Each precursor $l \in \{1,\dots,L\}$ is characterized by its radioactive decay constant λ_l, delayed neutron fraction $\beta_l(t,r,E)$, delayed fission spectrum $\chi_{d,l}(t,r,E)$ and concentration $C_l(t,r)$. The delayed fission operator for the precursor l writes

$$\mathbb{F}_{d,l}C_l(t,r) = \frac{\lambda_l}{4\pi}\chi_{d,l}(t,r,E)C_l(t,r). \tag{1.7}$$

We provide the Boltzmann equation with vacuum boundary conditions

$$\psi\left(t,r,\vec{\omega},E\right) = 0, \quad \text{on } \Gamma^-, \tag{1.8}$$

with

$$\Gamma^- = \left\{(r,\vec{\omega},E) \in \partial\mathcal{R}\times\mathcal{V},\, \vec{\omega}.\vec{n}(r) < 0\right\},$$

where $\vec{n}$ is the outward unit normal vector to the boundary of the core $\partial\mathcal{R}$. We refer to Section 1.1.3 of [97] and Chapter XXI, Section 3.1 of [42] to discuss the existence, uniqueness and positivity of Problem (1.2) with boundary conditions (1.8). This is the type of boundary conditions that we take into account throughout our studies. Among other boundary conditions for Problem (1.2), we find in the literature:

- non-homogeneous boundary conditions: it is particularly the case of an incoming angular flux ψ_{in}, i.e.,

$$\psi\left(t,r,\vec{\omega},E\right) = \psi_{\text{in}}\left(t,r,\vec{\omega},E\right), \quad \forall(r,\vec{\omega},E) \in \Gamma^-;$$

- reflective boundary conditions:

$$\psi\left(t,r,\vec{\omega},E\right) = \psi\left(t,r,\vec{\omega'},E'\right), \text{ with } \vec{\omega'} = \vec{\omega} - 2(\vec{\omega}.\vec{n})\vec{n}, \quad \forall(r,\vec{\omega},E) \in \Gamma^-;$$

- *albedo* boundary conditions:

$$\psi\left(t,r,\vec{\omega},E\right)=\int_{t'=0}^{t}\int_{\Gamma^+}\beta_0(t',r',\vec{\omega'},E',t,r,\vec{\omega},E)\psi\left(t',r',\vec{\omega'},E'\right)d\Gamma dt',\quad\forall(r,\vec{\omega},E)\in\Gamma^-,$$

  where the quantity $\beta_0(t',r',\vec{\omega'},E',t,r,\vec{\omega},E)$ is related to the flux that, at time t, enters the domain $\mathcal{R}$ at $r\in\partial\mathcal{R}$ with velocity $(\vec{\omega},E)$ as a result of the interaction of a unit flux of particles with the external media, and, at time t', exits the domain at $(r',\vec{\omega'},E')\in\Gamma^+$, with $\Gamma^+=\{(r,\vec{\omega},E)\in\partial\mathcal{R}\times\mathcal{V},\ \vec{\omega}.\vec{n}(r)>0\}$;

- periodic boundary conditions: for example for $\mathcal{R}=[0,L]^3$, with $L>0$,

$$\psi\left(t,r,\vec{\omega},E\right)=\psi\left(t,r+L\vec{e_i},\vec{\omega},E\right),\quad\forall(r,\vec{\omega},E)\in\Gamma^-,\quad\forall i=\{1,2,3\},$$

  where, $(\vec{e_i})_{i=1,2,3}$ is the canonical basis of $\mathbb{R}^3$.

1.2 The steady-state neutron transport equation: a generalized eigenvalue problem

The steady-state neutron transport equation may be formulated in two different ways as the k-, or α-eigenproblem [15]. The α-eigenproblem originates from the Laplace transform of the time-dependent neutron transport equation (1.2). It yields the following eigenvalue problem:

Find (ψ,α) such that

$$\begin{cases}\dfrac{\alpha}{|\vec{v}|}\psi\left(r,\vec{\omega},E\right)+\left(\mathbb{L}_0-\mathbb{H}_0\right)\psi\left(r,\vec{\omega},E\right)=\mathbb{F}_{p,0}\psi\left(r,\vec{\omega},E\right)+\displaystyle\sum_{l=1}^{L}\frac{\lambda_j}{\lambda_j+\alpha}\mathbb{F}_{d,l,0}\phi(r,E),\quad\text{in }\mathcal{D}\\ \psi\left(r,\vec{\omega},E\right)=0,\quad\text{on }\Gamma^-,\end{cases}\tag{1.9}$$

where

$$\mathbb{L}_0\psi\left(r,\vec{\omega},E\right):=\vec{\omega}.\nabla\psi\left(r,\vec{\omega},E\right)+\Sigma_t(r,E)\psi\left(r,\vec{\omega},E\right),\tag{1.10}$$

$$\mathbb{H}_0\psi\left(r,\vec{\omega},E\right):=\int_{E_{\min}}^{E_{\max}}\int_{\mathbb{S}_2}\Sigma_s\left(r,(\vec{\omega'},E')\to(\vec{\omega},E)\right)\psi\left(r,\vec{\omega'},E'\right)d\vec{\omega'}dE',\tag{1.11}$$

$$\mathbb{F}_{p,0}\psi\left(r,\vec{\omega},E\right):=\frac{\chi_p(r,E)}{4\pi}\int_{E_{\min}}^{E_{\max}}(1-\beta(r,E'))(\nu\Sigma_f)(r,E')\phi\left(r,E'\right)dE',\tag{1.12}$$

$$\mathbb{F}_{d,l,0}\phi(r,E):=\frac{\chi_{d,l}(r,E)}{4\pi}\int_{E_{\min}}^{E_{\max}}\beta_l(r,E')(\nu\Sigma_f)(r,E')\phi\left(r,E'\right)dE'.\tag{1.13}$$

The k-eigenproblem is also called the **criticality** problem. We strictly focus on the latter in this manuscript. It yields the following eigenvalue problem:

Find (ψ,k_{eff}) such that k_{eff} is an eigenvalue with maximal modulus and

$$\begin{cases}\mathbb{L}_0\psi\left(r,\vec{\omega},E\right)-\mathbb{H}_0\psi\left(r,\vec{\omega},E\right)=\dfrac{1}{k_{\text{eff}}}\mathbb{F}_0\psi\left(r,\vec{\omega},E\right),\quad\text{in }\mathcal{D}\\ \psi\left(r,\vec{\omega},E\right)=0,\quad\text{on }\Gamma^-,\end{cases}\tag{1.14}$$

where

$$\mathbb{F}_0\psi\left(r,\vec{\omega},E\right) := \frac{\chi(r,E)}{4\pi}\int_{E_{\min}}^{E_{\max}}(\nu\Sigma_f)(r,E')\phi\left(r,E'\right)dE', \tag{1.15}$$

with χ denoting the so-called total spectrum. Problem (1.14) is a **generalized eigenvalue problem** equivalent to

$$\begin{aligned}&\text{Find } (\psi,\lambda_{\text{eff}}) \text{ such that } \lambda_{\text{eff}} \text{ is an eigenvalue with minimal modulus and}\\ &\begin{cases}\left(\mathbb{L}_0-\mathbb{H}_0\right)\psi\left(r,\vec{\omega},E\right)=\lambda_{\text{eff}}\mathbb{F}_0\psi\left(r,\vec{\omega},E\right), & \text{in } \mathcal{D}\\ \psi\left(r,\vec{\omega},E\right)=0, & \text{on } \Gamma^-.\end{cases}\end{aligned} \tag{1.16}$$

with

$$\lambda_{\text{eff}} := \frac{1}{k_{\text{eff}}}. \tag{1.17}$$

1.2.1 Physical interpretation

The eigenvalue k_{eff} is called the effective multiplication factor, or k-effective, of the reactor core. This specific eigenvalue indicates whether the advection and scattering, or the fission dominates inside the core. Three main scenarios can asymptotically describe the reactor, depending on the value of k_{eff}:

- if $k_{\text{eff}} < 1$, the fission reaction is not the prevailing phenomenon, hence the total number of neutrons tends towards zero along time; the reactor is said to be subcritical;
- if $k_{\text{eff}} = 1$, both creation and absorption of neutrons prevail with same importance inside the system; the reactor is said to be critical;
- if $k_{\text{eff}} > 1$, the fission dominates the absorption phenomenon, therefore a chain reaction phenomenon takes place inside the system, and the total number of neutrons increases at an exponential rate, the system then tends to collapse; the reactor is said to be supercritical.

1.2.2 The Krein–Rutman theorem: existence and uniqueness of the solution to the criticality problem

As we are interested in the smallest eigenvalue in modulus of Problem (1.16), the Krein–Rutman theorem [77] ensures the existence and uniqueness of the solution. We start with a series of assumptions on the cross sections.

Assumption 1.2.1. *Let us recall the velocity variable $\vec{v}$ such that $\vec{\omega} = \vec{v}/|\vec{v}|$ and $E = \frac{1}{2}m|\vec{v}|^2$. We assume that:*

- $|\vec{v}|\Sigma_t \in L^\infty(\mathcal{R}\times\mathbb{S}_2\times[E_{min},E_{max}])$;
- $f_s\left(r,\vec{v'},\vec{v}\right) := m\frac{|\vec{v'}|}{|\vec{v}|}\Sigma_s\left(r,(\vec{\omega'},E')\to(\vec{\omega},E)\right)$ *and* $f_f\left(r,\vec{v'},\vec{v}\right) := m\frac{|\vec{v'}|}{|\vec{v}|}\chi(r,E)(\nu\Sigma_f)(r,E')$ *are real, nonnegative and measurable over $\vec{v}$ and $\vec{v'}$;*

- *There exists* $\alpha > 0$ *such that,*

$$\begin{cases} |\vec{v}|\Sigma_t(r,\vec{v}) - \displaystyle\int f_s\left(r,\vec{v'},\vec{v}\right) d\vec{v'} \geq \alpha, \\ |\vec{v'}|\Sigma_t(r,\vec{v'}) - \displaystyle\int f_s\left(r,\vec{v'},\vec{v}\right) d\vec{v} \geq \alpha, \end{cases} \quad a.e \text{ in } (r,\vec{v}) \in \mathcal{D};$$

- *There exists* $\beta, \beta' > 0$ *such that,*

$$\begin{cases} \displaystyle\int f_f\left(r,\vec{v'},\vec{v}\right) d\vec{v} + \int f_s\left(r,\vec{v'},\vec{v}\right) d\vec{v} \leq \beta, \\ \displaystyle\int f_f\left(r,\vec{v'},\vec{v}\right) d\vec{v'} + \int f_s\left(r,\vec{v'},\vec{v}\right) d\vec{v'} \leq \beta', \end{cases} \quad \text{in } \mathcal{D};$$

Theorem 1.2.2 (Krein–Rutman theorem, Theorem 1 of Appendix in Chapter VIII of [42])**.** *Let* $(\mathcal{X}, \|\cdot\|)$ *be a real Banach space. Let* $\mathcal{K} \subset \mathcal{X}$ *be a salient closed cone*[1] *satisfying* $\mathcal{X} = \mathcal{K} - \mathcal{K}$*, of non-empty interior* $\mathring{\mathcal{K}}$*. Let* B *be a compact operator, strongly positive on* $\mathcal{K}$*, i.e., for all* $v \in \mathcal{K}$*,* $v \neq 0$*,* $Bv \in \mathring{\mathcal{K}}$*. Let* B^* *be the dual operator of* B*. Then, its spectral radius* $\rho(B)$ *is a simple eigenvalue of* B *and* B^**, and is associated with a unique eigenvector* $u \in \mathring{\mathcal{K}}$ *that satisfies* $\|u\| = 1$ *(respectively* $u^* \in \mathring{\mathcal{K}}^* = \{v^* \in \mathcal{X}^* \,|\, \forall v \in \mathcal{K}, \langle v^*, v\rangle \geq 0\}$ *that satisfies* $\|u^*\| = 1$*).*

The application of Theorem 1.2.2 to the criticality problem yields the following result.

Theorem 1.2.3 (Theorem 1.2.1 of [12])**.** *Let* $1 < p < \infty$*. Under Assumption 1.2.1, Problem* (1.16) *has a countable number of eigenvalues and eigenvectors. The eigenvectors are elements of the Banach space*

$$W_p := \{u \in L^p(\mathcal{D}),\, \vec{v}.\nabla u \in L^p(\mathcal{D})\}\,.$$

Furthermore, if f_f *is positive, there exists a unique positive unit eigenvector of Problem* (1.16) *associated with the smallest eigenvalue* λ_{eff} *in modulus, which is simple and positive.*

1.3 Discretization of the steady-state neutron transport equation

The goal of this section is to consider some classical discretization of Problem (1.16) for the variables of energy E, direction ω, and space r.

1.3.1 Energy discretization

We start with the energy variable $E \in [E_{\min}, E_{\max}]$. A standard technique to discretize the energy variable is the so-called **multigroup approximation**, which enables the system to be described over G different energy groups or intervals, defined by the energy values $E_0 = E_{\max} > E_1 > \ldots > E_{G-1} > E_G = E_{\min}$, such that $[E_{\min}, E_{\max}] = \bigcup_{g=1}^{G} [E_g, E_{g-1}]$. Over each interval $\mathcal{I}_g = [E_g, E_{g-1}]$, the flux and cross sections are determined along mean

[1] i.e. $\mathcal{K}$ satisfies the three following properties: $0 \in \mathcal{K}$; $u, v \in \mathcal{K} \implies \alpha u + \beta v \in \mathcal{K}$, $\forall \alpha, \beta \geq 0$; $v \in \mathcal{K}$ and $-v \in \mathcal{K} \implies v = 0$.

evaluations. Indeed, suppose that there exists, for all $g \in \{1, \dots, G\}$, a function of energy $h^g(E)$ and a multigroup angular flux $\psi^g(r, \vec{\omega})$ such that

$$\psi(r, \vec{\omega}, E) \approx h^g(E)\psi^g(r, \vec{\omega}), \quad \forall E \in \mathcal{I}_g, \tag{1.18}$$

with

$$\int_{\mathcal{I}_g} h^g(E)dE = 1.$$

It is far from trivial to determine such functions h^g $(1 \leq g \leq G)$ in order to get reliable results with the multigroup approximation. We refer to [95] for studies on that matter. A main challenge is to take into account the strong oscillating behavior of the cross sections with respect to the energy, as illustrated in Figure 1.1. This phenomenon is modeled by so-called *self-shielding* techniques, developed in [39], for example. These methods involve homogenization of the cross sections in a medium with strong spatial and geometrical simplifications.

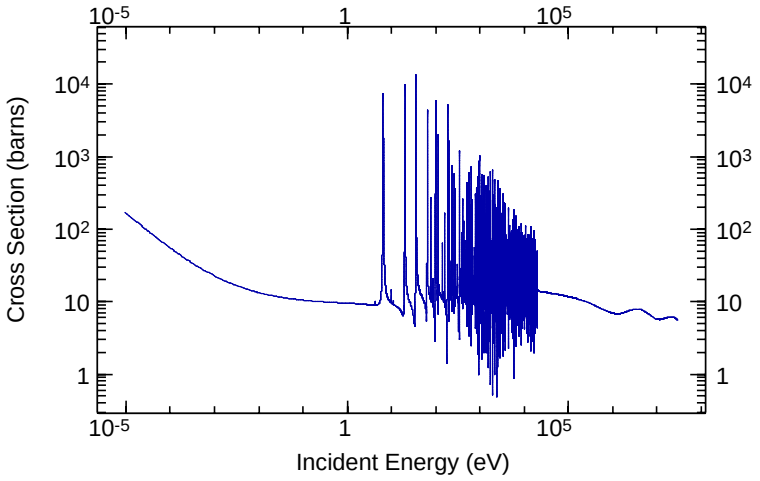

Figure 1.1: Total microscopic cross section Σ_t of Uranium-238 isotope as a function of the energy. Source: International Atomic Energy Agency (IAEA).

For all $1 \leq g \leq G$, we define

$$\Sigma_t^g(r) = \int_{\mathcal{I}_g} h^g(E)\Sigma_t(r, E)dE,$$

$$\Sigma_s^{g' \to g}\left(r, \vec{\omega'} \to \vec{\omega}\right) = \int_{\mathcal{I}_g} h^g(E)dE \int_{\mathcal{I}_{g'}} \Sigma_s\left(r, (\vec{\omega'}, E') \to (\vec{\omega}, E)\right) h^{g'}(E')dE',$$

$$(\nu\Sigma_f)^g(r) = \int_{\mathcal{I}_g} h^g(E)(\nu\Sigma_f)(r, E)dE,$$

$$\chi^g(r) = \int_{\mathcal{I}_g} h^g(E)\chi(r, E)dE.$$

The multigroup approximation of Problem (1.16) yields

$$\begin{array}{l}\text{Find } \left(\boldsymbol{\psi} = (\psi^1, \dots, \psi^G), \lambda\right) \text{ such that } \lambda \text{ is an eigenvalue with minimal modulus,} \\ \begin{cases} \mathbb{L}^g \psi^g (r, \vec{\omega}) - \mathbb{H}^g \boldsymbol{\psi} (r, \vec{\omega}) = \lambda \mathbb{F}^g \boldsymbol{\psi} (r, \vec{\omega}), \quad \text{in } \mathcal{R} \times \mathbb{S}_2, \quad \text{for all } 1 \leq g \leq G, \\ \psi^g (r, \vec{\omega}) = 0, \quad \text{on } \Gamma_0^-, \end{cases} \end{array} \tag{1.19}$$

with

$$\Gamma_0^- = \{(r, \vec{\omega}) \in \partial\mathcal{R} \times \mathbb{S}_2, \, \vec{\omega}.\vec{n}(r) < 0\},$$

and where, for all $1 \leq g \leq G$, and for all $E \in \mathcal{I}_g$,

$$\mathbb{L}^g \psi^g (r, \vec{\omega}) = \vec{\omega}.\nabla \psi^g (r, \vec{\omega}) + \Sigma_t^g(r) \psi^g (r, \vec{\omega}), \tag{1.20}$$

$$\mathbb{H}^g \boldsymbol{\psi} (r, \vec{\omega}) = \sum_{g'=1}^{G} \int_{\mathbb{S}_2} \Sigma_s^{g' \to g} \left(r, \vec{\omega'} \to \vec{\omega}\right) \psi^{g'}(r, \vec{\omega'}) d\vec{\omega'}, \tag{1.21}$$

$$\mathbb{F}^g \boldsymbol{\psi} (r, \vec{\omega}) = \frac{1}{4\pi} \sum_{g'=1}^{G} \chi^g(r) (\nu\Sigma_f)^{g'}(r) \int_{\mathbb{S}_2} \psi^{g'}(r, \vec{\omega}) d\vec{\omega}. \tag{1.22}$$

Among other approaches for the discretization of the energy variable, we note the probability table method [104] which is more accurate than the multigroup approximation, as it takes the variations of the cross sections on each interval $\mathcal{I}_g$ into account. An example of an application of the probability table method is the 1D neutron transport code SN1D [88]. Another discretization technique is the method of finite elements [4], mostly with a Galerkin projection of the angular flux along high-dimensional polynomial bases, so that the high-oscillating behavior in the cross sections' variations are taken into account. This approach generally implies expensive computational cost and a very large number of degrees of freedom in the resulting system. In addition to that, a wavelet Galerkin method has recently been carried out as it allows to solve sparse systems (see [116, 80, 53]).

1.3.2 Angular discretization

Throughout the rest of the document, we assume the **isotropic scattering hypothesis**, meaning that neutrons are scattered with no preferred direction, and the scattering cross section Σ_s only depends on the cosine of the incidental and scattered direction, i.e.,

$$\Sigma_s(r, \vec{\omega'} \to \vec{\omega}) \approx \Sigma_s(r, \vec{\omega'}.\vec{\omega}), \tag{1.23}$$

where we specifically omit the energy variable. As the quantity $\vec{\omega}'.\vec{\omega}$ ranges the interval $[-1, 1]$ over $\mathbb{S}_2$, it is standard to expand the scattering cross section along the basis of Legendre polynomials $(P_l)_{l \geq 0}$, i.e.,

$$\Sigma_s(r, \vec{\omega'}.\vec{\omega}) = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) \Sigma_{s,l}(r) P_l(\vec{\omega'}.\vec{\omega}), \tag{1.24}$$

where

$$\Sigma_{s,l}(r) = 2\pi \int_{-1}^{-1} \Sigma_s(r, \theta) P_l(\theta) dc$$

is the Legendre moment of order l of the scattering cross section.

To complete the discretization in the angular variable $\vec{\omega}$, we expand the unit sphere $\mathbb{S}_2$ along the basis of spherical harmonics $(Y_{l,m})_{l\geq 0,-l\leq m\leq l}$ and we introduce the flux moments, defined, for all $l \geq 0$, and all $m \in \{-l, \ldots, l\}$, by

$$\phi_{l,m}^g(r) = \int_{\mathbb{S}_2} \psi^g(r,\vec{\omega})\, Y_{l,m}(\vec{\omega}) d\vec{\omega}, \quad \text{for all } 1 \leq g \leq G. \tag{1.25}$$

Hence, using expressions (1.24) and (1.25), and the addition theorem[2], the multigroup neutron transport equations (1.19) result in solving the approximate problem

$$\vec{\omega}.\nabla\psi^g(r,\vec{\omega}) + \Sigma_t^g(r)\psi^g(r,\vec{\omega}) - \sum_{g'=1}^{G}\sum_{l=0}^{N}\Sigma_{s,l}^{g'\to g}(r)\sum_{m=-l}^{l}\phi_{l,m}^{g'}(r)Y_{l,m}(\vec{\omega}) = \lambda\mathbb{F}^g\psi(r,\vec{\omega}), \tag{1.26}$$

for all $1 \leq g \leq G$, where N is a given positive integer. We discuss here two standard methods to solve Problem (1.26): the S_N method [28] and the P_N method [85].

The S_N method

This first method relies on an approximation of integrals over the unit sphere $\mathbb{S}_2$ using a quadrature formula over $D = N(N+2)$ directions. More specifically, one discretizes the unit sphere $\mathbb{S}_2$ into the set of directions $\{\vec{\omega}_d,\, 1 \leq d \leq D\}$ associated with the weights $\{w_d\}_{1\leq d\leq D}$. For any measurable function f over $\mathbb{S}_2$, it holds

$$\int_{\mathbb{S}_2} f(\vec{\omega})d\vec{\omega} \approx \sum_{d=1}^{D} w_d f(\vec{\omega}_d).$$

Hence, the S_N approximation of Problem (1.26) writes

Find $\left(\left(\psi_1^1, \ldots, \psi_D^1, \ldots, \psi_1^G, \ldots, \psi_D^G\right), \lambda\right)$ such that λ is an eigenvalue with minimal modulus, and for all $1 \leq g \leq G$, for all $1 \leq d \leq D$,

$$\begin{cases} \vec{\omega_d}.\nabla\psi_d^g(r) + \Sigma_t^g(r)\psi_d^g(r) - \displaystyle\sum_{g'=1}^{G}\sum_{d'=1}^{D} w_{d'}\psi_{d'}^{g'}(r)\Theta_N^{g'\to g}(\vec{\omega_{d'}},\vec{\omega_d}) \\ \qquad = \dfrac{\lambda}{4\pi}\displaystyle\sum_{g'=1}^{G}\chi^g(r)(\nu\Sigma_f)^{g'}(r)\sum_{d'=1}^{D} w_{d'}\psi_{d'}^{g'}(r), \quad \text{in } \mathcal{R}, \\ \Theta_N^{g'\to g}(\vec{\omega_{d'}},\vec{\omega_d}) = \displaystyle\sum_{l=0}^{N}\Sigma_{s,l}^{g'\to g}(r)\sum_{m=-l}^{l} Y_{l,m}(\vec{\omega_{d'}})Y_{l,m}(\vec{\omega_d}), \\ \psi_d^g(r) = 0, \quad \text{on } \partial\mathcal{R}_d^- = \{r \in \partial\mathcal{R},\, \vec{\omega_d}.\vec{n}(r) < 0\}. \end{cases} \tag{1.27}$$

We refer to [9, 115, 101] for discussions on the convergence of such an angular discretization. Note that the choice of a quadrature rule on the unit sphere $\mathbb{S}_2$ remains a complex issue. We refer to [1] for recent advances on this topic. Several properties are targeted:

- the use of positive weights for stability and convergence issues;
- the ability of considering as many spherical harmonics as possible;

[2]It writes $P_l(\vec{\omega'}.\vec{\omega}) = \dfrac{4\pi}{2l+1}\displaystyle\sum_{m=-l}^{l} Y_{l,m}^*(\vec{\omega'})Y_{l,m}(\vec{\omega})$.

- an even distribution of the directions;
- the rotational invariance under some symmetry group in the set of directions.

The P_N method

This method relies on the expansion of any function in $L^2(\mathbb{S}_2)$ along the basis of spherical harmonics $(Y_{l,m})_{l\geq 0,-l\leq m\leq l}$. Then, for any $1 \leq g \leq G$, a truncated expansion of the neutron angular flux ψ^g at the order N reads as

$$\psi^g(r,\vec{\omega}) \approx \sum_{l=0}^{N} \sum_{m=-l}^{l} \phi_{l,m}^g(r) Y_{l,m}(\vec{\omega}), \tag{1.28}$$

where $(\phi_{l,m}^g(r))_{l\geq 0,-l\leq m\leq l}$ are the flux moments defined in (1.25).

Projections of Equation (1.26) onto each spherical harmonic $Y_{l,m}$ are carried out. Using expression (1.28) in (1.26), and the orthogonality of the spherical harmonics, the following $(N+1)^2$ coupled equations are obtained

Find $\left(\left(\phi_{l,m}^g\right)_{1\leq g\leq G,0\leq l\leq N,-l\leq m\leq l}, \lambda\right)$ such that λ is an eigenvalue with minimal modulus, and for all $1 \leq g \leq G$, all $0 \leq l \leq N$, and all $-l \leq m \leq l$,

$$\sum_{l'=0}^{N} \sum_{m'=-l'}^{l'} \left(\int_{\mathbb{S}_2} \vec{\omega} Y_{l',m'}(\vec{\omega}) Y_{l,m}(\vec{\omega}) d\vec{\omega} \right) . \nabla \phi_{l',m'}^g(r) + \Sigma_t^g(r) \phi_{l,m}^g(r) - \sum_{g'=1}^{G} \Sigma_{s,l}^{g'\to g}(r) \phi_{l,m}^{g'}(r)$$
$$= \lambda \sum_{g'=1}^{G} \chi^g(r) (\nu\Sigma_f)^{g'}(r) \delta_{l,0}\delta_{m,0} \phi_{0,0}^{g'}(r), \quad \text{in } \mathcal{R}, \tag{1.29}$$

supplemented with boundary conditions on $\partial\mathcal{R}$.

A simplified version of the P_N method, called the SP_N method [56], also enables the discretization of the angular variable under some hypotheses. It was first considered to decrease the computational complexity in the resolution of the P_N equations. It is based on the diffusion approximation theory (see Section 1.4), and on the assumption that the solution to the transport equation is locally planar, and thus, can be computed by solving 1D slab problems. We also note the existence of finite element discretizations for the unit sphere $\mathbb{S}_2$ [14] (see Section 1.3.3 for an introduction to the method).

1.3.3 Spatial discretization

In this section, we consider the mono-energetic neutron transport equation for an energy group $g \in [\![1, G]\!]$ along a direction $\vec{\omega}_d \in \mathbb{S}_2$ which reads

$$\begin{aligned} &\text{Find } \psi_d^g \in V_0 \text{ such that} \\ &\begin{cases} \vec{\omega_d}.\nabla\psi_d^g(r) + \Sigma_t^g(r)\psi_d^g(r) = q_d^g(r), & \text{in } \mathcal{R}, \\ \psi_d^g(r) = 0, & \text{on } \partial\mathcal{R}_d^-, \end{cases} \end{aligned} \tag{1.30}$$

where

$$V_0 = \left\{ v \in L^2(\mathcal{R}),\ \vec{\omega_d}.\nabla v \in L^2(\mathcal{R}),\ v_{|\partial\mathcal{R}_d^-} = 0 \right\},$$

and, in the case of an angular discretization with the S_N method as in (1.27),

$$q_d^g(r) := \sum_{g'=1}^{G}\sum_{d'=1}^{D} w_{d'}\psi_{d'}^{g'}\sum_{l=0}^{N}\Sigma_{s,l}^{g'\to g}(r)\sum_{m=-l}^{l} Y_{l,m}(\vec{\omega_{d'}})Y_{l,m}(\vec{\omega_d}) + \frac{\lambda}{4\pi}\sum_{g'=1}^{G}\chi^g(r)(\nu\Sigma_f)^{g'}(r)\sum_{d'=1}^{D} w_{d'}\psi_{d'}^{g'},$$

and the boundary condition applies over $\partial\mathcal{R}_d^- = \{r \in \partial\mathcal{R},\, \vec{\omega_d}.\vec{n}(r) < 0\}$, where $\vec{n}$ is the outward unit normal vector on $\partial\mathcal{R}$. Finally, in order to discretize along the space variable $r \in \mathcal{R}$, we use a Galerkin projection of Problem (1.30). The *finite element method* (FEM) [50] aims at projecting onto the sapce of piecewise polynomials. It is well-known that continuous finite elements raise stability issues because of the advection term [51, Chapter 61]. In order to alleviate this issue, a *discontinuous Galerkin (DG) finite element method* has been introduced in the case of the neutron transport equation [83] (see also Chapter 60 of [51]). This method also easily enables local mesh refinement due to the absence of the continuity assumption between two elements. Convergence analysis of the DG method is discussed in [82, 83, 73, 100].

Let $\mathcal{T}_\mathcal{N}$ be an affine simplicial mesh of the reactor core $\mathcal{R}$, such that $\overline{\mathcal{R}} = \bigcup_{K\in\mathcal{T}_\mathcal{N}} K$. We define the finite element approximation space $V_\mathcal{N}$ by

$$V_\mathcal{N} = \left\{v^\mathcal{N} \in L^2(\mathcal{R}),\, \forall K \in \mathcal{T}_\mathcal{N},\, v^\mathcal{N}_{|K} \in \mathbb{P}_k\right\}, \tag{1.31}$$

where $\mathbb{P}_k$ stands for the space of polynomials of degree at most $k \in \mathbb{N}$, and thus, $\mathcal{N}$ stands for the dimension of $V_\mathcal{N}$. We set

$$\mathcal{F}_\mathcal{N} = \bigcup_{K\in\mathcal{T}_\mathcal{N}} \partial K, \quad \text{and,} \quad v_\pm(r) = \lim_{\varepsilon\to\pm 0} v(r+\varepsilon\vec{\omega_d}), \quad \forall r \in \mathcal{F}_\mathcal{N},$$

with the convention that $v_\pm(r)$ is zero whenever r is at the boundary $\partial\mathcal{R}$ and that the limit is taken outside the domain.

A DG method has been introduced in [83]. For Problem (1.30), the discrete formulation writes

$$\text{Find } \psi^\mathcal{N} \in V_\mathcal{N} \text{ such that}$$
$$\sum_{K\in\mathcal{T}_\mathcal{N}}\int_K \left(\vec{\omega_d}.\nabla\psi^\mathcal{N} + \Sigma_t^g\psi^\mathcal{N}\right)v^\mathcal{N} + \int_{\mathcal{F}_\mathcal{N}\setminus\partial\mathcal{R}} (\vec{\omega_d}.\vec{n})\left(\psi_+^\mathcal{N} - \psi_-^\mathcal{N}\right)v_+^\mathcal{N} = \int_\mathcal{R} q_d^g v^\mathcal{N}, \quad \forall v^\mathcal{N} \in V_\mathcal{N}. \tag{1.32}$$

We give more details and we exhibit the whole discretization with finite elements for the resolution of the multigroup neutron diffusion equations in Section 1.4.2. Note that for the diffusion equation, we use a particular Discontinuous Galerkin method, namely the *Symmetric Interior Penalty Galerkin* method (SIPG) [46, Chapter 4].

Among other discretization methods for the space variable, we find the finite difference method [65] and the finite volume method [84]. If the former suffers from slow convergence rates and struggles to be generalized to non-cartesian and hexagonal geometries, the latter lacks some underlying mathematical framework to enable defining an associated adjoint problem. Another method used in the codes for neutronics is the method of characteristics.

It is based on an exact integration of Problem (1.30) along a trajectory generated by the direction $\vec{\omega_d}$. While it gives quite reliable approximations, the main challenge of this method is the memory imprint, especially in the case of complex geometries. We refer to [13, 66, 68] for applications of the method to deterministic codes in neutronics.

1.4 Transport and Diffusion

As we detailed in Section 1.3, the discretization of Problem (1.16) along the angular variable is not trivial and implies expensive calculations in terms of computational time and memory imprint. In real-world applications, it is usual to approximate the transport model by the **diffusion** problem, which results in a much less expensive discretized problem than the discretized transport problem. In this section, we introduce the diffusion problem, we show to what extent it is an asymptotic limit of the transport problem, and we provide physics-motivated arguments to illustrate the derivation of this model.

1.4.1 The diffusion approximation

Mathematical derivation On the one hand, some assumptions on the cross sections enable an asymptotic model for the transport problem. We consider here the time-dependent mono-energetic neutron transport equation with isotropic scattering which writes over $[0,T] \times \mathcal{R} \times \mathbb{S}_2$

$$\frac{\partial \psi}{\partial t}(t,r,\vec{\omega}) + \vec{\omega}.\nabla\psi(t,r,\vec{\omega}) + \Sigma_t(r)\psi(t,r,\vec{\omega}) - \frac{\Sigma_{s,0}(r)}{4\pi}\int_{\mathbb{S}_2}\psi(t,r,\vec{\omega})\,d\vec{\omega} = 0, \qquad (1.33)$$

where we make the assumption that the cross sections Σ_t and Σ_s do not evolve in time, and the scattering cross section Σ_s is expanded at the order 0 as in (1.24). We supplement the equation with vacuum boundary conditions and an initial condition ψ^0. Let us introduce a small parameter $\varepsilon > 0$ and we assume that there exist $\widetilde{\Sigma}_t$, $\widetilde{\Sigma}_{s,0}$ and $\widetilde{\Sigma}_a$, bounded functions over $\mathcal{R}$ such that

$$\begin{aligned}
\Sigma_t(r) &= \frac{\widetilde{\Sigma}_t(r)}{\varepsilon},\\
\Sigma_{s,0}(r) &= \frac{\widetilde{\Sigma}_{s,0}(r)}{\varepsilon},\\
\widetilde{\Sigma}_t(r) &= \widetilde{\Sigma}_{s,0}(r) - \varepsilon^2\widetilde{\Sigma}_a(r).
\end{aligned}$$

Therefore, under these assumptions, the following theorem [42, Chapter XXI, Section 5.2, Theorem 1] refers to the diffusion problem as an asymptotic limit of the transport problem.

Theorem 1.4.1 (Convergence of the diffusion problem in L^∞-norm)**.** *Let $\mathcal{R}$ be a bounded open subset of $\mathbb{R}^3$ with a regular boundary. For any $k \in \mathbb{N}$ and $\alpha \in]0,1]$, we denote by $\mathcal{C}^{k,\alpha}$ the Hölder space of functions with continuous derivatives up through order k and such that the k-th partial derivatives are α-Hölder continuous. Let us assume the following:*

- $\exists \beta, \beta' > 0$, $\beta \leq \widetilde{\Sigma}_{s,0}(r) \leq \beta'$, *over* $\mathcal{R}$;
- $\exists \alpha \in]0,1[$, $\widetilde{\Sigma}_{s,0} \in \mathcal{C}^{3,\alpha}(\overline{\mathcal{R}})$, $\widetilde{\Sigma}_a \in \mathcal{C}^{2,\alpha}(\overline{\mathcal{R}})$.

Let $\varepsilon > 0$. Then, there exists a symmetric positive definite matrix $D = (D_{ij})_{1\leq i,j\leq 3}$ such that, for any initial condition ψ^0 which satisfies

$$\psi^0 \in \mathcal{C}^{4,\alpha}(\overline{\mathcal{R}}), \quad \psi^0\big|_{\partial\mathcal{R}} = 0, \quad \text{and} \quad \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial \psi^0}{\partial x_j}\right)\Big|_{\partial\mathcal{R}} = 0,$$

the solution ψ_ε in $\mathcal{C}\left([0,+\infty[, L^\infty(\mathcal{R}\times\mathbb{S}_2)\right)$ of the **transport** *problem*

$$\begin{cases} \dfrac{\partial \psi_\varepsilon}{\partial t}(t,r,\vec{\omega}) + \dfrac{1}{\varepsilon}\vec{\omega}.\nabla\psi_\varepsilon(t,r,\vec{\omega}) - \widetilde{\Sigma}_a(r)\psi_\varepsilon(t,r,\vec{\omega}) \\ \quad + \dfrac{\widetilde{\Sigma}_{s,0}(r)}{\varepsilon^2}\left(\psi_\varepsilon(t,r,\vec{\omega}) - \dfrac{1}{4\pi}\displaystyle\int_{\mathbb{S}_2}\psi_\varepsilon\left(t,r,\vec{\omega'}\right)d\vec{\omega'}\right) = 0, \quad \text{in } [0,+\infty[\times\mathcal{R}\times\mathbb{S}_2, \\ \psi_\varepsilon(t,r,\vec{\omega}) = 0, \quad \text{on } \Gamma^-, \quad \text{for } t>0, \\ \psi_\varepsilon(0,\cdot) = \psi^0, \end{cases} \tag{1.34}$$

and the solution ϕ in $\mathcal{C}\left([0,+\infty[, L^\infty(\mathcal{R})\right)$ of the **diffusion** *problem*

$$\begin{cases} \dfrac{\partial\phi}{\partial t}(t,r) - \displaystyle\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial\phi}{\partial x_j}(t,r)\right) - \widetilde{\Sigma}_a(r)\phi(t,r) = 0, \quad \text{in } [0,+\infty[\times\mathcal{R}, \\ \phi(t,r) = 0, \quad \text{on } \partial\mathcal{R}, \quad \text{for } t>0, \\ \phi(0,\cdot) = \psi^0 \end{cases} \tag{1.35}$$

verify, for all $t \geq 0$,

$$\|\psi_\varepsilon(t,\cdot) - \phi(t,\cdot)\|_{L^\infty(\mathcal{R}\times\mathbb{S}_2)} \leq C_{\psi^0}\varepsilon e^{\delta t}(1+t),$$

where $\delta = \sup_{r\in\mathcal{R}} \widetilde{\Sigma}_a(r)$ and C_{ψ^0} is a positive constant independent of ε.

Physical derivation On the other hand, a physical argument actually allows to characterize the matrix $(D_{ij})_{1\leq i,j\leq 3}$ as a single positive coefficient $D > 0$. Let us consider the P_1 expansion of the angular flux as in (1.28)

$$\psi(r,\vec{\omega}) \approx \phi_{0,0}(r)Y_{0,0}(\vec{\omega}) + \phi_{1,-1}(r)Y_{1,-1}(\vec{\omega}) + \phi_{1,0}(r)Y_{1,0}(\vec{\omega}) + \phi_{1,1}(r)Y_{1,1}(\vec{\omega}). \tag{1.36}$$

Using the explicit expressions of the spherical harmonics, then there exists a so-called current vector $\vec{J}$ such that

$$\psi(r,\vec{\omega}) \approx \frac{\phi(r)}{4\pi} + \frac{3}{4\pi}\vec{\omega}.\vec{J}(r). \tag{1.37}$$

Then, substituting expression (1.37) into the P_1 equations (see Section 1.3.2), we obtain

$$\mathrm{div}\vec{J}(r) + \left(\Sigma_t(r) - \Sigma_{s,0}(r)\right)\phi(r) = 0. \tag{1.38}$$

We also assume that the so-called Fick's law holds, i.e.,

$$\vec{J}(r) = -D(r)\nabla\phi(r), \tag{1.39}$$

which states that the neutrons go from regions of high concentration to regions of low concentration with magnitude $D > 0$ that is proportional to the concentration gradient $\nabla\phi$. Combining (1.38) and (1.39), we obtain the diffusion problem.

1.4.2 The multigroup neutron diffusion equations

As for the neutron transport model, we can apply the same discretization techniques for the diffusion model, such as the multigroup approximation for the energy variable, and the finite element method for the spatial variable. The multigroup neutron diffusion equations [48, Chapter 7] write

$$\begin{aligned}&\text{Find } \left(\phi = (\phi^1, \dots, \phi^G), \lambda_{\text{eff}}\right) \in H_0^1(\mathcal{R})^G \times \mathbb{R}\\ &\text{such that } \lambda_{\text{eff}} \text{ is an eigenvalue with minimal modulus, and for all } 1 \le g \le G,\\ &\begin{cases} -\operatorname{div}\left(D^g \nabla \phi^g\right) + \displaystyle\sum_{g'=1}^{G} \Sigma^{gg'} \phi^{g'} = \lambda_{\text{eff}} \chi^g \sum_{g'=1}^{G} (\nu\Sigma_f)^{g'} \phi^{g'} & \text{in } \mathcal{R},\\ \phi^g = 0, \quad \text{on } \partial\mathcal{R}, \end{cases}\end{aligned} \tag{1.40}$$

where, for all $g, g' \in [\![1, G]\!]$:

- $D^g : \mathcal{R} \to \mathbb{R}_+$ is the diffusion coefficient of group g;
- $\Sigma^{gg'} : \mathcal{R} \to \mathbb{R}$ with $\Sigma^{gg'} = \begin{cases} \Sigma_t^g - \Sigma_{s,0}^{g \to g} & \text{if } g = g', \\ -\Sigma_{s,0}^{g' \to g} & \text{otherwise;} \end{cases}$
- $\Sigma_t^g : \mathcal{R} \to \mathbb{R}$ is the total cross section of group g;
- $\Sigma_{s,0}^{g' \to g} : \mathcal{R} \to \mathbb{R}$ is the scattering cross section of anisotropy order 0 from group g' to group g;
- $\chi^g : \mathcal{R} \to \mathbb{R}$ is the neutron total spectrum of group g;
- $\nu^g : \mathcal{R} \to \mathbb{R}$ is the average number of neutrons emitted per fission of group g;
- $\Sigma_f^g : \mathcal{R} \to \mathbb{R}$ is the fission cross section of group g.

Note that in the equations above, we used the short-hand notation $(\nu\Sigma_f)^g$ to refer to the product $\nu^g \Sigma_f^g$, for $g \in [\![1, G]\!]$.

Assumption 1.4.2. *We assume that, for all $g, g' \in [\![1, G]\!]$:*

- *The coefficients D^g, $\Sigma^{gg'}$, χ^g, $(\nu\Sigma_f)^g$ are all functions of $L^\infty(\mathcal{R})$;*
- *$\exists (D^g)_*, (D^g)^* > 0$, $\forall k \in [\![1, K]\!]$, $(D^g)_* \le (D^g)_{|\mathcal{R}_k} \le (D^g)^*$;*
- *$\exists (\Sigma^{gg})_*, (\Sigma^{gg})^* > 0$, $\forall k \in [\![1, K]\!]$, $(\Sigma^{gg})_* \le (\Sigma^{gg})_{|\mathcal{R}_k} \le (\Sigma^{gg})^*$;*
- *$\exists 0 \le \alpha \le G - 1$, $|\Sigma^{gg'}| \le \alpha \Sigma^{gg}$, a.e. in $\mathcal{R}$;*
- *$(\nu\Sigma_f)^g \ge 0$, a.e. in $\mathcal{R}$.*

We also assume that there exists $\tilde{g}, \tilde{g'} \in [\![1, G]\!]$ such that $\chi^{\tilde{g}} (\nu\Sigma_f)^{\tilde{g'}} \neq 0 \in L^\infty(\mathcal{R})$.

Under Assumption 1.4.2, Problem (1.40) is well-posed [62, Theorem 2.12] if it is supplemented with a normalization condition on the multigroup flux $\phi = (\phi^1, \dots, \phi^G)$.

The two-group neutron diffusion equations

Throughout our study, we mostly focus on the two-group neutron diffusion equations, that is Problem (1.40) with $G = 2$. The latter writes

$$\begin{array}{l}\text{Find } \left(\phi = (\phi^1, \phi^2), \lambda_{\text{eff}}\right) \in H_0^1(\mathcal{R})^2 \times \mathbb{R} \\ \text{such that } \lambda_{\text{eff}} \text{ is an eigenvalue with minimal modulus,} \\ \begin{cases} -\operatorname{div}\left(D^1 \nabla \phi^1\right) + \Sigma^{11}\phi^1 + \Sigma^{12}\phi^2 = \lambda_{\text{eff}}\chi^1 \left((\nu\Sigma_f)^1\phi^1 + (\nu\Sigma_f)^2\phi^2\right) & \text{in } \mathcal{R}, \\ -\operatorname{div}\left(D^2 \nabla \phi^2\right) + \Sigma^{21}\phi^1 + \Sigma^{22}\phi^2 = \lambda_{\text{eff}}\chi^2 \left((\nu\Sigma_f)^1\phi^1 + (\nu\Sigma_f)^2\phi^2\right) & \text{in } \mathcal{R}, \\ \phi^g = 0, \quad \text{on } \partial\mathcal{R}, \quad \text{for } g = \{1, 2\}. \end{cases} \end{array} \tag{1.41}$$

The two-group neutron diffusion equations: the weak formulation

We define the bilinear forms $a : H_0^1(\mathcal{R})^2 \times H_0^1(\mathcal{R})^2 \longrightarrow \mathbb{R}$ and $b : H_0^1(\mathcal{R})^2 \times H_0^1(\mathcal{R})^2 \longrightarrow \mathbb{R}$ by

$$\begin{aligned} a(\psi, \phi) := & \int_{\mathcal{R}} \left(D^1 \nabla \psi^1\right) \cdot \nabla \phi^1 + \int_{\mathcal{R}} \Sigma^{11}\psi^1\phi^1 + \int_{\mathcal{R}} \Sigma^{12}\psi^2\phi^1 \\ & + \int_{\mathcal{R}} \left(D^2 \nabla \psi^2\right) \cdot \nabla \phi^2 + \int_{\mathcal{R}} \Sigma^{21}\psi^1\phi^2 + \int_{\mathcal{R}} \Sigma^{22}\psi^2\phi^2, \end{aligned} \tag{1.42}$$

$$\begin{aligned} b(\psi, \phi) := & \int_{\mathcal{R}} \chi^1 \left((\nu\Sigma_f)^1\psi^1 + (\nu\Sigma_f)^2\psi^2\right) \phi^1 \\ & + \int_{\mathcal{R}} \chi^2 \left((\nu\Sigma_f)^1\psi^1 + (\nu\Sigma_f)^2\phi^2\right) \phi^2, \end{aligned} \tag{1.43}$$

for all $\psi = (\psi^1, \psi^2) \in H_0^1(\mathcal{R})^2$, $\phi = (\phi^1, \phi^2) \in H_0^1(\mathcal{R})^2$.

Therefore, the weak form (or variational form) of Problem (1.41) writes

$$\begin{array}{l}\text{Find } \left(\phi = (\phi^1, \phi^2), \lambda_{\text{eff}}\right) \in H_0^1(\mathcal{R})^2 \times \mathbb{R} \\ \text{such that } \lambda_{\text{eff}} \text{ is an eigenvalue with minimal modulus,} \\ a\left((\phi^1, \phi^2), (\varphi^1, \varphi^2)\right) = \lambda_{\text{eff}} b\left((\phi^1, \phi^2), (\varphi^1, \varphi^2)\right), \quad \forall (\varphi^1, \varphi^2) \in H_0^1(\mathcal{R})^2. \end{array} \tag{1.44}$$

We also introduce the associated adjoint problem that reads

$$\begin{array}{l}\text{Find } \left(\phi^* = (\phi^{*,1}, \phi^{*,2}), \lambda_{\text{eff}}\right) \in H_0^1(\mathcal{R})^2 \times \mathbb{R} \\ \text{such that } \lambda_{\text{eff}} \text{ is an eigenvalue with minimal modulus,} \\ a\left((\varphi^1, \varphi^2), (\phi^{*,1}, \phi^{*,2})\right) = \lambda_{\text{eff}} b\left((\varphi^1, \varphi^2), (\phi^{*,1}, \phi^{*,2})\right), \quad \forall (\varphi^1, \varphi^2) \in H_0^1(\mathcal{R})^2. \end{array} \tag{1.45}$$

The two-group neutron diffusion equations: the discrete form

We discretize the spatial domain $\mathcal{R}$ with finite elements as introduced in Section 1.3.3. To do so, we consider a shape-regular mesh $\mathcal{T}_\mathcal{N}$ of $\mathcal{R}$ and an associated conformal finite element approximation space $\widetilde{V}_\mathcal{N}$ of dimension $\widetilde{\mathcal{N}}$. We also define by $V_\mathcal{N} := (\widetilde{V}_\mathcal{N})^2$ which has dimension $\mathcal{N} = 2\widetilde{\mathcal{N}}$. We assume that the mesh is such that the cross sections are regular on each element.

The discrete variational formulation associated with Problem (1.44) writes

$$\begin{aligned} &\text{Find } \left(\phi^{\mathcal{N}}, \lambda^{\mathcal{N}}\right) \in V_{\mathcal{N}} \times \mathbb{R} \text{ such that } \lambda^{\mathcal{N}} \text{ is an eigenvalue with minimal modulus,} \\ &a\left(\phi^{\mathcal{N}}, \varphi^{\mathcal{N}}\right) = \lambda^{\mathcal{N}} b\left(\phi^{\mathcal{N}}, \varphi^{\mathcal{N}}\right), \quad \text{for all } \varphi^{\mathcal{N}} \in V_{\mathcal{N}}, \end{aligned} \tag{1.46}$$

as well as the discrete variational formulation associated with Problem (1.45) writes

$$\begin{aligned} &\text{Find } \left(\phi^{*,\mathcal{N}}, \lambda^{\mathcal{N}}\right) \in V_{\mathcal{N}} \times \mathbb{R} \text{ such that } \lambda^{\mathcal{N}} \text{ is an eigenvalue with minimal modulus,} \\ &a\left(\varphi^{\mathcal{N}}, \phi^{*,\mathcal{N}}\right) = \lambda^{\mathcal{N}} b\left(\varphi^{\mathcal{N}}, \phi^{*,\mathcal{N}}\right), \quad \text{for all } \varphi^{\mathcal{N}} \in V_{\mathcal{N}}. \end{aligned} \tag{1.47}$$

The two-group neutron diffusion equations: the matrix form

Let us denote by $(\theta_k)_{1\leq k\leq \mathcal{N}}$ a basis of $V_{\mathcal{N}}$. Let $u := (u_k)_{1\leq k\leq \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$ be the coordinates of $\phi^{\mathcal{N}}$ in the basis $(\theta_1, \ldots, \theta_{\mathcal{N}})$ so that

$$\phi^{\mathcal{N}} = \sum_{k=1}^{\mathcal{N}} u_k \theta_k. \tag{1.48}$$

Let us define the matrices $A := (a(\theta_j, \theta_i))_{1\leq i,j\leq \mathcal{N}}$ and $B := (b(\theta_j, \theta_i))_{1\leq i,j\leq \mathcal{N}}$. Then, Problem (1.46) is equivalent to the following generalized eigenvalue problem: Find $(u, \lambda) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R}$ such that λ is an eigenvalue with minimal modulus and

$$Au = \lambda B u. \tag{1.49}$$

Likewise, we expand the adjoint discrete flux $\phi^{*,\mathcal{N}}$ along the basis $(\theta_1, \ldots, \theta_{\mathcal{N}})$ so that

$$\phi^{*,\mathcal{N}} = \sum_{k=1}^{\mathcal{N}} u_k^* \theta_k, \tag{1.50}$$

where $u^* := (u_k^*)_{1\leq k\leq \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$ are the coordinates of $\phi^{*,\mathcal{N}}$ in the basis $(\theta_k)_{1\leq k\leq \mathcal{N}}$. Therefore, Problem (1.47) can be formulated as: Find $(u^*, \lambda) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R}$ such that λ is an eigenvalue with minimal modulus and

$$A^T u^* = \lambda B^T u^*. \tag{1.51}$$

***

In this chapter, we presented the standard techniques that are used to discretize eigenvalue problems arising in neutronics. The parametric dependency of the problem motivates the use of *a posteriori* error estimators for optimization purposes, for instance, when one needs to solve the eigenvalue problem of interest for a very large number of parameter values. In particular, to avoid considerable computational costs, one may consider to solve a cheaper approximate problem instead. Thus, these estimators can quantify the error of approximation without necessarily solving the problem of reference. Next section is dedicated to the development of these estimates in the case of a generalized non-symmetric eigenvalue problem.

Chapter 2

A posteriori error estimates for parameter-dependent non-symmetric generalized eigenvalue problems

In this chapter, we derive *a posteriori* error estimates for non-symmetric generalized eigenvalue problems. *A posteriori* error analysis of symmetric eigenvalue problems naturally relies on residuals with respect to the operator-induced energy norm. Our main goal here is to generalize the state-of-the-art *a posteriori* estimates to the case of non-symmetric generalized eigenvalue problems, such as the multigroup neutron diffusion equations, presented at the end of the previous chapter. We particularly exhibit a parameter-dependent *prefactor* in the error upper bound that must be taken into account in order to get reliable estimates. Computing an accurate and optimal value of this prefactor is not an easy task, compared to the case of symmetric eigenvalue problems where it can be expressed by means of the spectral gap of the considered operator.

We start, in Section 2.1, with recalling an *a priori* error result between eigenvalue error and left and right best approximation eigenvector errors, by I. Babuška and J. Osborn [11], in the case of Galerkin approximations of a generalized non-symmetric eigenvalue problem. In Section 2.2, we remind some classical residual-based *a posteriori* error estimates for both symmetric and non-symmetric eigenvalue problems. We then derive, in Section 2.3, some error bounds on the left and right eigenvectors, as well as on the eigenvalue, in the case of a generalized non-symmetric eigenvalue problem. As we need computable and reliable *a posteriori* error estimates that take the information contained in the prefactors into account, we propose, in Section 2.4, a practical heuristic method to estimate these prefactors. To do so, we provide some elements of theoretical analysis to illustrate the close link between the obtained expression of the prefactor and its well-known counterpart in the case of symmetric eigenvalue problems.

2.1 *A priori* analysis of Galerkin approximations of generalized eigenvalue problems

In this section, we remind some crucial results on the Galerkin approximation of an eigenvalue problem, introduced by I. Babuška and J. Osborn [11] in 1989. They emphasized the close link between eigenvalue error and best approximation error in the eigenvectors.

Here, we remind the results and their proofs in the case where we only consider finite-dimensional spaces. Let V be a real Hilbert space of finite dimension $\mathcal{N}$, equipped with an inner product $\langle \cdot, \cdot \rangle_V$ and associated norm $\| \cdot \|_V$. Let us consider the following discrete eigenvalue problem of reference

$$\begin{aligned} &\text{Find } (u, \lambda) \in V \times \mathbb{R} \text{ such that} \\ &a(u, v) = \lambda b(u, v), \quad \forall v \in V, \end{aligned} \tag{2.1}$$

where $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are two bilinear forms on $V \times V$. The associated adjoint problem writes

$$\begin{aligned} &\text{Find } (u^*, \lambda) \in V \times \mathbb{R} \text{ such that} \\ &a(v, u^*) = \lambda b(v, u^*), \quad \forall v \in V. \end{aligned} \tag{2.2}$$

Let V_N be a linear subspace of V of dimension $N < \mathcal{N}$. We consider the following eigenvalue problem on V_N

$$\begin{aligned} &\text{Find } (u_N, \lambda_N) \in V_N \times \mathbb{R} \text{ such that} \\ &a(u_N, v_N) = \lambda_N b(u_N, v_N), \quad \forall v_N \in V_N, \end{aligned} \tag{2.3}$$

and the associated adjoint problem on V_N

$$\begin{aligned} &\text{Find } (u_N^*, \lambda_N) \in V_N \times \mathbb{R} \text{ such that} \\ &a(v_N, u_N^*) = \lambda_N b(v_N, u_N^*), \quad \forall v_N \in V_N. \end{aligned} \tag{2.4}$$

From the bilinear form $a(\cdot, \cdot)$, we define the following operators:

$$\begin{aligned} \mathcal{A} \ &: V \longrightarrow V \\ &u \longmapsto \mathcal{A}u \in V \text{ such that for all } v \in V, \quad a(u, v) = \langle \mathcal{A}u, v \rangle_V, \end{aligned}$$

$$\begin{aligned} \mathcal{A}^T \ &: V \longrightarrow V \\ &u \longmapsto \mathcal{A}^T u \in V \text{ such that for all } v \in V, \quad a(v, u) = \langle v, \mathcal{A}^T u \rangle_V, \end{aligned}$$

$$\begin{aligned} \mathcal{A}^{-T} \ &: V \longrightarrow V \\ &u \longmapsto \mathcal{A}^{-T} u \in V \text{ such that for all } v \in V, \quad a(v, \mathcal{A}^{-T} u) = \langle v, u \rangle_V. \end{aligned}$$

We respectively refer to (u, λ) and (u^*, λ) as solutions to (2.1) and (2.2), and we define

$$\tilde{u}^* = \mathcal{A}^T u^*.$$

Similarly, we respectively refer to (u_N, λ_N) and (u_N^*, λ_N) as solutions to (2.3) and (2.4), and we define

$$\tilde{u}_N^* = \mathcal{A}^T u_N.$$

Assumption 2.1.1. *Let us assume the following:*

1. *We assume that λ is a **simple** eigenvalue of Problem (2.1), and therefore of Problem (2.2), and that $\|u\|_V = \|\tilde{u}^*\|_V = 1$; we also assume that $\langle u, \tilde{u}^* \rangle_V \neq 0$;*
2. *We assume that λ_N is a **simple** eigenvalue of Problem (2.3), and therefore of Problem (2.4), and that $\langle u_N, \tilde{u}_N^* \rangle_V \neq 0$;*

3. *The bilinear forms* $a : V \times V \longrightarrow \mathbb{R}$ *and* $b : V \times V \longrightarrow \mathbb{R}$ *are continuous, i.e. there exists two positive constants* C_a *and* C_b *such that*

$$|a(u,v)| \leqslant C_a \|u\|_V \|v\|_V, \quad \forall u, v \in V, \tag{2.5}$$
$$|b(u,v)| \leqslant C_b \|u\|_V \|v\|_V, \quad \forall u, v \in V, \tag{2.6}$$

noting that the continuity is here immediate since V *is a finite-dimensional space;*

4. *The bilinear form* $a(\cdot,\cdot)$ *admits the inf-sup condition on* V

$$\inf_{u \in V} \sup_{v \in V} \frac{|a(u,v)|}{\|u\|_V \|v\|_V} \geqslant \gamma > 0; \tag{2.7}$$

5. *The bilinear form* $a(\cdot,\cdot)$ *admits the inf-sup condition on* V_N

$$\inf_{u_N \in V_N} \sup_{v_N \in V_N} \frac{|a(u_N,v_N)|}{\|u_N\|_V \|v_N\|_V} \geqslant \gamma_N > 0. \tag{2.8}$$

We define the operator T on V by

$$\begin{aligned} T \ &: V \longrightarrow V \\ &f \longmapsto Tf \in V \text{ such that } a(Tf, v) = b(f, v), \quad \forall v \in V. \end{aligned}$$

Its adjoint operator T^* is defined by

$$\begin{aligned} T^* \ &: V \longrightarrow V \\ &f^* \longmapsto T^* f^* \in V \text{ such that } a(v, T^* f^*) = b(v, f^*), \quad \forall v \in V. \end{aligned}$$

Note that T (resp. T^*) has $k = \frac{1}{\lambda}$ as a simple eigenvalue, associated with the eigenvector u (resp. $\tilde{u}^*$). We introduce the operator $T_N : u \in V \longmapsto (T_N u) \in V_N$ and its associated adjoint T_N^* such that

$$\forall u \in V, \quad a\left(T_N u, v_N\right) = b(u, v_N), \quad \forall v_N \in V_N, \tag{2.9}$$
$$\forall u \in V, \quad a\left(v_N, T_N^* u\right) = b(v_N, u), \quad \forall v_N \in V_N. \tag{2.10}$$

Let $\mathcal{C}$ be a closed curve in the complex plane that encloses both k and $k_N := \frac{1}{\lambda_N}$, for all $N \in [\![1, \mathcal{N}]\!]$, which lies in $\left\{z \in \mathbb{C} \mid (z-T)^{-1} \text{ and } (z-T_N)^{-1} \text{ exist}\right\}$. We assume that $\mathcal{C}$ encloses neither any other points from the spectrum of T, nor any other points from the spectrum of T_N. Let us define the spectral projectors Π and Π^* respectively onto $E = \text{Span}\{u\}$ and $E^* = \text{Span}\{\tilde{u}^*\}$ by

$$\Pi = \frac{|u\rangle\langle\tilde{u}^*|}{\langle u, \tilde{u}^*\rangle} = \frac{1}{2\pi i} \int_{\mathcal{C}} (z-T)^{-1} \, dz$$
$$\Pi^* = \frac{|\tilde{u}^*\rangle\langle u|}{\langle u, \tilde{u}^*\rangle} = \frac{1}{2\pi i} \int_{\mathcal{C}} (z-T^*)^{-1} \, dz,$$

see Lemma 2.3.4 to justify the definition, as well as the projectors Π_N and Π_N^* respectively onto $E_N = \text{Span}\{u_N\}$ and $E_N^* = \text{Span}\{\tilde{u}_N^*\}$, defined by

$$\Pi_N = \frac{|u_N\rangle\langle\tilde{u}_N^*|}{\langle u_N, \tilde{u}_N^*\rangle} = \frac{1}{2\pi i} \int_{\mathcal{C}} (z-T_N)^{-1} \, dz$$
$$\Pi_N^* = \frac{|\tilde{u}_N^*\rangle\langle u_N|}{\langle u_N, \tilde{u}_N^*\rangle} = \frac{1}{2\pi i} \int_{\mathcal{C}} (z-T_N^*)^{-1} \, dz.$$

Assumption 2.1.2. *There exists an integer* $1 \leqslant N_\delta < \mathcal{N}$ *such that*

$$\forall N \geqslant N_\delta, \quad \|\Pi - \Pi_N\| < 1/2. \tag{2.11}$$

The general result to consider in this section is the following.

Theorem 2.1.3 (*A priori* error analysis for generalized eigenvalue problems, Theorem 8.2 of [11], p. 695)**.** *Let* (u, λ) *and* (u^*, λ) *respectively be solutions to* (2.1) *and* (2.2)*. Let* (u_N, λ_N) *and* (u_N^*, λ_N) *respectively be solutions to* (2.3) *and* (2.4)*. Then, under Assumptions 2.1.1 and 2.1.2, there exists a constant* $C > 0$ *independent of* N *such that*

$$|\lambda_N - \lambda| \leqslant C \frac{1}{\gamma_N} \varepsilon_N \varepsilon_N^*,$$

where

$$\begin{aligned} \varepsilon_N &:= \inf_{v_N \in V_N} \|u - v_N\|_V , \\ \varepsilon_N^* &:= \inf_{v_N \in V_N} \|u^* - v_N\|_V . \end{aligned}$$

In order to apprehend the proof of Theorem (2.1.3), we include in the following a series of technical lemmas (see Section 7 of [11], pp. 685-691).

Lemma 2.1.4. *Let the* V*-orthogonal projectors onto* $E = \mathrm{Span}\{u\}$ *and* $E_N = \mathrm{Span}\{u_N\}$ *be respectively denoted by* π_E *and* π_{E_N}*. We set*

$$\hat{\delta}(E, E_N) := \max\left(\|u - \pi_{E_N} u\|, \|u_N - \pi_E u_N\|\right).$$

Then, under Assumptions 2.1.1 and 2.1.2, there exists two constants $C > 0$ *and* $C^* > 0$ *independent of* N *such that, for* $N \geqslant N_\delta$*,*

$$\begin{aligned} \hat{\delta}(E, E_N) &\leqslant C \left\| (T - T_N)_{|E} \right\|, \\ \hat{\delta}(E^*, E_N^*) &\leqslant C^* \left\| (T^* - T_N^*)_{|E^*} \right\|, \end{aligned}$$

and thus,

$$\begin{aligned} \|\Pi - \Pi_N\| &\leqslant C \left\| (T - T_N)_{|E} \right\|, \\ \|\Pi^* - \Pi_N^*\| &\leqslant C^* \left\| (T^* - T_N^*)_{|E^*} \right\|. \end{aligned}$$

Proof. As $\|u\| = 1$, we have

$$\begin{aligned} \|u - \pi_{E_N} u\| &\leqslant \|u - \Pi_N u\| \\ &\leqslant \|(\Pi - \Pi_N) u\| \\ &\leqslant \|\Pi - \Pi_N\| . \end{aligned}$$

Using (2.11), it holds that $\|u - \pi_{E_N} u\| < 1/2$, for $N \geqslant N_\delta$. V is a finite-dimensional Hilbert space, then, for $N \geqslant N_\delta$,

$$\begin{aligned} \|u_N - \pi_E u_N\| &\leqslant \frac{\|u - \pi_{E_N} u\|}{1 - \|u - \pi_{E_N} u\|}, \quad \text{(cf. Theorem 6.1 of [11])} \\ &\leqslant 2 \|u - \pi_{E_N} u\| , \end{aligned}$$

and hence

$$\hat{\delta}(E, E_N) \leqslant 2 \left\| u - \pi_{E_N} u \right\|.$$

Moreover, we have

$$\begin{aligned}
\|u - \Pi_N u\| &= \|\Pi u - \Pi_N u\| \\
&= \left\| \frac{1}{2\pi i} \int_{\mathcal{C}} \left[(z - T)^{-1} - (z - T_N)^{-1} \right] u \, dz \right\| \\
&= \left\| \frac{1}{2\pi i} \int_{\mathcal{C}} (z - T_N)^{-1} \left[(z - T_N)(z - T)^{-1} - I \right] u \, dz \right\| \\
&= \left\| \frac{1}{2\pi i} \int_{\mathcal{C}} (z - T_N)^{-1} \left[(z - T_N) - (z - T) \right] (z - T)^{-1} u \, dz \right\| \\
&= \left\| \frac{1}{2\pi i} \int_{\mathcal{C}} (z - T_N)^{-1} (T - T_N) (z - T)^{-1} u \, dz \right\|.
\end{aligned}$$

Let us recall that E is invariant for T, then

$$\|u - \Pi_N u\| \leqslant \frac{|\mathcal{C}|}{2\pi} \sup_{z \in \mathcal{C}} \left\| (z - T_N)^{-1} \right\| \left\| (T - T_N)_{|E} \right\| \sup_{z \in \mathcal{C}} \left\| (z - T)^{-1} \right\| \|u\|.$$

Therefore, it holds

$$\hat{\delta}(E, E_N) \leqslant C \left\| (T - T_N)_{|E} \right\|,$$

with $C = \dfrac{|\mathcal{C}|}{\pi} \max\limits_{N_\delta \leqslant N \leqslant \mathcal{N}} \left(\sup\limits_{z \in \mathcal{C}} \left\| (z - T_N)^{-1} \right\| \right) \sup\limits_{z \in \mathcal{C}} \left\| (z - T)^{-1} \right\|$.

We prove similarly that

$$\hat{\delta}(E^*, E_N^*) \leqslant C^* \left\| (T^* - T_N^*)_{|E^*} \right\|.$$

This proof also shows that

$$\begin{aligned}
\|\Pi - \Pi_N\| &\leqslant C \left\| (T - T_N)_{|E} \right\|, \\
\|\Pi^* - \Pi_N^*\| &\leqslant C^* \left\| (T^* - T_N^*)_{|E^*} \right\|.
\end{aligned}$$

□

Lemma 2.1.5. *Under Assumptions 2.1.1 and 2.1.2, there exists a constant $C > 0$ such that for $N \geqslant N_\delta$,*

$$|k_N - k| \leqslant |\langle \tilde{u}^*, (T - T_N) u \rangle| + C \left\| (T - T_N)_{|E} \right\| \left\| (T^* - T_N^*)_{|E^*} \right\|.$$

Proof. From (2.11), for $N \geqslant N_\delta$, using that $\|u\| = 1$, there holds

$$1 - \|\Pi_N u\| = \|\Pi u\| - \|\Pi_N u\| \leqslant \|\Pi - \Pi_N\| < 1/2.$$

Then, we have

$$\|\Pi_N u\| \geqslant 1/2.$$

Let us consider the operator $\left(\Pi_{N_{|E}}\right)^{-1} : E_N \longrightarrow E$, that is well defined since Assumption 2.1.2 holds. Therefore, the quantity $\left\|\left(\Pi_{N_{|E}}\right)^{-1}\right\|$ is bounded, as $\left\|\left(\Pi_{N_{|E}}\right)^{-1}\right\| \leqslant 1/2$. Let us simply denote the operator $\left(\Pi_{N_{|E}}\right)^{-1}$ by Π_N^{-1}. We now define the operators

$$\begin{aligned}\widehat{T} &= T_{|E} \ : \ E \longrightarrow E,\\ \widehat{T}_N &= \left(\Pi_N^{-1} T_N \Pi_N\right)_{|E} \ : \ E \longrightarrow E.\end{aligned}$$

Using $T_N \Pi_N = \Pi_N T_N$ and $\Pi_N^{-1}\Pi_{N_{|E}} = I_{|E}$,

$$\begin{aligned}k - k_N &= \mathrm{trace}(\widehat{T} - \widehat{T}_N)\\ &= \langle \tilde{u}^*, \left(\widehat{T} - \widehat{T}_N\right) u\rangle\\ &= \langle \tilde{u}^*, Tu - \Pi_N^{-1} T_N \Pi_N u\rangle\\ &= \langle \tilde{u}^*, \Pi_N^{-1}\Pi_N \left(T - T_N\right) u\rangle\\ &= \langle \tilde{u}^*, \left(T - T_N\right) u\rangle + \langle \tilde{u}^*, \left(\Pi_N^{-1}\Pi_N - I\right)\left(T - T_N\right) u\rangle.\end{aligned}$$

Let $L_N := \Pi_N^{-1}\Pi_N$. There holds $\mathrm{Ran}(L_N) = E$ and $\mathrm{Ker}(L_N) = \mathrm{Ker}(\Pi_N) = (E_N^*)^\perp$. In other terms, L_N is the projection onto E along $(E_N^*)^\perp$. Let L_N^* be the dual of L_N, i.e. the projection onto E_N^* along $E^\perp$. Thus, since $\Pi^* \tilde{u}^* = \tilde{u}^*$, and $\left(L_N^* - I\right)\Pi_N^* \tilde{u}^* = 0$, then

$$\langle \tilde{u}^*, \left(\Pi_N^{-1}\Pi_N - I\right)\left(T - T_N\right) u\rangle = \langle \left(\Pi^* - \Pi_N^*\right)\tilde{u}^*, \left(L_N - I\right)\left(T - T_N\right) u\rangle.$$

Since $\|L_N\| \leqslant \left\|\left(\Pi_{N_{|E}}\right)^{-1}\right\| \|\Pi_N\| \leqslant 1/2$, L_N is bounded in norm for $N \geqslant N_\delta$. Then, using Lemma 2.1.4, there exists a constant $C > 0$ such that for $N \geqslant N_\delta$,

$$\begin{aligned}\left|\langle \tilde{u}^*, \left(\Pi_N^{-1}\Pi_N - I\right)\left(T - T_N\right) u\rangle\right| &\leqslant \left(\max_{1\leqslant N\leqslant \mathcal{N}} \|L_N - I\|\right) \left\|\left(T - T_N\right)_{|E}\right\|\\ &\qquad \left\|\left(\Pi^* - \Pi_N^*\right)_{|E^*}\right\| \|\tilde{u}^*\| \|u\|\\ &\leqslant C \left\|\left(T - T_N\right)_{|E}\right\| \left\|\left(T^* - T_N^*\right)_{|E^*}\right\|.\end{aligned}$$

Thus, it yields, for $N \geqslant N_\delta$,

$$|k_N - k| \leqslant |\langle \tilde{u}^*, \left(T - T_N\right) u\rangle| + C \left\|\left(T - T_N\right)_{|E}\right\| \left\|\left(T^* - T_N^*\right)_{|E^*}\right\|.$$

□

Lemma 2.1.6. *Under Assumptions 2.1.1 and 2.1.2, where it is recalled that γ_N is the constant from the inf-sup condition of the bilinear form $a(\cdot,\cdot)$ on V_N, and*

$$\begin{aligned}\varepsilon_N &= \inf_{v_N \in V_N} \|u - v_N\|_V,\\ \varepsilon_N^* &= \inf_{v_N \in V_N} \|u^* - v_N\|_V,\end{aligned}$$

there exists a constant $C > 0$ independent of N such that

$$\left\|\left(T - T_N\right)_{|E}\right\| \leqslant C \frac{1}{\gamma_N} \varepsilon_N,$$

and

$$\left\|\left(T^* - T_N^*\right)_{|E^*}\right\| \leqslant C \frac{1}{\gamma_N} \varepsilon_N^*.$$

Proof. First, since E is invariant for T, then for all $u \in E$,

$$\begin{aligned}\inf_{v_N \in V_N} \|Tu - v_N\| &= \inf_{v_N \in V_N} \left\| \frac{Tu}{\|Tu\|} - \frac{v_N}{\|Tu\|} \right\| \|Tu\| \\ &= \varepsilon_N \|Tu\| \\ \inf_{v_N \in V_N} \|Tu - v_N\| &\leqslant \varepsilon_N \|T\| \|u\|. \end{aligned} \tag{2.12}$$

On the one hand, according to Céa's lemma, as $\gamma_N \leqslant C_a$, where C_a is defined in (2.5), for all $v \in V$,

$$\|(T - T_N)\, v\|_V \leqslant \left(1 + \frac{1}{\gamma_N}\right) \inf_{v_N \in V_N} \|Tv - v_N\|_V \leqslant \frac{(C_a + 1)}{\gamma_N} \inf_{v_N \in V_N} \|Tv - v_N\|_V . \tag{2.13}$$

On the other hand, it holds that for all $v \in V$,

$$a(Tv, Tv) = b(v, Tv).$$

Using (2.6) and (2.7),

$$a(Tv, Tv) \geqslant \gamma \|Tv\|_V^2,$$

and

$$b(v, Tv) \leqslant C_b \|v\|_V \|Tv\|_V,$$

hence

$$\|T\| \leqslant \frac{C_b}{\gamma}. \tag{2.14}$$

Therefore, using (2.12), (2.13) and (2.14),

$$\left\| (T - T_N)_{|E} \right\| \leqslant \sup_{u \in V} \frac{\|(T - T_N)\, u\|_V}{\|u\|_V} \leqslant \frac{C_b}{\gamma} \frac{(C_a + 1)}{\gamma_N} \varepsilon_N.$$

We then proceed similarly for the adjoint operators. □

Proof of Theorem (2.1.3). Let us write, for all $v_N \in V_N$,

$$\begin{aligned} |\langle \tilde{u}^*, (T - T_N)\, u \rangle| &= \left| a\left((T - T_N)\, u, \mathcal{A}^{-T} \tilde{u}^* \right) \right| \\ &= \left| a\left((T - T_N)\, u, \mathcal{A}^{-T} \tilde{u}^* - v_N \right) \right| \\ &\leqslant C_a \|(T - T_N)\, u\|_V \, \|u^* - v_N\|_V \\ &\leqslant C_a \left\| (T - T_N)_{|E_h} \right\| \|u^*\| \varepsilon_N^*, \end{aligned}$$

with $u^* = \mathcal{A}^{-T} \tilde{u}^*$. Using Lemma 2.1.5 and Lemma 2.1.6, there exists a constant $C > 0$ independent of N such that for $N \geqslant N_\delta$,

$$|k_N - k| \leqslant C \frac{1}{\gamma_N} \varepsilon_N \varepsilon_N^*.$$

This concludes the proof. □

We write a similar result for the eigenvectors of Problems (2.1) and (2.2).

Theorem 2.1.7. *Let* (u, λ) *and* (u^*, λ) *respectively be solutions to* (2.1) *and* (2.2). *Let* (u_N, λ_N) *and* (u_N^*, λ_N) *respectively be solutions to* (2.3) *and* (2.4) *such that*

$$\langle u, u_N \rangle_V \geq 0,$$
$$\langle u^*, u_N^* \rangle_V \geq 0.$$

Let ε_N *and* ε_N^* *as defined in Theorem* (2.1.3). *Then, under Assumptions 2.1.1 and 2.1.2 listed above, there exists two constants* $C > 0$ *and* $C^* > 0$ *independent of* N *such that*

$$\|u_N - u\|_V \leqslant C\varepsilon_N,$$
$$\|u_N^* - u^*\|_V \leqslant C^*\varepsilon_N^*.$$

2.2 State-of-the-art *a posteriori* error estimation for eigenvalue problems

We review in this section some existing results on *a posteriori* error estimation for both symmetric and non-symmetric eigenvalue problems. We remind Problem (2.1) as our reference eigenvalue problem, as well as Problem (2.2) as the associated adjoint problem. Note that the problem of interest actually depends on a parameter $\mu \in \mathcal{P}$ (e.g., see Problem (1.40)), $\mathcal{P} \subset \mathbb{R}^p$, for some $p \geq 1$, which is omitted in this section, as only one parameter value μ is considered. We respectively refer to (u, λ) and (u^*, λ) as solutions to (2.1) and (2.2), where λ is the smallest eigenvalue in modulus, which assumed to be simple and positive. Let $(\theta_1, \ldots, \theta_{\mathcal{N}})$ be a basis of V, we also consider the matrix forms of Problems (2.1) and (2.2), which respectively write

$$\begin{aligned} &\text{Find } (u, \lambda) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R} \text{ such that} \\ &Au = \lambda Bu, \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} &\text{Find } (u^*, \lambda) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R} \text{ such that} \\ &A^T u^* = \lambda B^T u^*, \end{aligned} \tag{2.16}$$

where

$$A := \left(a(\theta_j, \theta_i)\right)_{1 \leq i,j \leq \mathcal{N}}, \quad \text{and} \quad B := \left(b(\theta_j, \theta_i)\right)_{1 \leq i,j \leq \mathcal{N}}.$$

We make the assumption that A is invertible and B is nonnegative.

2.2.1 In the symmetric case

Let us first focus on the case of symmetric eigenvalue problems. In a more general context, upper bounds on the eigenvalue error, based on the perturbation theory, emerged at the beginning of the 1950s, notably with the Kato–Temple theorem [74, 113], which gives a residual-based error estimate of second-order in the residual norm. In the case of a diagonalizable matrix, F. L. Bauer and C. T. Fike suggested, in 1960, a weaker upper bound of first-order in the residual norm, stating that the sensitivity of the eigenvalues is proportional to the norm of the residual vector and the condition number of the matrix of eigenvectors.

In the context of finite element approximation theory, the main case study is the Laplace eigenvalue problem, where error bonds were derived in [29] and [86]. Nevertheless, these error estimates loose accuracy if the diameter of the largest mesh element does

not tend to zero. Both eigenvalue and eigenvector error bounds have been derived for nonconforming methods, e.g. in [41] for nonconforming finite elements, in [59] for discontinuous Galerkin finite elements, and in [49, 72] for mixed finite elements. However, all these estimates present either unknown, solution-dependent, or non-computable terms.

In the context of reduced bases, a first residual-based *a posteriori* error estimate on the smallest eigenvalue of a symmetric parametrized eigenvalue problem was developed at the end of the 1990s by Y. Maday and A. Patera, in their pioneering work [92], and was used to certify an associated reduced-basis approximation [89]. Then, *a posteriori* error estimates for multiple eigenvalues were developed in [70] and [69] for similar reduced-order applications.

Let us consider $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be two symmetric positive-definite bilinear forms. Then, their associated matrices A and B are both symmetric and positive-definite. Problem (2.1) is then equivalent to (2.2), and Problem (2.15) is equivalent to (2.16). The eigenvalues are all real and positive and can be ordered as

$$0 < \lambda < \lambda_2 \leq \ldots \leq \lambda_{\mathcal{N}}.$$

We start with some algebraic (in $\mathbb{R}^{\mathcal{N}}$) error estimates, for the case where $B = I$, discussed in [106]. The first one is an immediate consequence of the Kato–Temple theorem [74, 113].

Proposition 2.2.1 (Corollary 3.4 of [106])**.** *Let A be a symmetric positive definite matrix, and $B = I$. Let $u_N \in \mathbb{R}^{\mathcal{N}}$, such that $\|u_N\|_2 = 1$, and $\lambda_N := \langle Au_N, u_N\rangle$. We define the residual vector R_N*

$$R_N = Au_N - \lambda_N u_N,$$

and the gap

$$\delta = \min_{2\leq i\leq \mathcal{N}} |\lambda_i - \lambda_N|.$$

Then,

$$|\lambda - \lambda_N| \leq \frac{\|R_N\|_2^2}{\delta}.$$

An associated *a posteriori* error estimate for the eigenvector writes as follows.

Proposition 2.2.2 (Theorem 3.9 of [106])**.** *Let us consider the same assumptions as in Proposition 2.2.1. Let us consider two scalars $\alpha, \beta \in [0,1]$ such that $u_N = \alpha u + \beta u^{\perp}$, where $u^{\perp}$ is an orthogonal vector to u. Then,*

$$\beta\langle u, u_N\rangle \leq \frac{\|R_N\|_2}{\delta}.$$

We continue with the works of T. Horger, et al. [69] and S. Giani, et al. [60] which aimed at estimating multiple eigenvalues, and not only the smallest one, as it was also investigated in [27, 26]. These error estimates were carried out with respect to a so-called energy norm, induced by the symmetric positive-definite operator, associated with the eigenvalue problem. Although their results were derived with respect to an infinite-dimensional space V, we rewrite them to consider V of finite dimension $\mathcal{N} \in \mathbb{N}^*$, and target only the smallest eigenvalue.

Proposition 2.2.3 (Theorem 3.1 and Remark 3.2 of [69], Theorem 2.2 of [60])**.** *Let $a(\cdot,\cdot)$ be a symmetric positive definite bilinear form on $V \times V$. Let*

$$b(u,v) := \langle u, v\rangle_V, \quad \forall (u,v) \in V \times V.$$

Let u, $u_N \in V$ satisfy the following normalization condition

$$b(u,u) = 1, \quad \text{and} \quad b(u_N,u_N) = 1.$$

We assume that $\lambda_N := a(u_N,u_N) \geq \lambda$. Let define the residual

$$Res_N(v) := a(u_N,v) - \lambda_N b(u_N,v), \quad \forall v \in V.$$

We denote by $res_N \in V$ the Riesz representation of Res_N such that

$$a(res_N,v) = Res_N(v), \quad \forall v \in V.$$

Let λ_2 be the closest eigenvalue to λ_N, aside from λ, and let $\widetilde{\lambda}_2$ be the eigenvalue which satisfies

$$\max_{2\leq i\leq \mathcal{N}} \left|\frac{\lambda_i}{\lambda_i - \lambda_N}\right| = \left|\frac{\widetilde{\lambda}_2}{\widetilde{\lambda}_2 - \lambda_N}\right|.$$

Then,

$$|\lambda - \lambda_N| \leq \left|\frac{\lambda_2}{\lambda_2 - \lambda_N}\right|^2 \|Res_N\|^2_{-a},$$
$$|\lambda - \lambda_N| \leq \left|\frac{\widetilde{\lambda}_2}{\widetilde{\lambda}_2 - \lambda_N}\right|^2 \|Res_N\|^2_{-a},$$

where $\|Res_N\|_{-a} = \|res_N\|_a$, and $\|\cdot\|_a := a(\cdot,\cdot)^{1/2}$.

Moreover, if we denote by Π_u the L^2-orthogonal projection onto $\mathrm{Span}\{u\}$*, the following error estimates hold*

$$\|u_N - \Pi_u u_N\|^2_a \leq \left|\frac{\lambda_2}{\lambda_2 - \lambda_N}\right|^2 \|Res_N\|^2_{-a},$$

and

$$\sup_{v_N \in \mathrm{Span}\{u_N\}} \frac{\|v_N - \Pi_u v_N\|^2_a}{\|v_N\|^2_a} \leq \left|\frac{\widetilde{\lambda}_2}{\widetilde{\lambda}_2 - \lambda_N}\right|^2 \frac{\|Res_N\|^2_{-a}}{\lambda_N}.$$

Remark 2.2.4. *Let u_N be an approximation of u and let us consider the residual vector*

$$R_N = Au_N - \lambda_N B u_N.$$

Then, the residual norm $\|Res_N\|^2_{-a}$ introduced in Proposition 2.2.3 would write

$$\|Res_N\|^2_{-a} = R_N^T A^{-1} R_N.$$

As a consequence, computing the residual norm would then require the computation of the matrix A^{-1}, which becomes tedious in the case where the operation needs to be carried out for a very large number of parameters μ. In addition to that, if A exhibits an affine decomposition along the parameter μ (see Section 3.2), its inversion would break the offline/online decomposition, as the inverse operation is neither linear nor affine. We then must consider the residual with respect to a norm that is independent of any parameter-dependent operator (see Section 2.3).

2.2.2 In the non-symmetric case

In the early stages of *a posteriori* error analysis for non-symmetric eigenvalue problems, we note here a pioneering work from S. Giani, A. Międlar, et al. [61], who developed error estimates based on Kato's square root theorem [10, 75]. It states that any maximal accretive operator $\mathcal{A}$ (which is the case for all convection-diffusion-reaction operators) admits a unique maximal accretive operator $\mathcal{A}^{1/2}$ which solves $\mathcal{Z}^2 = \mathcal{A}$. Furthermore, as the eigenvectors do not provide immediate orthogonality properties, we use the properties of spectral projectors, that are defined in the result below. Note that their approach was limited to the case where $b(\cdot,\cdot)$ is symmetric (and positive definite), which is not the case for eigenvalue problems arising in neutronics (cf. Problem (1.40)). Let us start with the following estimates for the eigenvector errors.

Proposition 2.2.5 (Proposition 4 of [61])**.** *Let $\mathcal{A}$ and its associated adjoint $\mathcal{A}^T$ be the maximal accretive operators defined as*

$$a(u,v) = \langle \mathcal{A}u, v\rangle_V = \langle u, \mathcal{A}^T v\rangle_V, \quad \forall (u,v) \in V \times V.$$

Let us denote by $\mathcal{A}^{1/2}$ (resp. $(\mathcal{A}^T)^{1/2}$) the square root of the operator $\mathcal{A}$ (resp. $\mathcal{A}^T$). Let

$$b(u,v) := \langle u, v\rangle_V, \quad \forall (u,v) \in V \times V.$$

Let $u_N, u_N^* \in V$ and $\lambda_N \in \mathbb{R}$ with corresponding residuals

$$\begin{aligned} Res_N(v) &= a(u_N, v) - \lambda_N b(u_N, v), \quad \forall v \in V, \\ Res_N^*(v) &= a(v, u_N^*) - \lambda_N b(v, u_N^*), \quad \forall v \in V, \end{aligned}$$

and the following residual norms

$$\begin{aligned} \|Res_N\|_{\mathcal{A}^T,-1/2} &:= \left\|\mathcal{A}^{1/2}u_N - \lambda_N \mathcal{A}^{-1/2}u_N\right\|, \\ \|Res_N^*\|_{\mathcal{A},-1/2} &:= \left\|(\mathcal{A}^T)^{1/2}u_N^* - \lambda_N (\mathcal{A}^T)^{-1/2}u_N^*\right\|. \end{aligned}$$

Moreover, let us define the following so-called spectral projectors

$$\Pi_{\text{int}} = \int_{\mathcal{C}_\lambda} (z - \mathcal{A})^{-1} dz, \quad \textit{and} \quad \Pi_{\text{int}}^* = \int_{\mathcal{C}_\lambda} (z - \mathcal{A}^T)^{-1} dz,$$

where $\mathcal{C}_\lambda$ is a curve in the complex plane which encloses both λ and λ_N, and no other eigenvalues of $\mathcal{A}$ and $\mathcal{A}^T$. Then,

$$\begin{aligned} \inf_{v \in \text{Span}\{u\}} \|v - u_N\| &\leq \left\|\mathcal{A}^{1/2}\left(\lambda_N - \mathcal{A}_{|(\text{Span}\{u\})^\perp}\right)^{-1}(I - \Pi_{\text{int}})\right\| \|Res_N\|_{\mathcal{A}^T,-1/2}, \\ \inf_{v^* \in \text{Span}\{u^*\}} \|v^* - u_N^*\| &\leq \left\|(\mathcal{A}^T)^{1/2}\left(\lambda_N - \mathcal{A}^T_{|(\text{Span}\{u^*\})^\perp}\right)^{-1}(I - \Pi_{\text{int}}^*)\right\| \|Res_N^*\|_{\mathcal{A},-1/2}. \end{aligned}$$

An associated *a posteriori* estimate of the eigenvalue error reads as follows.

Proposition 2.2.6 (Theorem 14 of [61])**.** *Let us consider the same assumptions as in Proposition 2.2.5. We assume that $b(u_N, u_N^*) \neq 0$ and we set*

$$\lambda_N = \frac{a(u_N, u_N^*)}{b(u_N, u_N^*)}.$$

Under some additional assumptions[1]*, there exists a constant* $C_N > 0$ *such that*

$$|\lambda - \lambda_N| \leq \frac{C_N|\lambda|}{|b(u_N, u_N^*)|} \frac{\|Res_N\|_{\mathcal{A}^T,-1/2}}{\|u_N\|} \frac{\|Res_N^*\|_{\mathcal{A},-1/2}}{\|u_N^*\|}.$$

Note that in the estimates presented above, no specific assumptions are made on how the approximate eigenfunctions are chosen, although most of their applications deal with finite element or reduced basis approximations. Furthermore, most of these results analytically provide reliable *a posteriori* estimates. Nevertheless, the major drawback of the use of these estimates in practice lies in their implementation and, if applicable in the context of parameter dependent eigenvalue problems, the associated cost, as illustrated below. In particular, the residual norm $\|Res_{\mu,N}\|_{\mathcal{A}_\mu{}^T,-1/2}$ introduced in Proposition 2.2.5 cannot be directly used in practical calculations.

2.3 *A posteriori* error estimation

The goal of this section is to build *a posteriori* error bounds on the error between the solutions of the eigenvalue problems (2.15) and (2.16), denoted by (u, u^*, λ), and their approximations, denoted by (u_N, u_N^*, λ_N). Again, to simplify notation, the subscript μ is omitted in this section. In all the following, $\mathbb{R}^{\mathcal{N}}$ is equipped with the Euclidean inner product[2] denoted by $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. We assume that

$$\|u\| = \|u^*\| = \|u_N\| = \|u_N^*\| = 1.$$

We also make the following additional assumption which is satisfied in the problems we are eventually interested in for neutronics applications.

Assumption 2.3.1. *A is invertible and there exists a positive eigenvalue λ which realizes the smallest modulus solution to* (2.15)*. Moreover, the eigenvalue λ is* ***simple****.*

As in neutronics, one quantity of interest is the k-effective which is the inverse of the smallest eigenvalue (see (1.17)), we introduce the scalars

$$k := \frac{1}{\lambda}, \quad \text{and} \quad k_N := \frac{1}{\lambda_N}.$$

Lemma 2.3.2. *Under Assumption 2.3.1,* $\langle u^*, Au \rangle \neq 0$*.*

Proof. Given A invertible, we consider the matrix $M = A^{-1}B$, and we set $v^* = A^T u^*$. Then, k is an eigenvalue of M associated with the right eigenvector u and the left eigenvector v^*. We have

$$\langle u^*, Au \rangle = \langle u, A^T u \rangle = \langle u, v^* \rangle.$$

Let us then show that $u \notin \text{Span}\{v^*\}^\perp$.

We first show that $\text{Span}\{v^*\}^\perp$ is invariant by M. We denote by $(e_1, \ldots, e_{\mathcal{N}-1})$ an orthonormal basis of $\text{Span}\{v^*\}^\perp$. For all $i \in [\![1, \mathcal{N}-1]\!]$, we have

$$\langle Me_i, v^* \rangle = \langle e_i, M^T v^* \rangle = k \langle e_i, v^* \rangle = 0,$$

[1] We refer to Theorem 14 of [61] for more details on these assumptions which are not trivial to write in our context.

[2] It is easy to generalize the results presented here to any Hilbertian norm.

hence, $Me_i \in \text{Span}\{v^*\}^\perp = \text{Span}\{e_1, \dots, e_{\mathcal{N}-1}\}^\perp$.

Therefore, if we write the matrix M with respect to the basis $\left(e_1, \dots, e_{\mathcal{N}-1}, e_{\mathcal{N}} = \frac{v^*}{\|v^*\|}\right)$, the last row of the matrix M is then $(0, \dots, 0, k)$, as we have

$$Me_{\mathcal{N}} = \sum_{i=1}^{\mathcal{N}} \langle Me_{\mathcal{N}}, e_i \rangle e_i,$$

and

$$\langle Me_{\mathcal{N}}, e_{\mathcal{N}} \rangle = \langle e_{\mathcal{N}}, M^T e_{\mathcal{N}} \rangle = k.$$

As k is a simple eigenvalue of M, the restriction $M_{|\text{Span}\{v^*\}^\perp}$ does not have k as an eigenvalue, therefore, $u \notin \text{Span}\{v^*\}^\perp$. □

Remark 2.3.3. *Note that without Assumption 2.3.1, it is possible that* $\langle u^*, Au \rangle = 0$. *Indeed, a simple example is to take* $A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ *and* $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. *Let* $u = (1, 0)$ *and* $u^* = (0, 1)$. *Equations* (2.15) *and* (2.16) *are satisfied with* $\lambda = 1$ *while* $\langle u^*, Au \rangle = 0$.

Therefore, our goal now is to derive bounds for the quantities $e_N^k := |k - k_N|$, $e_N^u := \|u - u_N\|$, and $e_N^{u^*} := \|u^* - u_N^*\|$. In order to estimate these errors, we first define the following residual vector quantities

$$R_N = (B - k_N A)\, u_N, \tag{2.17}$$

$$R_N^* = \left(B^T - k_N A^T\right) u_N^*. \tag{2.18}$$

Moreover, we introduce the vector

$$\tilde{u}^* = \frac{A^T u^*}{\|A^T u^*\|}, \tag{2.19}$$

and the matrix

$$M = A^{-1} B, \tag{2.20}$$

which is well-defined since A is invertible from Assumption 2.3.1. Note that it then holds that

$$Mu = ku, \quad M^T \tilde{u}^* = k\tilde{u}^*.$$

2.3.1 Error estimates on the eigenvectors

Let $P \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ and $P^* \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ be the matrices associated with the spectral projection operators onto $\text{Span}\{\tilde{u}^*\}^\perp$ and $\text{Span}\{u\}^\perp$, respectively. More precisely, P and P^* are defined as

$$P = I - \frac{u(\tilde{u}^*)^T}{\langle u, \tilde{u}^* \rangle}, \tag{2.21}$$

$$P^* = I - \frac{\tilde{u}^* u^T}{\langle u, \tilde{u}^* \rangle}, \tag{2.22}$$

where I denotes the identity matrix in $\mathbb{R}^{\mathcal{N} \times \mathcal{N}}$. Before presenting the *a posteriori* error estimates, we first collect a few useful auxiliary lemmas.

Lemma 2.3.4. *The spectral projector onto the eigenspace of* M *associated with the simple eigenvalue* k *is* $I - P$, *where* P *is defined by* (2.21).

Proof. Let us introduce the spectral projector $P_{\text{int}} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ of M associated with the eigenvalue k, i.e.,

$$\forall v \in \mathbb{R}^{\mathcal{N}}, \quad P_{\text{int}} v = \int_{\mathcal{C}_k} (z - M)^{-1} v \, dz,$$

where $\mathcal{C}_k$ is a closed contour in the complex plane such that k is the only eigenvalue of M contained inside the contour. Let us show that $P_{\text{int}} = \dfrac{u(\tilde{u}^*)^T}{\langle u, \tilde{u}^* \rangle}$. As the eigenvalue k is simple, it holds that Ran $P_{\text{int}} = \text{Span}\{u\}$, and $P_{\text{int}}^T \tilde{u}^* = \tilde{u}^*$ by noting that P_{int}^T is the spectral projector associated with M^T and the eigenvalue k. Let us show that Ker $P_{\text{int}} = (\text{Span}\{\tilde{u}^*\})^\perp$. Indeed, for all $v \in \mathbb{R}^{\mathcal{N}}$,

$$P_{\text{int}} v = 0 \iff \langle \tilde{u}^*, v \rangle = \langle P_{\text{int}}^T \tilde{u}^*, v \rangle = \langle \tilde{u}^*, P_{\text{int}} v \rangle = 0.$$

As Ran P_{int} + Ker $P_{\text{int}} = \mathbb{R}^{\mathcal{N}}$, we have $\text{Span}\{u\} + [\text{Span}\{\tilde{u}^*\}]^\perp = \mathbb{R}^{\mathcal{N}}$. The identity $P_{\text{int}} = I - P$ is then an immediate consequence of this decomposition. □

Lemma 2.3.5. *There holds*

(i) $P^2 = P$;

(ii) Ker $P = \text{Span}\{u\}$, Ran $P = [\text{Span}\{\tilde{u}^*\}]^\perp$ *and these two spaces are invariant by* P *and* M;

(iii) $MP = PM$.

Proof. (i) Let $v \in \mathbb{R}^{\mathcal{N}}$. Noting that $Pu = 0$, there holds

$$P^2 v = P\left(v - \frac{\langle v, \tilde{u}^* \rangle}{\langle u, \tilde{u}^* \rangle} u\right) = Pv - \frac{\langle v, \tilde{u}^* \rangle}{\langle u, \tilde{u}^* \rangle} Pu = Pv,$$

hence $P^2 = P$.

(ii) The proof of the fact that Ker $P = \text{Span}\{u\}$ and Ran $P = [\text{Span}\{\tilde{u}^*\}]^\perp$ is immediate from the proof of the previous lemma. The fact that Ker P is invariant by P and M is also obvious. Now, let $v \in \text{Span}\{\tilde{u}^*\}^\perp$, i.e. such that $\langle \tilde{u}^*, v \rangle = 0$. Then

$$\langle \tilde{u}^*, Pv \rangle = \langle \tilde{u}^*, v \rangle - \langle \tilde{u}^*, v \rangle \frac{\langle \tilde{u}^*, u \rangle}{\langle \tilde{u}^*, u \rangle} = 0,$$

and

$$\langle \tilde{u}^*, Mv \rangle = \langle M^T \tilde{u}^*, v \rangle = k \langle \tilde{u}^*, v \rangle = 0.$$

Therefore $Pv \in \text{Span}\{\tilde{u}^*\}^\perp$ and $Mv \in \text{Span}\{\tilde{u}^*\}^\perp$.

(iii) It is obvious that for all $v \in$ Ker P, $PMv = MPv = 0$. Besides, for all $v \in$ Ran P, it holds that $Pv = v$, and $Mv \in$ Ran P from (ii) so that $PMv = Mv$. As a consequence, $PMv = Mv = MPv$ for any $v \in \mathbb{R}^{\mathcal{N}}$, hence the desired result.

□

It is easy to check that the following Lemma holds for P^*, using similar arguments as in the proof of Lemma 2.3.5.

Lemma 2.3.6. *There holds*

(i) $(P^*)^2 = P^*$;

(ii) $\mathrm{Ker}\ P^* = \mathrm{Span}\{\tilde{u}^*\}$, $\mathrm{Ran}\ P^* = [\mathrm{Span}\{u\}]^\perp$ *and these two spaces are invariant by* P^* *and* M^T;

(iii) $M^T P^* = P^* M^T$.

Let us introduce some additional notation. By Lemma 2.3.5, the operator $PMP - k_N I$ leaves $\mathrm{Ran}\ \mathrm{P} = [\mathrm{Span}\{\tilde{u}^*\}]^\perp$ invariant. Moreover, provided that $k_N \notin \sigma\left(PMP|_{[\mathrm{Span}\{\tilde{u}^*\}]^\perp}\right)$, the operator $(PMP - k_N I)\,|_{[\mathrm{Span}\{\tilde{u}^*\}]^\perp}$ is invertible, since it is an operator from $[\mathrm{Span}\{\tilde{u}^*\}]^\perp$ onto $[\mathrm{Span}\{\tilde{u}^*\}]^\perp$. We can thus define the Moore–Penrose inverse of this operator, denoted by $(PMP - k_N I)^+$ as follows

$$\begin{aligned} \forall v \in [\mathrm{Span}\{\tilde{u}^*\}]^\perp, (PMP - k_N I)^+ v &= (PMP - k_N I)\,|^{-1}_{[\mathrm{Span}\{\tilde{u}^*\}]^\perp} v, \\ \forall v \in \mathrm{Span}\{u\}, (PMP - k_N I)^+ v &= 0. \end{aligned}$$

We define in a similar way the operator $\left(P^* M^T P^* - k_N I\right)^+$.

Proposition 2.3.7 (Eigenvector error estimates)**.** *Let* $u_N, u_N^* \in \mathbb{R}^{\mathcal{N}} \setminus \{0\}$ *and let* $k_N \in \mathbb{R}$ *such that* $k_N \notin \sigma((PMP)|_{[\mathrm{Span}\{\tilde{u}^*\}]^\perp})$ *and* $k_N \notin \sigma((P^* M^T P^*)|_{[\mathrm{Span}\{u\}]^\perp}$. *Then, the following estimates hold:*

$$\inf_{v \in \mathrm{Span}\{u\}} \|u_N - v\| \leq C_N^u \|R_N\|, \tag{2.23}$$

$$\inf_{v^* \in \mathrm{Span}\{u^*\}} \|u_N^* - v^*\| \leq C_N^{u^*} \|R_N^*\|, \tag{2.24}$$

with

$$\begin{aligned} C_N^u &:= \left\| P\left(PMP - k_N I\right)^+ PA^{-1} \right\|, \\ C_N^{u^*} &:= \left\| A^{-T} P^* \left(P^* M^T P^* - k_N I\right)^+ P^* \right\|. \end{aligned}$$

Here and in what follows, with a slight abuse of notation, we denote by $\|\cdot\|$ the operator norm associated with the vector norm $\|\cdot\|$ on $\mathbb{R}^{\mathcal{N}}$.

Remark 2.3.8. *The notations* C_N^u *and* $C_N^{u^*}$ *allow to introduce prefactors in the upper bound of the error on the eigenvectors* u *and* u^**, respectively.*

Remark 2.3.9. *Note that* u_N, u_N^* *and* k_N *do not have to be respectively related to* u*,* u^* *and* k *for the above estimates to hold. However, in practice,* u_N *will be an approximation of* u*,* u_N^* *will be an approximation of* u^* *and* k_N *will be an approximation of* k*, so that the norms of the residuals* $\|R_N\|$ *and* $\|R_N^*\|$ *will be small.*

Remark 2.3.10. *The results obtained in Proposition 2.3.7 match those of [61, Proposition 4] for* $k = 0$*, with a slightly different definition of the residual to take into account the generalized eigenvalue problem.*

Proof of Proposition (2.3.7). First,

$$\inf_{v \in \mathrm{Span}\{u\}} \|u_N - v\| \leq \|P u_N\|.$$

Second, let us show that $P\left(PMP - k_N I\right)^+ \left(PMP - k_N I\right) P = P$. Indeed for $v \in \mathrm{Span}\{u\}$,
$P\left(PMP - k_N I\right)^+ \left(PMP - k_N I\right) Pv = 0$ and $Pv = 0$. Moreover, for $v \in [\mathrm{Span}\{\tilde{u}^*\}]^\perp$, $Pv = v$, and $\left(PMP - k_N I\right) Pv \in [\mathrm{Span}\{\tilde{u}^*\}]^\perp$ from Lemma 2.3.5 (ii). As a consequence, since $k_N \notin \sigma((PMP)_{|[\mathrm{Span}\tilde{u}^*]^\perp}$, $\left(PMP - k_N I\right)$ is invertible on $[\mathrm{Span}\{\tilde{u}^*\}]^\perp$. Hence for $v \in [\mathrm{Span}\{\tilde{u}^*\}]^\perp$,

$$\begin{aligned} P\left(PMP - k_N I\right)^+ \left(PMP - k_N I\right) Pv &= P\left(PMP - k_N I\right)|^{-1}_{\mathrm{Span}\{\tilde{u}^*\}} \left(PMP - k_N I\right)|_{\mathrm{Span}\{\tilde{u}^*\}} Pv \\ &= Pv. \end{aligned}$$

We conclude this part of the proof by noting that $\mathbb{R}^{\mathcal{N}} = \mathrm{Span}\{u\} + [\mathrm{Span}\{\tilde{u}^*\}]^\perp$.

Then using Lemma 2.3.5 (iii), we have

$$\begin{aligned} Pu_N &= P\left(PMP - k_N I\right)^+ \left(PMP - k_N I\right) Pu_N \\ &= P\left(PMP - k_N I\right)^+ P\left(M - k_N I\right) u_N. \end{aligned}$$

Using (2.17), we obtain

$$Pu_N = P\left(PMP - k_N I\right)^+ PA^{-1} R_N. \tag{2.25}$$

Thus,

$$\|Pu_N\| \leqslant \left\|P\left(PMP - k_N I\right)^+ PA^{-1}\right\| \|R_N\|. \tag{2.26}$$

To show the second bound, we first note that

$$P^*\left(A^T u_N^*\right) = A^T u_N^* - \frac{\langle u, A^T u_N^* \rangle}{\langle u, \tilde{u}^* \rangle} \tilde{u}^*,$$

so that (2.19) and (2.22) yield

$$\begin{aligned} \inf_{v^* \in \mathrm{Span}\{u^*\}} \|u_N^* - v^*\| &= \inf_{v^* \in \mathrm{Span}\{u^*\}} \left\|A^{-T}\left(A^T u_N^* - A^T v^*\right)\right\| \\ &= \inf_{\tilde{v}^* \in \mathrm{Span}\{\tilde{u}^*\}} \left\|A^{-T}\left(A^T u_N^* - \tilde{v}^*\right)\right\| \\ &\leqslant \left\|A^{-T} P^* A^T u_N^*\right\|. \end{aligned}$$

Now, using Lemma 2.3.6 (iii) and similar arguments as above, we get

$$\begin{aligned} P^*\left(A^T u_N^*\right) &= P^*\left(P^* M^T P^* - k_N I\right)^+ \left(P^* M^T P^* - k_N I\right) P^* A^T u_N^* \\ &= P^*\left(P^* M^T P^* - k_N I\right)^+ P^*\left(M^T - k_N I\right) A^T u_N^* \\ &= P^*\left(P^* M^T P^* - k_N I\right)^+ P^*\left(B - k_N A\right)^T u_N^*. \end{aligned}$$

Hence

$$P^*\left(A^T u_N^*\right) = P^*\left(P^* M^T P^* - k_N I\right)^+ P^* R_N^*. \tag{2.27}$$

Then,

$$\left\|A^{-T} P^* A^T u_N^*\right\| \leqslant \left\|A^{-T} P^*\left(P^* M^T P^* - k_N I\right)^+ P^*\right\| \|R_N^*\|,$$

which proves (2.24). □

2.3.2 Error estimate on the eigenvalue

We now provide an estimate for the eigenvalue error.

Proposition 2.3.11 (Eigenvalue error estimate)**.** *Let $u_N, u_N^* \in \mathbb{R}^{\mathcal{N}}$. Under Assumption 2.3.1 and the fact that $k_N := \dfrac{\langle u_N^*, Bu_N\rangle}{\langle u_N^*, Au_N\rangle}$ is such that $k_N \notin \sigma((PMP)|_{[\mathrm{Span}\{\tilde{u}^*\}]^\perp})$ and $k_N \notin \sigma((P^*M^TP^*)|_{[\mathrm{Span}\{u\}]^\perp}$, there holds*

$$|k_N - k| \leq C_N^k \eta_N^k, \tag{2.28}$$

where

$$\eta_N^k := \frac{\|R_N\|\|R_N^*\|}{|\langle u_N^*, Au_N\rangle|}, \tag{2.29}$$

and

$$C_N^k := \left\| [P^* \left(P^*M^TP^* - k_N I\right)^+ P^*]^T (M - kI) P \left(PMP - k_N I\right)^+ PA^{-1} \right\|. \tag{2.30}$$

Remark 2.3.12. *The notations C_N^k and η_N^k allow to respectively refer to a prefactor in the upper bound and a residual-based error estimator for the eigenvalue k.*

Remark 2.3.13. *Note that in this result, the vectors u_N and u_N^* may not be solutions of a reduced eigenvalue problem of the form (3.6) or (3.7). The only requirement of Proposition 2.3.11 is that k_N has to be related to u_N and u_N^* by the formula stated in the proposition.*

Proof of Proposition 2.3.11. For any α, $\beta \in \mathbb{R}$,

$$\begin{aligned}\langle A^T \left(u_N^* - \alpha u^*\right), \left(M - kI\right)\left(u_N - \beta u\right)\rangle = {} & \langle A^T u_N^*, Mu_N\rangle - \beta\langle A^T u_N^*, Mu\rangle - \alpha\langle A^T u^*, Mu_N\rangle \\ & + \alpha\beta\langle A^T u^*, Mu\rangle - k\langle A^T u_N^*, u_N\rangle + \beta k\langle A^T u_N^*, u\rangle \\ & + \alpha k\langle A^T u^*, u_N\rangle - \alpha\beta k\langle A^T u^*, u\rangle.\end{aligned}$$

Noting that $Mu = ku$, $M^TA^Tu^* = kA^Tu^*$ and recalling that $M = A^{-1}B$, we obtain

$$\begin{aligned}\langle A^T \left(u_N^* - \alpha u^*\right), \left(M - kI\right)\left(u_N - \beta u\right)\rangle &= \langle A^T u_N^*, Mu_N\rangle - k\langle u_N^*, Au_N\rangle \\ &= \left(k_N - k\right)\langle u_N^*, Au_N\rangle.\end{aligned}$$

According to Lemma 2.3.2, we can set

$$\alpha = \frac{1}{\|A^Tu^*\|}\frac{\langle A^Tu_N^*, u\rangle}{\langle \tilde{u}^*, u\rangle}, \quad \beta = \frac{\langle u_N, \tilde{u}^*\rangle}{\langle u, \tilde{u}^*\rangle},$$

and show that

$$k_N - k = \frac{1}{\langle u_N^*, Au_N\rangle}\langle P^*\left(A^Tu_N^*\right), \left(M - kI\right)Pu_N\rangle.$$

Using (2.25) and (2.27) finishes the proof. □

2.4 Practical *a posteriori* error estimators

In view of Proposition 2.3.7 and Proposition 2.3.11, it is natural to estimate the actual errors $e_N^k = |k - k_N|$, $e_N^u = \|u_{\mu,N} - u_\mu\|$ and $e_N^{u^*} = \left\|u_{\mu,N}^* - u_\mu^*\right\|$ by the quantities respectively defined as follows,

$$\Delta_N^k := \overline{C}_N^k \eta_N^k, \qquad \Delta_N^u := \overline{C}_N^u \|R_N\|, \qquad \Delta_N^{u^*} := \overline{C}_N^{u^*} \|R_N^*\|, \tag{2.31}$$

where $\overline{C}_N^k$, $\overline{C}_N^u$ and $\overline{C}_N^{u^*}$ are some constants which are good estimates of the efficiencies $\dfrac{e_N^k}{\eta_N^k(\mu)}$, $\dfrac{e_N^u}{\|R_N(\mu)\|}$, and $\dfrac{e_N^{u^*}}{\|R_N^*(\mu)\|}$. For example, one could use practical (computable) estimations of the constants C_N^k, C_N^u and $C_N^{u^*}$ appearing in Proposition 2.3.7 and Proposition 2.3.11.

For the applications we are interested in, as will be illustrated in Section 4.1, we observe that the operators are perturbations of symmetric operators. This is why we investigate in Section 2.4.2 the links between the prefactor C_N^k introduced above and its well-known counterpart in the symmetric case. However, this does not yield practical formulas for estimating/computing the prefactors. This is why, in Section 2.4.3, we propose a practical heuristic approach to compute some prefactors $\overline{C}_N^k$, $\overline{C}_N^u$ and $\overline{C}_N^{u^*}$ in the reduced basis context, that we will use in the numerical results to build practical *a posteriori* error estimators in the greedy algorithm to select the reduced space (see Section 3.3). This heuristic approach gives very interesting numerical results for neutronics applications as will be illustrated in Chapter 4.

Let us briefly go back to the symmetric case, i.e. in the case where A is a positive definite symmetric matrix, and $B = I$, i.e., $M = A^{-1}$. In this case, all the eigenvalues of M are real and positive, with k being a largest one. We still assume that k is a simple eigenvalue of M and denote by k_2 the second largest eigenvalue of M so that $k > k_2$. Note that in the symmetric case, $u = u^*$ and $P = P^* = P^T$. As a consequence, for a given vector u_N and the value $k_N = \dfrac{\langle u_N, Bu_N\rangle}{\langle u_N, Au_N\rangle} > 0$, we have (from (2.30))

$$\begin{aligned} C_N^k &= \left\|\left[P\left(PMP - k_N I\right)^+ P\right]^T (M - kI) P\left(PMP - k_N I\right)^+ PA^{-1}\right\| \\ &= \left\|P\left(PMP - k_N I\right)^+ P(M - kI) P\left(PMP - k_N I\right)^+ PA^{-1}\right\|. \end{aligned}$$

In the spirit of Proposition 2.2.3, we get the following result.

Proposition 2.4.1. *Let A be symmetric positive definite and $B = I$. Let k be the largest eigenvalue of $M = A^{-1}$, let us assume that it is simple, and let us denote by k_2 its second largest eigenvalue. Let us also assume that*

$$k > k_N > k_2 > 0. \tag{2.32}$$

Then,

$$C_N^k = \frac{k_2(k - k_2)}{(k_N - k_2)^2} = \frac{\|PA^{-1}\|\|M - kI\|}{\operatorname{dist}(k_N, \sigma((PMP)|_{\operatorname{Span}\{u\}^\perp}))^2}. \tag{2.33}$$

Proof. Let $k^{(1)} = k > k^{(2)} = k_2 \geq \cdots \geq k^{(\mathcal{N})}$ denote the eigenvalues of $M = A^{-1}$ and $u^{(1)}, u^{(2)}, \ldots, u^{(\mathcal{N})}$ be corresponding eigenvectors,

$$A^{-1} = M = \sum_{i=1}^{\mathcal{N}} k^{(i)} u^{(i)} u^{(i)^T},$$

and

$$P = \sum_{i=2}^{\mathcal{N}} u^{(i)} u^{(i)^T}.$$

Then, using functional calculus, there yields

$$P\left(PMP - k_N I\right)^+ P(M - kI) P\left(PMP - k_N I\right)^+ PA^{-1} = \sum_{i=2}^{\mathcal{N}} \frac{k^{(i)}(k - k^{(i)})}{(k_N - k^{(i)})^2} u_i u_i^T.$$

Since the operator norm is associated with the Euclidean vector norm, the operator norm corresponds to the largest eigenvalue of the (symmetric) operator, such that

$$C_N^k = \max_{2 \leq i \leq \mathcal{N}} \frac{k^{(i)}(k - k^{(i)})}{(k_N - k^{(i)})^2} = \frac{k_2(k - k_2)}{(k_N - k_2)^2}.$$

Since $\|PA^{-1}\| = k_2$, $\|M - kI\| = k - k_2$ using (2.32), and $\operatorname{dist}(k_N, \sigma((PMP)|_{\operatorname{Span}\{u\}^\perp}))^2 = (k_N - k_2)^2$, we easily obtain the second equality. □

The constant C_N^k is therefore strongly linked to the spectral gap between the first and second eigenvalue of M in this particular symmetric case. However, this notion of spectral gap is not clear in the non-symmetric context and we provide two points of view which enable to draw a comparison between the symmetric and non-symmetric case.

2.4.1 Numerical range

In this section, we prove that in the general non-symmetric case, the value of the prefactor C_N^k can be estimated using the so-called numerical range [99] of the non-symmetric operator. In this work, the computation of the numerical range was not investigated, but we refer to [22] for some guidelines. Let us first define the notion of numerical range.

Definition 2.4.2 (Numerical range)**.** *Let $Q \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$. The numerical range of the matrix Q is defined by*

$$\operatorname{Num}(Q) = \overline{\{\langle v, Qv \rangle,\ \|v\| = 1\}}.$$

Lemma 2.4.3. *Let $Q \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ let and $z \notin \sigma(Q)$. Then,*

$$\left\|(Q - zI)^{-1}\right\| \leqslant \frac{1}{\operatorname{dist}\left(z, \operatorname{Num}\left(Q\right)\right)}.$$

Proof. Let $w \in \mathbb{R}^{\mathcal{N}}$ be a unit vector. Then,

$$\begin{aligned}
\operatorname{dist}\left(z, \operatorname{Num}\left(Q\right)\right) &\leqslant \left| z - \frac{\langle w, Qw \rangle}{\|w\|^2} \right| \\
&\leqslant \frac{|\langle w, \left(Q - zI\right) w \rangle|}{\|w\|^2} \\
&\leqslant \frac{\|\left(Q - zI\right) w\|}{\|w\|}.
\end{aligned}$$

Taking $u = \left(Q - zI\right) w$, then,

$$\left\|(Q - zI)^{-1}\right\| \leqslant \frac{1}{\operatorname{dist}\left(z, \operatorname{Num}\left(Q\right)\right)}.$$

□

Proposition 2.4.4. *Under the same assumptions as in Proposition 2.3.11,*

$$C_N^k \leq \frac{\|M - kI\| \|PA^{-1}\|}{\operatorname{dist}(k_N, \operatorname{Num}((PMP)_{|[\operatorname{Span}\{\tilde{u}^*\}]^\perp})) \operatorname{dist}(k_N, \operatorname{Num}(P^*M^TP^*)_{|[\operatorname{Span}\{u\}]^\perp}))}$$

Note that the bound given in Proposition 2.4.4 is exactly equal to the value of the prefactor in the symmetric case since, when M is symmetric and non-negative, $\operatorname{Num}((PMP)_{|[\operatorname{Span}\{\tilde{u}^*\}]^\perp}) = \operatorname{Num}(P^*M^TP^*)_{|[\operatorname{Span}\{u\}]^\perp} = [k_2, k_{\mathcal{N}}]$ where $k > k_2 \geq \ldots \geq k_{\mathcal{N}}$ are the ordered eigenvalues of M.

Proof. Starting from (2.30), it holds that

$$C_N^k \leq \left\| P^* \left(P^*M^TP^* - k_N I\right)^+ P^* \right\| \|M - kI\| \left\| P \left(PMP - k_N I\right)^+ P \right\| \|PA^{-1}\|.$$

Using Lemma 2.4.3, this completes the proof. □

The upper bound on C_N^k derived in Proposition 2.4.4 goes to infinity if k_N (which is supposed to be an approximation of k) gets close to $\operatorname{Num}((PMP)_{|[\operatorname{Span}\{\tilde{u}^*\}]^\perp})$ or $\operatorname{Num}(P^*M^TP^*)_{|[\operatorname{Span}\{u\}]^\perp}$, which can be seen as an underlying spectral gap condition.

2.4.2 A perturbative approach

The aim of this section is to propose another connection between the estimation of the prefactor C_N^k in the non-symmetric case with its well-known expression in the symmetric case. In all this section, we assume that

$$\begin{cases} A := A^\varepsilon = S + \varepsilon T \text{ with } S^T = S,\ T^T = -T,\ \varepsilon > 0, \\ B := I. \end{cases} \tag{2.34}$$

In other words, the matrix A is a perturbation of a symmetric positive definite matrix $S \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$, since $\varepsilon > 0$ is intended to be a small parameter. We still assume here that $B = I$ for the sake of simplicity.

We also assume that the positive definite symmetric matrix S has a simple positive lowest eigenvalue λ_S, and that u_S is an associated eigenvector. We then denote by $\lambda_S < \lambda_{S,2} \leq \ldots \leq \lambda_{S,\mathcal{N}}$ all the eigenvalues of S. We also define $k_S := \frac{1}{\lambda_S}$ and $k_{S,i} := \frac{1}{\lambda_{S,i}}$ for $2 \leq i \leq \mathcal{N}$. By a perturbative argument, for any $\varepsilon > 0$ small enough, there exists a simple nonzero eigenvalue λ^ε of smallest modulus of A^ε, and we denote by u^ε an associated right eigenvector, $u^{*,\varepsilon}$ an associated left eigenvector, $\tilde{u}^{*,\varepsilon}$ defined as in (2.19) and $k^\varepsilon := \frac{1}{\lambda^\varepsilon}$.

For the sake of simplicity of the perturbative analysis, we assume that the approximate value k_N is independent of ε. This for example makes sense if one uses a reduced-order model constructed from the one-dimensional reduced space $\mathcal{V} = \operatorname{Span}\{u_S\}$. In that case, $u_N = u_N^* = u_S$ and thus $k_N = k_S$.

In this section, using obvious notation, we would like to study the convergence of the prefactor

$$C_N^{k,\varepsilon} := \left\| [P^{*,\varepsilon} \left(P^{*,\varepsilon}(M^\varepsilon)^T P^{*,\varepsilon} - k_N I\right)^+ P^{*,\varepsilon}]^T (M^\varepsilon - k^\varepsilon I) P^\varepsilon \left(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I\right)^+ P^\varepsilon (A^\varepsilon)^{-1} \right\|$$

to the value

$$C_N^{k,\mathrm{sym}} = \frac{k_{S,2}(k_S - k_{S,2})}{(k_N - k_{S,2})^2} \tag{2.35}$$

as ε goes to 0.

We first perform a first-order expansion of the eigenvectors and eigenvalues in ε (cf. Chapter 2 of [76]).

Lemma 2.4.5. *Let us assume* (2.34) *and*

$$\|u^\varepsilon\|^2 = \|u_S\|^2 = 1 \quad \text{and} \quad \langle u^\varepsilon, u_S\rangle > 0. \tag{2.36}$$

Then, as ε *goes to* 0,

$$\begin{aligned}
\lambda^\varepsilon &= \lambda_S + O(\varepsilon^2),\\
k^\varepsilon &= k_S + O(\varepsilon^2),\\
u^\varepsilon &= u_S - \varepsilon \left(S - \lambda_S I\right)^{-1}_{|\mathrm{Span}\{u_S\}^\perp} T u_S + O(\varepsilon^2),\\
u^{*,\varepsilon} &= u_S + \varepsilon \left(S - \lambda_S I\right)^{-1}_{|\mathrm{Span}\{u_S\}^\perp} T u_S + O(\varepsilon^2),\\
\tilde{u}^{*,\varepsilon} &= u_S + \varepsilon \left(S - \lambda_S I\right)^{-1}_{|\mathrm{Span}\{u_S\}^\perp} T u_S + O(\varepsilon^2).
\end{aligned}$$

Proof. Using the results of [76, Chapter 2], we decompose λ^ε, u^ε, and $u^{*,\varepsilon}$ at first order as

$$\begin{aligned}
\lambda^\varepsilon &= \lambda_{A,0} + \varepsilon\lambda_{A,1} + O(\varepsilon^2),\\
u^\varepsilon &= u_{A,0} + \varepsilon u_{A,1} + O(\varepsilon^2),\\
u^{*,\varepsilon} &= u^*_{A,0} + \varepsilon u^*_{A,1} + O(\varepsilon^2).
\end{aligned}$$

Using this decomposition, the eigenvalue problem writes

$$Su_{A,0} + \varepsilon(Su_{A,1} + Tu_{A,0}) = \lambda_{A,0}u_{A,0} + \varepsilon(\lambda_{A,0}u_{A,1} + \lambda_{A,1}u_{A,0}) + O(\varepsilon^2).$$

At order 0 in ε, we obtain $u_{A,0} = u_S$ and $\lambda_{A,0} = \lambda_S$. Then, at first order,

$$Su_{A,1} + Tu_{A,0} = \lambda_{A,0}u_{A,1} + \lambda_{A,1}u_{A,0}. \tag{2.37}$$

Using (2.36), one can write

$$\|u^\varepsilon\|^2 = \|u_{A,0}\|^2 + \varepsilon(\langle u_{A,0}, u_{A,1}\rangle + \langle u_{A,1}, u_{A,0}\rangle) + O(\varepsilon^2),$$

which implies that

$$\langle u_{A,0}, u_{A,1}\rangle = 0. \tag{2.38}$$

Using the latter and projecting (2.37) onto $u_{A,0}$ gives

$$\langle Su_{A,1}, u_{A,0}\rangle + \langle Tu_{A,0}, u_{A,0}\rangle = \lambda_{A,1}\langle u_{A,0}, u_{A,0}\rangle = \lambda_{A,1}.$$

As T is skew-symmetric, it holds $\langle Tu_{A,0}, u_{A,0}\rangle = 0$, so that

$$\langle Su_{A,1}, u_{A,0}\rangle = \langle u_{A,1}, Su_{A,0}\rangle = \lambda_{A,0}\langle u_{A,1}, u_{A,0}\rangle = 0.$$

Hence $\lambda_{A,1} = 0$. Then (2.37) transforms into

$$\left(S - \lambda_S I\right) u_{A,1} = -Tu_S.$$

The latter has a solution since $Tu_S \in \mathrm{Span}\{u_S\}^\perp$ and $\left(\mathrm{Ker}\left(S - \lambda_S I\right)\right)^\perp = \mathrm{Ran}\left(S - \lambda_S I\right)$. Hence

$$u_{A,1} = -\left(S - \lambda_S I\right)^{-1}_{|\mathrm{Span}\{u_S\}^\perp} Tu_S. \tag{2.39}$$

We can apply the same procedure for the adjoint eigenvector to obtain the result. Finally,

$$\begin{aligned}\tilde{u}^{*,\varepsilon} &:= \frac{(A^\varepsilon)^T u^{*,\varepsilon}}{\|(A^\varepsilon)^T u^{*,\varepsilon}\|} \\ &= \frac{(S-\varepsilon T)(u_S - \varepsilon u_{A,1})}{\|(S-\varepsilon T)(u_S-\varepsilon u_{A,1})\|} + O(\varepsilon^2) \\ &= \frac{\lambda_S u_S - \varepsilon(Su_{A,1}+Tu_S)}{\|\lambda_S u_S - \varepsilon(Su_{A,1}+Tu_S)\|} + O(\varepsilon^2) \\ &= \left(u_S - \frac{\varepsilon}{\lambda_S}(Su_{A,1}+Tu_S)\right)\left(1+\frac{\varepsilon}{\lambda_S}\langle u_S,(Su_{A,1}+Tu_S)\rangle\right) + O(\varepsilon^2) \\ &= u_S - \varepsilon u_{A,1} + O(\varepsilon^2),\end{aligned}$$

which concludes the proof. □

We now provide first-order expansions of operators which will be needed in the subsequent estimation of the prefactor.

Lemma 2.4.6. *Let us assume* (2.34) *and* (2.36). *Then, as* ε *goes to* 0,

$$P^\varepsilon = P_S + \varepsilon P_T + O(\varepsilon^2), \quad \textit{and} \quad P^{*,\varepsilon} = P_S - \varepsilon P_T + O(\varepsilon^2),$$

where

$$P_S = I - u_S u_S^T, \quad \textit{and} \quad P_T = u_{A,1}u_S^T - u_S u_{A,1}^T,$$

$u_{A,1}$ *being defined in* (2.39).

Proof. We have

$$\begin{aligned}P^\varepsilon &= I - \frac{u^\varepsilon(\tilde{u}^{*,\varepsilon})^T}{\langle u^\varepsilon, \tilde{u}^{*,\varepsilon}\rangle}, \\ &= I - (u_S + \varepsilon u_{A,1} + O(\varepsilon^2))(u_S - \varepsilon u_{A,1} + O(\varepsilon^2))^T \\ &= I - u_S u_S^T + \varepsilon(u_{A,1}u_S^T - u_S u_{A,1}^T) + O(\varepsilon^2).\end{aligned}$$

Similarly,

$$\begin{aligned}P^{*,\varepsilon} &= I - \frac{\tilde{u}^{*,\varepsilon}(u^\varepsilon)^T}{\langle u^\varepsilon, \tilde{u}^{*,\varepsilon}\rangle} \\ &= I - (u_S - \varepsilon u_{A,1} + O(\varepsilon^2))(u_S + \varepsilon u_{A,1} + O(\varepsilon^2))^T \\ &= I - u_S u_S^T + \varepsilon(-u_{A,1}u_S^T + u_S u_{A,1}^T) + O(\varepsilon^2).\end{aligned}$$

This concludes the proof. □

We now provide a first order expansion of the operator entering the prefactor in (2.30), namely

$$\mathcal{M}^\varepsilon = [P^\varepsilon\,(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I)^+\,P^\varepsilon](M^\varepsilon - k^\varepsilon I)P^\varepsilon\,(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I)^+\,P^\varepsilon (A^\varepsilon)^{-1}. \tag{2.40}$$

Lemma 2.4.7. *Let us assume* (2.34) *and* (2.36). *Then, as* ε *goes to* 0,

$$\mathcal{M}^\varepsilon = \mathcal{M}_0 + \varepsilon\mathcal{M}_1 + O(\varepsilon^2), \tag{2.41}$$

with

$$\begin{aligned}\mathcal{M}_0 &= \Gamma_S(S^{-1} - k_S I)\Gamma_S S^{-1},\\ \mathcal{M}_1 &= -\Gamma_S S^{-1}TS^{-1}\Gamma_S S^{-1} + \Gamma_S(S^{-1} - k_S I)\Gamma_T S^{-1}\\ &\quad - \Gamma_S(S^{-1} - k_S I)\Gamma_S S^{-1}TS^{-1} + \Gamma_T(S^{-1} - k_S I)\Gamma_S S^{-1},\end{aligned}$$

and

$$\begin{aligned}\Gamma_S &= P_S\left(P_S S^{-1}P_S - k_N I\right)^+ P_S,\\ \Gamma_T &= P_S\left(P_S S^{-1}P_S - k_N I\right)^+ P_T + P_T\left(P_S S^{-1}P_S - k_N I\right)^+ P_S\\ &\quad - P_S\left(P_S S^{-1}P_S - k_N I\right)^+ \left(P_S S^{-1}TS^{-1}P_S + P_T S^{-1}P_S + P_S S^{-1}P_T\right)\\ &\quad \left(P_S S^{-1}P_S - k_N I\right)^+ P_S.\end{aligned}$$

Proof. First,

$$M^\varepsilon = (A^\varepsilon)^{-1} = S^{-1} - \varepsilon S^{-1}TS^{-1} + O(\varepsilon^2).$$

Therefore, as we have $P^\varepsilon = P_S + \varepsilon P_T + O(\varepsilon^2)$, there holds

$$\begin{aligned}P^\varepsilon M^\varepsilon P^\varepsilon &= \left(P_S + \varepsilon P_T + O(\varepsilon^2)\right)\left(S^{-1} - \varepsilon S^{-1}TS^{-1} + O(\varepsilon^2)\right)\left(P_S + \varepsilon P_T + O(\varepsilon^2)\right)\\ &= \left(P_S S^{-1} - \varepsilon P_S S^{-1}TS^{-1} + \varepsilon P_T S^{-1} + O(\varepsilon^2)\right)\left(P_S + \varepsilon P_T + O(\varepsilon^2)\right)\\ &= P_S S^{-1}P_S - \varepsilon P_S S^{-1}TS^{-1}P_S + \varepsilon P_T S^{-1}P_S + \varepsilon P_S S^{-1}P_T + O(\varepsilon^2)\\ &= P_S S^{-1}P_S + \varepsilon\left(P_S S^{-1}TS^{-1}P_S + P_T S^{-1}P_S + P_S S^{-1}P_T\right) + O(\varepsilon^2).\end{aligned}$$

Using a first-order expansion of the pseudo-inverse in ε, there holds

$$\begin{aligned}(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I)^+ &= \left[\left(P_S S^{-1}P_S - k_N I\right) + \varepsilon\left(P_S S^{-1}TS^{-1}P_S + P_T S^{-1}P_S + P_S S^{-1}P_T\right)\right.\\ &\qquad \left.+O(\varepsilon^2)\right]^+\\ &= \left(P_S S^{-1}P_S - k_N I\right)^+\\ &\quad - \varepsilon\left(P_S S^{-1}P_S - k_N I\right)^+\left(P_S S^{-1}TS^{-1}P_S + P_T S^{-1}P_S + P_S S^{-1}P_T\right)\left(P_S S^{-1}P_S - k_N I\right)^+\\ &\quad + O(\varepsilon^2).\end{aligned}$$

Hence, one can write

$$\begin{aligned}P^\varepsilon\left(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I\right)^+ P^\varepsilon &= P_S\left(P_S S^{-1}P_S - k_N I\right)^+ P_S + \varepsilon\Bigg[P_S\left(P_S S^{-1}P_S - k_N I\right)^+ P_T\\ &- P_S\left(P_S S^{-1}P_S - k_N I\right)^+\left(P_S S^{-1}TS^{-1}P_S + P_T S^{-1}P_S + P_S S^{-1}P_T\right)\\ &\quad \times\left(P_S S^{-1}P_S - k_N I\right)^+ P_S + P_T\left(P_S S^{-1}P_S - k_N I\right)^+ P_S\Bigg] + O(\varepsilon^2).\end{aligned}$$

Defining

$$\Gamma^\varepsilon := P^\varepsilon\left(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I\right)^+ P^\varepsilon,$$

we have just obtained that

$$\Gamma^\varepsilon = \Gamma_S + \varepsilon\Gamma_T + O(\varepsilon^2).$$

Using that

$$\mathcal{M}^\varepsilon = \Gamma^\varepsilon(M^\varepsilon - k^\varepsilon I)\Gamma^\varepsilon(A^\varepsilon)^{-1} = \left(\Gamma_S + \varepsilon\Gamma_T + O(\varepsilon^2)\right)\left(S^{-1} - k_S I - \varepsilon S^{-1}TS^{-1} + O(\varepsilon^2)\right) \times \left(\Gamma_S + \varepsilon\Gamma_T + O(\varepsilon^2)\right)\left(S^{-1} - \varepsilon S^{-1}TS^{-1} + O(\varepsilon^2)\right),$$

we easily obtain (2.41). □

We then estimate the prefactor C_N^k in the perturbative case using the previous results.

Proposition 2.4.8. *Let us assume* (2.34) *and* (2.36)*. Let us also assume that*

$$k_S \geq k_N > k_{S,2} > 0, \tag{2.42}$$

and that $k_{S,2}$ is not degenerate. Then, for $\varepsilon > 0$ sufficiently small,

$$C_N^{k,\varepsilon} = C_N^{k,\mathrm{sym}} + O(\varepsilon^2),$$

where $C_N^{k,\mathrm{sym}}$ is defined as in (2.35)*.*

Proof. Starting from (2.41), let us first note that $\mathcal{M}_0 = \Gamma_S(S^{-1} - k_S I)\Gamma_S S^{-1}$ has the same spectral decomposition as S, that is eigenvectors $u_{S,i}$ with corresponding eigenvalues

$$\begin{cases} 0 & \text{for } i = 1 \\ \frac{(k_{S,i}-k_S)k_{S,i}}{(k_{S,i}-k_N)^2} & \text{for } 2 \leq i \leq \mathcal{N}. \end{cases}$$

From this, we deduce that

$$\|\mathcal{M}_0\| = \max_{2\leq i\leq \mathcal{N}} \frac{|k_{S,i} - k_S|k_{S,i}}{|k_{S,i} - k_N|^2} = \frac{|k_{S,2} - k_S|k_{S,2}}{|k_{S,2} - k_N|^2} = C_N^{k,\mathrm{sym}}.$$

Note that the same holds for Γ_S with eigenvalues

$$\begin{cases} 0 & \text{for } i = 1 \\ \frac{1}{k_{S,i}-k_N} & \text{for } 2 \leq i \leq \mathcal{N}. \end{cases}$$

Then, since $k_{S,2}$ is a simple eigenvalue, we can write down the Taylor expansion of the spectral norm as

$$C_N^{k,\varepsilon} = \|\mathcal{M}_0 + \varepsilon\mathcal{M}_1 + O(\varepsilon^2)\| = \|\mathcal{M}_0\| + \varepsilon u_{\mathcal{M},0}^T \mathcal{M}_1 u_{\mathcal{M},0} + O(\varepsilon^2),$$

where $u_{\mathcal{M},0}$ the unit eigenvector corresponding to the largest eigenvalue of $\mathcal{M}_0$, that is $u_{\mathcal{M},0} = \pm u_{S,2}$. For simplicity, let us choose $u_{\mathcal{M},0} = u_{S,2}$.

Then

$$\begin{aligned} u_{\mathcal{M},0}^T \mathcal{M}_1 u_{\mathcal{M},0} &= -u_{\mathcal{M},0}^T \Gamma_S S^{-1}TS^{-1}\Gamma_S S^{-1} u_{\mathcal{M},0} + u_{\mathcal{M},0}^T \Gamma_S(S^{-1} - k_S I)\Gamma_T S^{-1} u_{\mathcal{M},0} \\ &\quad - u_{\mathcal{M},0}^T \Gamma_S(S^{-1} - k_S I)\Gamma_S S^{-1}TS^{-1} u_{\mathcal{M},0} + u_{\mathcal{M},0}^T \Gamma_T(S^{-1} - k_S I)\Gamma_S S^{-1} u_{\mathcal{M},0} \\ &= - \frac{k_{S,2}^3}{(k_{S,2} - k_N)^2} u_{\mathcal{M},0}^T T u_{\mathcal{M},0} + \frac{k_{S,2}(k_{S,2} - k_S)}{k_{S,2} - k_N} u_{\mathcal{M},0}^T \Gamma_T u_{\mathcal{M},0} \\ &\quad - \frac{k_{S,2}^2(k_{S,2} - k_S)}{(k_{S,2} - k_N)^2} u_{\mathcal{M},0}^T T u_{\mathcal{M},0} + \frac{k_{S,2}(k_{S,2} - k_S)}{k_{S,2} - k_N} u_{\mathcal{M},0}^T \Gamma_T u_{\mathcal{M},0} \\ &= 0, \end{aligned}$$

where we used the fact that the matrices T and Γ_T are skew-symmetric. This concludes the proof. □

We now illustrate the above bounds on a toy numerical example. Let us introduce the following matrices $S, T \in \mathbb{R}^{4\times 4}$:

$$S = \begin{pmatrix} 2000 & 0 & 0 & 0 \\ 0 & 1500 & 0 & 0 \\ 0 & 0 & 1000 & 0 \\ 0 & 0 & 0 & 0.02 \end{pmatrix}, \quad T = \frac{\|S\|}{\|T_0\|} T_0, \quad \text{with} \quad T_0 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix}.$$

It then holds that $k_S = 50$ and $k_{S,2} = 0.001$. Let us consider $k_N = k_S$. A second-order convergence of the difference $|C_N^{k,\varepsilon} - C_N^{k,\mathrm{sym}}|$ as a function of ε is depicted in Figure 2.1. This is a strong indication that the estimate of Proposition 2.4.8 is sharp.

In our practical applications of interest, we will indeed observe that the operator of reference is a perturbation of a symmetric operator, but the estimate of the prefactor $C_N^{k,\varepsilon}$ by $C_N^{k,\mathrm{sym}}$ is not sufficiently good over a large range of the values of the parameters μ, in particular because the spectral gap (see Assumption (2.42)) is not uniformly bounded from below (see Section 4.1 for a discussion). This is why we will resort to a practical heuristic method to approximate the prefactor, as explained in the next section 2.4.3.

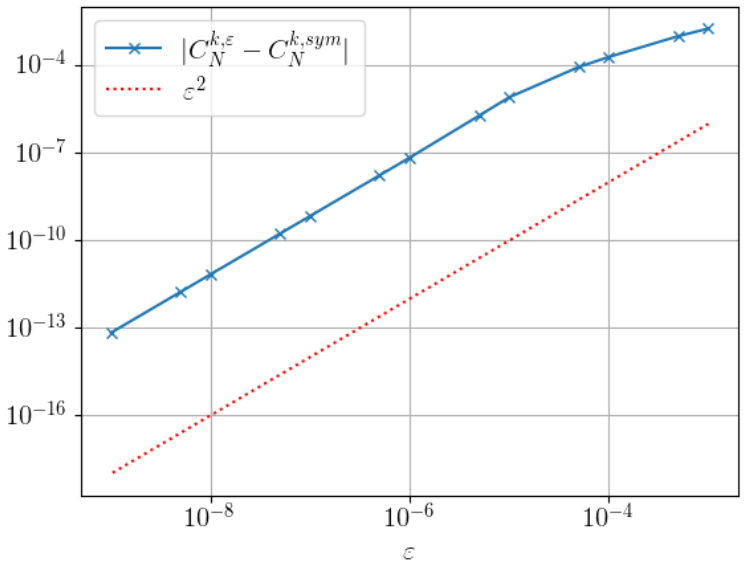

Figure 2.1: $|C_N^{k,\varepsilon} - C_N^{k,sym}|$ as a function of ε

2.4.3 Heuristic estimation of prefactors

The aim of this section is to present a heuristic algorithm to estimate the prefactors C_N^k, C_N^u and $C_N^{u^*}$ defined in Proposition 2.3.11 and Proposition 2.3.7 respectively. The algorithm then yields approximations of these constants, denoted by $\overline{C}_N^k$, $\overline{C}_N^u$ and $\overline{C}_N^{u^*}$, which are then used to build *a posteriori* error estimates for the greedy algorithm presented in Chapter 3.

This heuristic procedure is based on the use of a set of parameters values $\mathcal{P}_{\mathrm{pref}} \subset \mathcal{P}$, containing a finite number of elements, which does not contain any values of the parameters belonging to the training set $\mathcal{P}_{\mathrm{train}}$, in the context of a reduced-basis approach, as detailed in Chapter 3, i.e., $\mathcal{P}_{\mathrm{pref}}$ is chosen such that $\mathcal{P}_{\mathrm{train}} \cap \mathcal{P}_{\mathrm{pref}} = \emptyset$.

Let us introduce the efficiency ratios, i.e., for all $\mu \in \mathcal{P}$,

$$\mathcal{E}_N^k(\mu) := \frac{|k_{\mu,N} - k_\mu|}{\eta_N^k(\mu)}, \quad \mathcal{E}_N^u(\mu) := \frac{\|u_{\mu,N} - u_\mu\|}{\|R_N(\mu)\|} \text{ and } \mathcal{E}_N^{u^*}(\mu) := \frac{\left\|u^*_{\mu,N} - u^*_\mu\right\|}{\|R^*_N(\mu)\|}. \tag{2.43}$$

By definition, for all $\mu \in \mathcal{P}$,

$$\mathcal{E}_N^k(\mu) \leq C_N^k(\mu), \quad \mathcal{E}_N^u(\mu) \leq C_N^u(\mu) \text{ and } \mathcal{E}_N^{u^*}(\mu) \leq C_N^{u^*}(\mu). \tag{2.44}$$

Our heuristic approach aims at estimating the constants $C_N^k(\mu)$, $C_N^u(\mu)$ and $C_N^{u^*}(\mu)$ for all $\mu \in \mathcal{P}$ by their maximum values over $\mathcal{P}_{\text{pref}}$. More precisely, defining

$$\overline{C}_N^k := \max_{\mu \in \mathcal{P}_{\text{pref}}} \mathcal{E}_N^k(\mu), \qquad \overline{C}_N^u := \max_{\mu \in \mathcal{P}_{\text{pref}}} \mathcal{E}_N^u(\mu), \text{ and } \qquad \overline{C}_N^{u^*} := \max_{\mu \in \mathcal{P}_{\text{pref}}} \mathcal{E}_N^{u^*}(\mu), \tag{2.45}$$

the practical *a posteriori* error estimates that are used in the greedy reduced basis method, introduced in the next chapter, are then defined by

$$\Delta_N^k(\mu) := \overline{C}_N^k \eta_N^k(\mu), \qquad \Delta_N^u(\mu) := \overline{C}_N^u \|R_N(\mu)\|, \text{ and } \qquad \Delta_N^{u^*}(\mu) := \overline{C}_N^{u^*} \|R^*_N(\mu)\|. \tag{2.46}$$

The efficiency of this practical approach will be illustrated in Chapter 4, where numerical results obtained in neutronics applications are presented.

***

In this chapter, we developed residual-based *a posteriori* error estimates for a given non-self-adjoint generalized eigenvalue problem. These estimates all exhibit parameter-dependent *prefactors* that are too expensive to compute in practice for a large number of parameters. Therefore, we first provided some elements of theoretical analysis to illustrate the close link between the obtained expression of the *prefactor* and its well-known counterpart in the case of symmetric eigenvalue problems. Particularly, we have considered perturbative arguments to give a first order development of the *prefactor* when the operator is a small perturbation of a symmetric operator, as is the case of eigenvalue problems arising in neutronics, although it is more complex in practice, as we must tackle a generalized eigenvalue problem that does not satisfiy all the assumptions that we raised. We finally came up with a data-driven heuristic approach to estimate the *prefactor*, assuming that its parametric dependency can be disregarded. The derivation of *a posteriori* error estimates for non-symmetric generalized eigenvalue problems is a crucial step in the development of an inexpensive reduced basis method. In that context, these estimates enable an efficient construction of the approximation space over an iterative procedure, and also certify the accuracy of the solutions to the reduced problem.

Chapter 3

Implementation of an iterative reduced basis method based on *a posteriori* error estimates for parameter-dependent non-symmetric generalized eigenvalue problems

Model order reduction methods such as reduced basis (RB) techniques [21, 67, 102] are useful to accelerate the computation of approximate solutions of parameterized problems. In the context of neutronics, parameterized problems naturally occur when optimizing the loading pattern of a nuclear core [44, 47, 114]. Mathematically, this amounts to optimizing an objective function which involves the solution to a generalized non-symmetric eigenvalue problem. The aim of this chapter is then to set up a methodology for the implementation of an *offline/online* reduced basis procedure for parameter-dependent non-self-adjoint generalized eigenvalue problems in this context. It can be seen as a generalization of [69, 54, 20], where reduced basis methods for symmetric eigenvalue problems have been developed. To do so, we rely on the practical *a posteriori* error estimates developed in the previous chapter, which allow, in the *offline* stage, to build the reduced space with the greedy algorithm [25], by breaking the dependence on the *high-fidelity* solver, unlike POD procedures (see [16] for a general introduction), for instance; in the *online* stage, these estimates certify the approximation and enable a convergence analysis of the reduced problem. To do so, we consider a particular situation where the linear operators exhibit an affine expansion along their parametric dependency. This should be considered for the iterative reduced basis to compete with POD. This gives an example of RB implementation where quantities of interest are efficiently assembled along the parameter in a generic way.

In Section 3.1, we introduce the reduced (RB) eigenvalue problem as a Galerkin approximation of the high-fidelity (HF) eigenvalue problem. In Section 3.2, we consider the affine parametric expansion of the high-fidelity matrices, and we explain how this key assumption enables an efficient implementation of the reduced basis method and the *a posteriori* error estimates. Section 3.3 details the *offline* stage of the RB procedure, through a *greedy* algorithm, and shows how the reduced space is thoroughly built. Section 3.4 presents the *online* stage of the RB procedure that consists of the Galerkin projection onto the reduced space, and assembling the algebraic quantities of interest

along the parameters.

3.1 The high-fidelity and the reduced problem

Let us recall the parametrized generalized eigenvalue problem for which we wish to build a reduced-order problem. Let $\mathcal{N} \in \mathbb{N}^*$ be a large positive integer. As before, $\mathbb{R}^{\mathcal{N}}$ is equipped with the Euclidean inner product[1] denoted by $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. For all values of the vector of parameters $\mu \in \mathcal{P} \subset \mathbb{R}^K$, for some $K \geq 1$, we consider two matrices A_μ and B_μ in $\mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ and the following generalized eigenvalue problem: Find $(u_\mu, \lambda_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{C}$ with λ_μ of the smallest modulus such that:

$$A_\mu u_\mu = \lambda_\mu B_\mu u_\mu, \quad \|u_\mu\| = 1. \tag{3.1}$$

We refer to Problem (3.1) as the ***high-fidelity* (HF) problem**. Imposing Assumption 2.3.1 on Problem (3.1), it yields a uniquely defined (up to a sign) u_μ, and **real and strictly positive** λ_μ. Moreover, it ensures a spectral gap between λ_μ and the other eigenvalues of Problem (3.1), a property that was essential in the *a posteriori* analysis developed in Chapter 2. The associated HF adjoint problem then reads: Find $(u_\mu^*, \lambda_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R}_+ \setminus \{0\}$ such that

$$A_\mu^T u_\mu^* = \lambda_\mu B_\mu^T u_\mu^*, \quad \|u_\mu^*\| = 1. \tag{3.2}$$

Remark 3.1.1. *Let us mention here that, for any $A \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$, the adjoint matrix $A^T \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ is defined relatively to the inner product $\langle \cdot, \cdot \rangle$ as follows:*

$$\forall u, v \in \mathbb{R}^{\mathcal{N}}, \quad \langle v, Au \rangle = \langle A^T v, u \rangle.$$

Similarly, for any column vector $u \in \mathbb{R}^{\mathcal{N}}$, we denote by u^T the unique row vector of $\mathbb{R}^{\mathcal{N}}$ such that

$$\forall v \in \mathbb{R}^{\mathcal{N}}, \quad u^T v = \langle u, v \rangle.$$

Let us notice that, from Assumption 2.3.1, the eigenvectors u_μ and u_μ^* can indeed be chosen to be vectors with real components. In practice, the solutions to (3.1) and (3.2) are approximated by the inverse power method [23, Chapter 4], described in Algorithm 2, with $\mathtt{A} = A_\mu$ and $\mathtt{B} = B_\mu$ for the right eigenproblem, and with $\mathtt{A} = A_\mu^T$ and $\mathtt{B} = B_\mu^T$ for the left eigenproblem. We also define, for all $\mu \in \mathcal{P}$, the so-called effective multiplication factor

$$k_\mu := \frac{1}{\lambda_\mu},$$

such that

$$k_\mu = \frac{\langle u_\mu^*, B_\mu u_\mu \rangle}{\langle u_\mu^*, A_\mu u_\mu \rangle}. \tag{3.3}$$

Remark 3.1.2. *On the one hand, Assumption 2.3.1 holds for instance if A_μ is invertible and the matrix $A_\mu^{-1} B_\mu$ coming from Problem* (3.1) *satisfies the assumptions of the Perron–Frobenius theorem [3]. Note that under the assumption that A_μ is invertible, λ_μ is solution to* (3.1) *if and only if k_μ is an eigenvalue associated with the matrix $A_\mu^{-1} B_\mu$. On the other hand, in the context of neutronics applications detailed in Chapters 4 and 5,* (3.1) *is obtained as an appropriate discretization of a continuous problem where the associated*

[1] It is easy to generalize the results presented below to any Hilbertian norm.

resolvent operator satisfies the assumptions of the Krein–Rutman theorem (Theorem 1.2.2) and thus admits a simple real largest eigenvalue in modulus denoted k_μ^{ex}. *Since* $1/k_\mu^{\mathrm{ex}}$ *is solution to the continuous problem, the smallest eigenvalue of* (3.1) *in modulus is also expected to be simple and positive for fine enough discretization, i.e. large enough* $\mathcal{N}$.

We are interested in situations where one has to solve the reference high-fidelity problem (3.1) quickly and for many values of μ. The idea is to build a reduced basis using some HF solutions of (3.1) (so-called *snapshots*) computed *offline*, and to use a Galerkin method to project Problem (3.1) onto this reduced basis. This requires *a posteriori* estimators, developed in Chapter 2, to wisely select the parameters μ used to build the reduced basis, as well as to certify the numerical results obtained *online* using the reduced basis.

Let us now present the reduced-order model obtained from a given reduced basis to get an approximation of (3.1). Let us consider a reduced linear subspace $\mathcal{V}$ of $\mathbb{R}^{\mathcal{N}}$ of dimension N much smaller than $\mathcal{N}$, built in such a way that any solution of Problem (3.1) can be accurately approximated by an element of the space $\mathcal{V}$ (the construction of such a subspace will be discussed in Section 3.3). A reduced-order model for Problem (3.1) can then be obtained from the reduced space $\mathcal{V}$ as follows. Let $(\xi_i)_{1\leqslant i\leqslant N}$ be an orthonormal basis of $\mathcal{V}$. The reduced matrices $A_{\mu,N} \in \mathbb{R}^{N\times N}, B_{\mu,N} \in \mathbb{R}^{N\times N}$ are defined as follows: for all $1 \leq i, j \leq N$,

$$(A_{\mu,N})_{ij} := \langle \xi_i, A_\mu \xi_j \rangle, \tag{3.4}$$
$$(B_{\mu,N})_{ij} := \langle \xi_i, B_\mu \xi_j \rangle. \tag{3.5}$$

The reduced-order model is then used to solve the following problem: Find $(c_{\mu,N}, \lambda_{\mu,N}) \in \mathbb{R}^N \times \mathbb{C}$ such that $\lambda_{\mu,N}$ is an eigenvalue with smallest modulus of

$$A_{\mu,N} c_{\mu,N} = \lambda_{\mu,N} B_{\mu,N} c_{\mu,N}, \quad u_{\mu,N} = \sum_{i=1}^{N} c_{\mu,N}^i \xi_i, \quad \text{and} \quad \|u_{\mu,N}\| = 1, \tag{3.6}$$

where for all $1 \leq i \leq N$, $c_{\mu,N}^i$ is the i^{th} component of the vector $c_{\mu,N}$. We refer to Problem (3.6) as the **reduced (RB) problem**. Similarly as for the HF problem (see Assumption 2.3.1), we make the following assumption.

Assumption 3.1.3. *For any parameter* $\mu \in \mathcal{P}$, *the matrix* $A_{\mu,N}$ *is invertible and there exists a unique positive eigenvalue* $\lambda_{\mu,N}$ *which is the smallest modulus solution to* (3.6). *Moreover, the eigenvalue* $\lambda_{\mu,N}$ *is simple.*

Under this assumption, $c_{\mu,N}$ and $u_{\mu,N}$ are uniquely defined up to a sign and $\lambda_{\mu,N}$ is **real**. Endowing the space $\mathbb{R}^N$ with the canonical Euclidean inner product $\langle \cdot, \cdot \rangle_{\ell^2}$, we can consider the solution to the associated reduced adjoint problem: Find $(c_{\mu,N}^*, \lambda_{\mu,N}) \in \mathbb{R}^N \times \mathbb{R}_+ \setminus \{0\}$ such that the eigenvalue $\lambda_{\mu,N}$ is the smallest in modulus and

$$A_{\mu,N}^T c_{\mu,N}^* = \lambda_{\mu,N} B_{\mu,N}^T c_{\mu,N}^*, \quad u_{\mu,N}^* = \sum_{i=1}^{N} c_{\mu,N}^{*,i} \xi_i, \quad \text{and} \quad \|u_{\mu,N}^*\| = 1. \tag{3.7}$$

where for all $1 \leq i \leq N$, $c_{\mu,N}^{*,i}$ is the i^{th} component of the vector $c_{\mu,N}^*$ and $A_{\mu,N}^T$ and $B_{\mu,N}^T$ are respectively the transpose of the matrix $A_{\mu,N}$ and $B_{\mu,N}$. Moreover, under this

assumption, we have $\langle c^*_{\mu,N}, A_{\mu,N}c_{\mu,N}\rangle_{\ell^2} = \langle u^*_{\mu,N}, A_\mu u_{\mu,N}\rangle \neq 0$ (see Lemma 2.3.2), and we define

$$k_{\mu,N} = \frac{\langle c^*_{\mu,N}, B_{\mu,N}c_{\mu,N}\rangle_{\ell^2}}{\langle c^*_{\mu,N}, A_{\mu,N}c_{\mu,N}\rangle_{\ell^2}} = \frac{\langle u^*_{\mu,N}, B_\mu u_{\mu,N}\rangle}{\langle u^*_{\mu,N}, A_\mu u_{\mu,N}\rangle}. \tag{3.8}$$

In practice, we use the inverse power method described in Algorithm 2 to solve (3.6) and (3.7). If both algorithms converge, we refer to the outputs $c_{\mu,N}$ and $c^*_{\mu,N}$ as the right and left eigenvectors of the reduced problem. If one of the sequences does not converge, the reduced basis is enriched using the high-fidelity left and right eigenvectors for the considered parameter value (see the construction of the reduced space in Section 3.3). Note that the power methods applied to (3.6) and (3.7) (resp. to (3.1) and (3.2)) are guaranteed to converge if Assumption 3.1.3 (resp. Assumption 2.3.1) is satisfied. In the numerical examples presented in Chapter 4, we observe that both algorithms (for the right and left reduced eigenvalue problems) indeed converge and that $\langle c^*_{\mu,N}, A_{\mu,N}c_{\mu,N}\rangle_{\ell^2} \neq 0$ as soon as the reduced space has a sufficiently large dimension (typically $N \geq 4$ is sufficient in the numerical results presented in Chapter 4).

3.2 The affine expansion and an efficient implementation of *a posteriori* error estimates

Before presenting the RB methodology, let us discuss some elements which allow an **efficient** implementation of the procedure, by taking advantage of the decomposition of the HF matrices along their parametric dependency. Indeed, we assume that $\mathcal{P} \subset \mathbb{R}^K$, for some $K \geq 1$, and that there exists two non-zero integers P_A and Q_B such that, for all $\mu = (\mu_1, \ldots, \mu_K) \in \mathcal{P}$, the matrices A_μ and B_μ write

$$A_\mu = \sum_{k=1}^{K}\sum_{p=1}^{P_A} f_p(\mu_k)A_{k,p} + M_{bc}, \tag{3.9}$$

$$B_\mu = \sum_{k=1}^{K}\sum_{q=1}^{Q_B} g_q(\mu_k)B_{k,q}, \tag{3.10}$$

where, for all $1 \leq k \leq K$, $1 \leq p \leq P_A$ and $1 \leq q \leq Q_B$, $f_p(\mu_k) \in \mathbb{R}$ and $g_q(\mu_k) \in \mathbb{R}$, and $A_{k,p} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$, $B_{k,q} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ and $M_{bc} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ are μ-independent matrices.

Remark 3.2.1. *In neutronics applications, such as the ones presented in Chapter 4, the dimension K of the parameter space $\mathcal{P}$ stems from a partition of the core $\mathcal{R}$; the vectors f and g contain the parameter-dependent coefficients of the equations, i.e. the coefficients $D^1, \Sigma^{11}, \Sigma^{12}, D^2, \Sigma^{21}, \Sigma^{22}, \chi^1, \chi^2, \Sigma^1_f, \Sigma^2_f$ ($P_A = 6$, $Q_B = 4$) for the two-group neutron diffusion equations (see Section 1.4.2); M_{bc} comes from the discretization of the boundary conditions, and therefore does not depend on μ.*

As a consequence, the parameter-independent matrices $A_{k,p}$, $B_{k,q}$ ($1 \leq k \leq K$, $1 \leq p \leq P_A$, $1 \leq q \leq Q_B$) and M_{bc} can be pre-computed in order to efficiently assemble the Galerkin projection of the matrices A_μ and B_μ *online*, and the residuals $R_N(\mu)$ and $R^*_N(\mu)$ defined in (2.17) and (2.18). Indeed, thanks to the affine decomposition of the matrices A_μ and B_μ above, the residual norm is easily computable online (see Section 3.3), as it only requires algebraic operations over vectors of the size of the (small) reduced basis, which is N.

3.3 The *offline* stage: the greedy algorithm

Let us start with the *offline* stage of the RB procedure, which is dedicated to the construction of the reduced space. In this section, the parametric exploration is restricted to the parametric subspace denoted by $\mathcal{P}_{\text{train}} \subset \mathcal{P}$, namely the *training set*. In practice, the reduced space $\mathcal{V}$ used in the reduced-order model described in Section 3.1 is built following the standard procedure of the reduced basis technique [93]. We first initialize the reduced space $\mathcal{V}_0$ as a very low-dimensional space spanned by a few snapshots of the direct and adjoint HF problems. A sequence of parameter values $(\mu_n)_{n\geq 1}$ is then selected from a greedy procedure [25] described below, from which nested reduced spaces $(\mathcal{V}_n)_{n\geq 1}$ are built as follows:

$$\forall n \geq 1, \quad \mathcal{V}_n = \mathcal{V}_0 + \text{Span}\left\{u_{\mu_1}, \ldots, u_{\mu_n}, u^*_{\mu_1}, \ldots, u^*_{\mu_n}\right\}. \tag{3.11}$$

In the following, we denote by $N_n := \dim\mathcal{V}_n$ and by u_{μ,N_n}, u^*_{μ,N_n}, λ_{μ,N_n} and k_{μ,N_n} the solutions of the RB problems described in the Section 3.1, for $\mathcal{V} = \mathcal{V}_n$.

The choice made in (3.11) to enrich the sequence of reduced spaces with both the eigenvector of the direct and of the adjoint HF problem stems from the *a priori* error analysis of Galerkin approximations of generalized eigenvalue problems (see Section 2.1). Hence, it appears natural when it comes to the design of a greedy procedure in the present reduced basis context to enrich the Galerkin approximation space with snapshots of both direct and adjoint eigenvalue problems, in order to at least get the reference solution for the reduced problem when considering the parameter $\mu \in \mathcal{P}_{\text{train}}$.

In the greedy procedure, the parameters $(\mu_n)_{\geq 1}$ need to be selected satisfying some specific criteria. In practice, we select snapshots over $\mathcal{P}_{\text{train}}$ maximizing some error surrogate Δ_{N_n} for the error between solutions of the HF model and the RB model, as is described in Algorithm 1. In an *ideal greedy* procedure, we would choose the exact error as the error surrogate Δ_{N_n}. In that case, two possible choices for the definition of Δ_{N_n} would be:

a) either the eigenvalue error: $\Delta_{N_n}(\mu) := e^k_{N_n}(\mu)$

b) or the eigenvector errors: $\Delta_{N_n}(\mu) := e^u_{N_n}(\mu) + e^{u^*}_{N_n}(\mu)$

with

$$e^u_{N_n}(\mu) := \|u_\mu - u_{\mu,N_n}\|, \quad e^{u^*}_{N_n}(\mu) := \|u^*_\mu - u^*_{\mu,N_n}\| \text{ and } e^k_{N_n}(\mu) := |k_\mu - k_{\mu,N_n}|.$$

However, these quantities are of course not available in general, so one has to resort to *a posteriori* error estimates for an *efficient greedy* algorithm. Therefore, the strategies developed in Chapter 2 allow to define an *a posteriori* error estimator $\Delta_{N_n}(\mu)$ in order to obtain an estimation of the errors on the eigenvalues and the eigenvectors for any reduced space $\mathcal{V}$ without having to compute the solutions of the HF eigenvalue problem.

3.3.1 Initialization of the reduced basis

The idea is to start with a very low-dimensional space $\mathcal{V}_0 \subset \mathbb{R}^{\mathcal{N}}$ such that

$$\mathcal{V}_0 = \text{Span}\left\{u_{\hat{\mu}_1}, \ldots, u_{\hat{\mu}_{n_s}}, u^*_{\hat{\mu}_1}, \ldots, u^*_{\hat{\mu}_{n_s}}\right\},$$

where $\{\hat{\mu}_1, \ldots, \hat{\mu}_{n_s}\} \subset \mathcal{P}_{\text{train}}$, with $n_s \in \mathbb{N}^*$ small.

Algorithm 1 GREEDY ALGORITHM FOR BUILDING THE REDUCED SUBSPACE

Input: $\mathcal{P}_{\text{train}} \subset \mathcal{P}$: training set of parameters, $\tau > 0$: error tolerance threshold, $\mathcal{V}_0 \subset \mathbb{R}^{\mathcal{N}}$: initial reduced space
$N_0 := \dim \mathcal{V}_0$
$\tau_0 := \max_{\mu \in \mathcal{P}_{\text{train}}} \Delta_{N_0}(\mu)$
$n := 0$
while $\tau_n > \tau$ **do**
  $\mu_{n+1} := \operatorname{argmax}_{\mu \in \mathcal{P}_{\text{train}}} \Delta_{N_n}(\mu)$
  Compute $u_{\mu_{n+1}}$ and $u^*_{\mu_{n+1}}$.
  $\mathcal{V}_{n+1} := \mathcal{V}_n + \text{Span}\{u_{\mu_{n+1}}, u^*_{\mu_{n+1}}\}$
  $N_{n+1} := \dim \mathcal{V}_{n+1}$
  $\tau_{n+1} := \max_{\mu \in \mathcal{P}_{\text{train}}} \Delta_{N_{n+1}}(\mu)$
  $n := n + 1$
end while
Output: Reduced space $\mathcal{V} := \mathcal{V}_n \subset \mathbb{R}^{\mathcal{N}}$

Remark 3.3.1. *We may start the greedy iterative procedure with only one parameter, i.e. with* $\mathcal{V}_0 = Span\{u_{\hat{\mu}_1}, u^*_{\hat{\mu}_1}\}$, $\hat{\mu}_1 \in \mathcal{P}_{train}$. *However, we observe in practice that the greedy algorithm may suffer from stability issues in the first steps of the procedure, which are crucial, as the initial reduced space is not rich enough and thus does not ensure a good enough convergence.*

To do so, we apply a Singular Value Decomposition (SVD) [110] to the so-called *matrix of snapshots* composed of left and right eigenvectors

$$S = \left(u_{\hat{\mu}_1} | u^*_{\hat{\mu}_1} | \cdots | u_{\hat{\mu}_{n_s}} | u^*_{\hat{\mu}_{n_s}}\right) \in \mathbb{R}^{\mathcal{N} \times 2n_s},$$

to obtain an $\mathbb{X}$-orthonormal basis of $\mathcal{V}_0$, where $\mathbb{X}$ stands for the Gram matrix of size $\mathcal{N} \times \mathcal{N}$ for the considered inner product $\langle \cdot, \cdot \rangle$, e.g., $\mathbb{X} = I$ if $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product.

We compute the SVD of $\tilde{S} = \mathbb{X}^{1/2} S$ which writes

$$\begin{aligned} S &= U \Sigma Z^T, \\ U &= \left(\tilde{\xi}_1 | \ldots | \tilde{\xi}_{\mathcal{N}}\right) \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}, \\ \Sigma &= \text{diag}\left(\sigma_1, \ldots, \sigma_{\min(2n_s, \mathcal{N})}\right), \\ Z &= \left(\tilde{\psi}_1 | \ldots | \tilde{\psi}_{2n_s}\right) \in \mathbb{R}^{n_s \times 2n_s}, \end{aligned}$$

where the σ_i are the singular values of $\tilde{S}$, sorted in decreasing order, and U and Z are two orthogonal matrices. We then choose

$$\mathcal{V}_0 = \text{Span}\left\{\mathbb{X}^{-1/2}\tilde{\xi}_1, \ldots, \mathbb{X}^{-1/2}\tilde{\xi}_{N_0}\right\},$$

with $1 \leq N_0 \leq 2n_s$.

Finally, in order to get reliable and computable *a posteriori* estimates, as those developed in Chapter 2, which enable an efficient greedy algorithm, the prefactors in the obtained upper bounds lead us to compute the HF solutions (u_μ, u^*_μ, k_μ) to (3.1) and (3.2), for $\mu \in \mathcal{P}_{\text{pref}}$, where $\mathcal{P}_{\text{pref}} \subset \mathcal{P}$ is small enough not to penalize the cost of the *offline* stage, and such that $\mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{pref}} = \emptyset$.

3.3.2 An iteration of the greedy algorithm: searching for the next basis function

Let us assume that n iterations of the greedy algorithm have successfully been completed, and let us denote by $\mathcal{V}_n$ the reduced space built so far. In this section, we detail the procedure for one iteration, i.e. how to get an updated reduced space $\mathcal{V}_{n+1}$ from $\mathcal{V}_n$. Let $(\xi_1, \dots, \xi_{N_n})$ be an orthonormal basis of $\mathcal{V}_n$ (for the inner product $\langle \cdot, \cdot \rangle$), and let $V_n \in \mathbb{R}^{\mathcal{N} \times N_n}$ be the matrix containing the coordinates of the basis $(\xi_1, \dots, \xi_{N_n})$ in the canonical basis of $\mathbb{R}^{\mathcal{N}}$.

Step #1: Compute the parameter-independent matrices

We use the μ-affine expansion of the HF matrices A_μ and B_μ (see Section 3.2) in order to prepare the computation of the RB matrices A_{μ,N_n} and B_{μ,N_n}, defined in (3.4) and (3.5), as they also exhibit an affine decomposition along the parameter μ. Hence, for all $1 \leq k \leq K$, $1 \leq p \leq P_A$, $1 \leq q \leq Q_B$, we compute the following reduced matrices of dimension $N_n \times N_n$

$$A_{k,p}^{N_n} := V_n^T A_{k,p} V_n,$$
$$B_{k,q}^{N_n} := V_n^T B_{k,q} V_n,$$

and

$$M_{bc}^{N_n} := V_n^T M_{bc} V_n.$$

Remark 3.3.2. *In the case of the resolution of the G-group neutron diffusion equations (see Section 1.4.2 for an introduction; Chapter 4 for applications), for $k \in [\![1, K]\!]$, the matrices $A_{k,p}$ and $B_{k,q}$ are local stiffness and mass matrices, and thus, each matrix $B_{k,q}$ corresponds to a certain matrix $A_{k,p}$. Even after projecting onto $\mathcal{V}_n$, the same applies to the matrices $B_{k,q}^{N_n}$. Therefore, we have to pre-compute $K \times G$ stiffness matrices and $K \times G^2$ mass matrices, that is in total $K \times (G + G^2)$ matrices independent of μ of size $N_n \times N_n$.*

Step #2: Prepare the computation of the residual norm

As discussed in Section 3.2, the residual norm can also be computed along an *offline/online* strategy. To do so, we compute, for $1 \leq k \leq K$, $1 \leq p \leq P_A$, $1 \leq q \leq Q_B$,

$$\hat{A}_{k,p}^{N_n} := A_{k,p} V_n,$$
$$\hat{B}_{k,q}^{N_n} := B_{k,q} V_n,$$

and

$$\hat{M}_{bc}^{N_n} := M_{bc} V_n.$$

Then, we compute, for $1 \leq k, l \leq K$, $1 \leq p, p' \leq P_A$, $1 \leq q, q' \leq Q_B$, the following reduced

matrices of dimension $N_n \times N_n$

$$\begin{aligned}
D^{N_n}_{k,l,p,p'} &:= (\hat{A}^{N_n}_{k,p})^T \mathbb{X}^{-1} \hat{A}^{N_n}_{l,p'} \\
E^{N_n}_{k,l,p,q} &:= (\hat{A}^{N_n}_{k,p})^T \mathbb{X}^{-1} \hat{B}^{N_n}_{l,q} \\
F^{N_n}_{k,l,q,q'} &:= (\hat{B}^{N_n}_{k,q})^T \mathbb{X}^{-1} \hat{B}^{N_n}_{l,q'} \\
D^{N_n}_{bc,k,p} &:= (\hat{M}^{N_n}_{bc})^T \mathbb{X}^{-1} \hat{A}^{N_n}_{k,p} \\
E^{N_n}_{bc,k,q} &:= (\hat{M}^{N_n}_{bc})^T \mathbb{X}^{-1} \hat{B}^{N_n}_{k,q} \\
F^{N_n}_{bc} &:= (\hat{M}^{N_n}_{bc})^T \mathbb{X}^{-1} \hat{M}^{N_n}_{bc},
\end{aligned}$$

where $\mathbb{X}$ stands for the Gram matrix of size $\mathcal{N} \times \mathcal{N}$ for the considered inner product $\langle \cdot, \cdot \rangle$ ($\mathbb{X} = I$ if $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product).

Remark 3.3.3. *As the reduced spaces $(\mathcal{V}_n)_{n \geq 1}$ are nested, we do not need to compute from scratch the reduced matrices above at each step of the algorithm. Indeed, in the case where the relation $V_n = (V_{n-1} \mid \xi_n \mid \xi_n^*)$ (i.e. V_n is the addition of the two column basis vectors ξ_n and ξ_n^* to the previous reduced matrix V_{n-1}) holds, the computation of one matrix $D^{N_n}_{k,l,p,p'}$ is carried out as follows*

$$\begin{aligned}
v_{1,k,p} &:= A_{k,p}\left[\xi_n, \xi_n^*\right] \\
v_{2,k,p} &:= \hat{A}^{N_n}_{k,p} \\
w_{1,k,p} &:= \mathbb{X}^{-1} v_{1,k,p} \\
w_{2,k,p} &:= \mathbb{X}^{-1} v_{2,k,p}
\end{aligned}$$

$$D^{N_n}_{k,l,p,p'} = \begin{pmatrix} D^{N_{n-1}}_{k,l,p,p'} & v_{2,k,p}^T w_{1,k,p'} \\ v_{1,k,p}^T w_{2,k,p'} & v_{1,k,p}^T w_{1,k,p'} \end{pmatrix},$$

and we update

$$\hat{A}^{N_{n+1}}_{k,p} \leftarrow \left[v_{2,k,p}, v_{1,k,p}\right].$$

The computation of the matrices $E^{N_n}_{k,l,p,q}, F^{N_n}_{k,l,q,q'}, D^{N_n}_{bc,k,p}, E^{N_n}_{bc,k,q}, F^{N_n}_{bc}$ then follow similar recurrence relations.

Then, the following steps #3 and #4 are carried out for all $\mu \in \mathcal{P}_{\text{train}}$.

Step #3: Solve the RB problem

From the parameter-independent matrices computed in Step #1, we first assemble *online* the reduced matrices A_{μ,N_n} and B_{μ,N_n} of size $N_n \times N_n$ as in (3.4) and (3.5) respectively. For $\mu \in \mathcal{P}_{\text{train}}$, the computation of these reduced matrices then reads as

$$A_{\mu,N_n} = \sum_{k=1}^{K} \sum_{p=1}^{P_A} f_p(\mu_k) A^{N_n}_{k,p} + M^{N_n}_{bc}, \tag{3.12}$$

$$B_{\mu,N_n} = \sum_{k=1}^{K} \sum_{q=1}^{Q_B} g_q(\mu_k) B^{N_n}_{k,q}. \tag{3.13}$$

We then solve the RB problems (3.6) and (3.7), using Algorithm 2. Here, two cases arise:

- if both iterative procedures converge, then we get the reduced outputs c_{μ,N_n}, c^*_{μ,N_n}, k_{μ,N_n}, and we check that the relation (3.8) holds,

$$k_{\mu,N_n} = \frac{\langle c^*_{\mu,N}, B_{\mu,N} c_{\mu,N}\rangle_{\ell^2}}{\langle c^*_{\mu,N}, A_{\mu,N} c_{\mu,N}\rangle_{\ell^2}};$$

- if one of the iterative procedures does not converge, then we break the iteration, and we choose $\mu_{n+1} = \mu$.

Step #4: Compute the residual norm

Let $\mu \in \mathcal{P}_{\text{train}}$. We use the independent-parameter matrices computed in Step #2 to assemble *online* the residual norm as

$$\|R_{N_n}(\mu)\| := \|\left(B_\mu - k_{\mu,N_n} A_\mu\right) u_{\mu,N_n}\| = \sqrt{c^T_{\mu,N_n} G_{\mu,N_n} c_{\mu,N_n}},$$

with

$$\begin{aligned} G_{\mu,N_n} = |k_{\mu,N_n}|^2 &\left(\sum_{k,l=1}^{K} \sum_{p,p'=1}^{P_A} f_p(\mu_k) f_{p'}(\mu_l) D^{N_n}_{k,l,p,p'} + \sum_{k=1}^{K} \sum_{p=1}^{P_A} f_p(\mu_k) \left(D^{N_n}_{bc,k,p} + (D^{N_n}_{bc,k,p})^T \right) + F^{N_n}_{bc} \right) \\ &- k_{\mu,N_n} \left(\sum_{k,l=1}^{K} \sum_{p=1}^{P_A} \sum_{q=1}^{Q_B} f_p(\mu_k) g_q(\mu_l) \left(E^{N_n}_{k,l,p,q} + (E^{N_n}_{k,l,p,q})^T \right) \right) \\ &+ \sum_{k=1}^{K} \sum_{q=1}^{Q_B} g_q(\mu_k) \left(E^{N_n}_{bc,k,q} + (E^{N_n}_{bc,k,q})^T \right) + \sum_{k,l=1}^{K} \sum_{q,q'=1}^{Q_B} g_q(\mu_k) g_{q'}(\mu_l) F^{N_n}_{k,l,q,q'}. \end{aligned}$$

A similar construction is readily possible for $\|R^*_{N_n}(\mu)\|$.

In practice, the residual norm $\|R_{N_n}(\mu)\|$ is computed as follows. Define:

$$r_1 = \sum_{k,l=1}^{K} \sum_{q,q'=1}^{Q_B} g_q(\mu_k) g_{q'}(\mu_l) c^T_{\mu,N_n} F^{N_n}_{k,l,q,q'} c_{\mu,N_n},$$

$$r_2 = \sum_{k,l=1}^{K} \sum_{p=1}^{P_A} \sum_{q=1}^{Q_B} f_p(\mu_k) g_q(\mu_l) c^T_{\mu,N_n} E^{N_n}_{k,l,p,q} c_{\mu,N_n},$$

$$r_3 = \sum_{k,l=1}^{K} \sum_{p,p'=1}^{P_A} f_p(\mu_k) f_{p'}(\mu_l) c^T_{\mu,N_n} D^{N_n}_{k,l,p,p'} c_{\mu,N_n},$$

$$r_4 = \sum_{k=1}^{K} \sum_{q=1}^{Q_B} g_q(\mu_k) c^T_{\mu,N_n} E^{N_n}_{bc,k,q} c_{\mu,N_n},$$

$$r_5 = \sum_{k=1}^{K} \sum_{p=1}^{P_A} f_p(\mu_k) c^T_{\mu,N_n} D^{N_n}_{bc,k,p} c_{\mu,N_n},$$

$$r_6 = c^T_{\mu,N_n} (F^{N_n}_{bc})^T c_{\mu,N_n}.$$

We then have

$$\|R_{N_n}(\mu)\|^2 = r_1 - 2 \times (r_2 + r_4) \times k_{\mu,N_n} + (r_3 + 2r_5 + r_6) \times |k_{\mu,N_n}|^2. \tag{3.14}$$

Remark 3.3.4. *Note that some rounding errors may occur while assembling* (3.14)*, because of the large number of terms in the sum (there are* $K^2(P_A^2+Q_B^2+P_AQ_B)+K(P_A+Q_B)+1$ *inner products of size* $N_n \times N_n$*). In that case, we may get* $\|R_{N_n}(\mu)\|^2 < 0$*, in which case we set* $\|R_{N_n}(\mu)\| = 0$*.*

Remark 3.3.5. *In the case of the resolution of the G-group neutron diffusion equations (cf. Remark 3.3.2), we need to compute a total of* $K^2(G+G^2)^2 + K(G+G^2)+1$ *inner products of size* $N_n \times N_n$*.*

We may also use the quantity defined in (2.29) as

$$\eta_{N_n}^k(\mu) := \frac{\|R_{N_n}(\mu)\|\|R_{N_n}^*(\mu)\|}{\left|\langle c_{N_n}^*, A_{\mu,N_n} c_{N_n}\rangle\right|}. \tag{3.15}$$

Step #5: Compute the *a posteriori* error estimate

We compute the *a posteriori* error estimate Δ_{N_n} using the residual norm computed in the previous step, as well as the prefactor estimations (see Section 2.4.3) from the pre-computed HF solutions in Section 3.3.1. We first compute, for all $\mu \in \mathcal{P}_{\text{pref}}$, the efficiency ratios defined in (2.43), in order to get *practical* prefactors $\overline{C}_{N_n}^k$, $\overline{C}_{N_n}^u$ and $\overline{C}_{N_n}^{u^*}$ as in (2.45). Then, two possible choices for Δ_{N_n} are:

a) $\Delta_{N_n}(\mu) := \Delta_{N_n}^k(\mu)$

b) $\Delta_{N_n}(\mu) := \Delta_{N_n}^u(\mu) + \Delta_{N_n}^{u^*}(\mu)$

with

$$\Delta_{N_n}^k(\mu) = \overline{C}_{N_n}^k \eta_{N_n}^k(\mu), \tag{3.16}$$

$$\Delta_{N_n}^u(\mu) = \overline{C}_{N_n}^u \|R_{N_n}(\mu)\|, \tag{3.17}$$

$$\Delta_{N_n}^{u^*}(\mu) = \overline{C}_{N_n}^{u^*} \|R_{N_n}^*(\mu)\|. \tag{3.18}$$

$$\tag{3.19}$$

Step #6: Select the next basis function

We then select the parameter $\mu_{n+1} \in \mathcal{P}_{\text{train}}$ that maximizes the error estimate Δ_{N_n}:

$$\mu_{n+1} = \underset{\mu \in \mathcal{P}_{\text{train}}}{\operatorname{argmax}}\, \Delta_{N_n}(\mu).$$

As discussed in Step #3, there might exist a parameter $\mu \in \mathcal{P}_{\text{train}}$ which yields non-convergence of the reduced problem. Also, it may be possible that various parameters in $\mathcal{P}_{\text{train}}$ realize the maximum value of Δ_{N_n}. In that case, the parameter μ_{n+1} is chosen by the following order of priority:

1. if the reduced problem does not converge for a parameter μ, we choose $\mu_{n+1} = \mu$;
2. otherwise, if for example, μ^1 and μ^2 both realize the maximum value of the error surrogate Δ_{N_n}, there is no preferred choice, and we either consider $\mu_{n+1} = \mu^1$, or $\mu_{n+1} = \mu^2$.

Step #7: Enrich the reduced space

To do so, we need to compute the HF solutions $u_{\mu_{n+1}}$, $u^*_{\mu_{n+1}}$ and $k_{\mu_{n+1}}$ by solving the HF problems (3.1) and (3.2), for $\mu = \mu_{n+1}$. Then, the idea is to orthogonalize the vectors $u_{\mu_{n+1}}$ and $u^*_{\mu_{n+1}}$ to the basis $(\xi_1, \ldots, \xi_{N_n})$ of $\mathcal{V}_n$. We then apply a Gram–Schmidt process and we compute the quantities

$$\Pi_{\mathrm{GS}} := u_{\mu_{n+1}} - V_n V_n^T \mathbb{X} u_{\mu_{n+1}},$$
$$\Pi^*_{\mathrm{GS}} := u^*_{\mu_{n+1}} - V_n V_n^T \mathbb{X} u^*_{\mu_{n+1}}.$$

Let us denote by $\varepsilon_{\mathrm{GS}}$ a very small tolerance parameter for the Gram–Schmidt orthogonalization procedure. If $\|\Pi_{\mathrm{GS}}\| < \varepsilon_{\mathrm{GS}}$ (resp. $\|\Pi^*_{\mathrm{GS}}\| < \varepsilon_{\mathrm{GS}}$), we do not take the contribution of $u_{\mu_{n+1}}$ (resp. $u^*_{\mu_{n+1}}$) into account in the next reduced space $\mathcal{V}_{n+1}$. Otherwise, we compute

$$\xi_{n+1} := \frac{u_{\mu_{n+1}} - V_n V_n^T \mathbb{X} u_{\mu_{n+1}}}{\left\| u_{\mu_{n+1}} - V_n V_n^T \mathbb{X} u_{\mu_{n+1}} \right\|},$$
$$\xi^*_{n+1} := \frac{u^*_{\mu_{n+1}} - V_n V_n^T \mathbb{X} u^*_{\mu_{n+1}}}{\left\| u^*_{\mu_{n+1}} - V_n V_n^T \mathbb{X} u^*_{\mu_{n+1}} \right\|},$$

and we build

$$V_{n+1} = \left(V_n \mid \xi_{n+1} \mid \xi^*_{n+1} \right),$$

so that

$$\mathcal{V}_{n+1} = \mathcal{V}_n + \mathrm{Span}\{u_{\mu_{n+1}}, u^*_{\mu_{n+1}}\}.$$

3.3.3 The stopping criterion

Let us consider a given tolerance threshold $\tau > 0$. The idea is to let the greedy iterative procedure keep running as long as the maximum error (on the eigenvalue and/or the eigenvectors, depending on the most relevant quantity of interest) does not overcome τ. Therefore, the choice of the error surrogate Δ_N is crucial regarding the quality of the reduced space $\mathcal{V}$ as an output of the algorithm, in the case it is endowed with such a stopping criterion. Note that we could just choose a given maximum number of iterations, denoted by $n_{\max}$ and then expect to get the reduced space $\mathcal{V}_{n_{\max}}$, but without having any idea on the quality of approximation brought by the space $\mathcal{V}_{n_{\max}}$.

Let us consider the space $\mathcal{V}_n$ built after $n \geq 1$ iterations of the greedy algorithm, with $N_n = \dim \mathcal{V}_n$. A criterion for the quality of the approximation brought by $\mathcal{V}_n$ is to verify that

$$\tau_n := \max_{\mu \in \mathcal{P}_{\mathrm{train}}} \Delta_{N_n}(\mu) \leq \tau.$$

Otherwise, we search for a next parameter μ_{n+1} in order to enrich the reduced space, until the criterion is satisfied.

3.4 The *online* stage: solving the reduced problem

Once the greedy algorithm generates a reduced space $\mathcal{V}$, the associated reduced-order model can be tested as we assemble and solve, in an *online* stage, the reduced problems (3.6) and (3.7), for all $\mu \in \mathcal{P}_{\mathrm{test}}$, where $\mathcal{P}_{\mathrm{test}} \subset \mathcal{P}$ is a parameter subspace, namely the *test set*, given as an input by the user at the beginning of the procedure, and which

verifies $\mathcal{P}_{\text{test}} \cap \mathcal{P}_{\text{train}} = \emptyset$.

Let $\mu \in \mathcal{P}_{\text{test}}$, and $N = \dim\mathcal{V}$. First, from the matrices $A^N_{k,p}$, $B^N_{k,q}$, M^N_{bc} computed *offline*, we compute the reduced matrices $A_{\mu,N}$ and $B_{\mu,N}$ of size $N \times N$ as in (3.12) and (3.13) respectively, with complexity of $O(N^2)$. Then, we solve Problems (3.6) and (3.7) using Algorithm 2, as described below.

Algorithm 2 INVERSE POWER METHOD - SOLVE $\mathtt{A}u = \lambda \mathtt{B}u$

Input: $\mathtt{A} \in \mathbb{R}^{M\times M}$, $\mathtt{B} \in \mathbb{R}^{M\times M}$, τ_u: acceptance criterion for the eigenvector, τ_λ: acceptance criterion for the eigenvalue
Choose a random positive unit vector u_0 and $k_0 \neq 0$
Set $i = 0$ and STOP=**false**
while (STOP==**false**) **do**
  Solve $\mathtt{A}v_{i+1} = \mathtt{B}u_i$
  $u_{i+1} = \dfrac{v_{i+1}}{\|v_{i+1}\|}$
  $k_{i+1} = \langle v_{i+1}, u_i \rangle$
  STOP= $\left[\dfrac{\|u_{i+1} - u_i\|}{\|u_i\|} \leq \tau_u \textbf{ and } \dfrac{|k_{i+1} - k_i|}{|k_i|} \leq \tau_\lambda \right]$
  $i = i + 1$
end while
Output: $(u, \lambda) = \left(u_i, \frac{1}{k_i}\right)$

We then compute the residual norms $\|R_N(\mu)\|$, $\|R_N(\mu)\|$ and the quantity $\eta^k_N(\mu)$, respectively defined in (2.17), (2.18) and (2.29), as in Steps #2 and #4 of the *offline* stage (see Section 3.3).

***

Thanks to the practical *a posteriori* error estimates developed in Chapter 2, we implemented a reduced basis routine for the resolution of a generalized non-symmetric and parameter-dependent eigenvalue problem. Based on the greedy algorithm, and under the assumption that the matrices of the high-fidelity problem exhibit an affine decomposition along the parameter, it enables an efficient implementation of the reduced basis method which provides a reduced model whose computational cost competes with other reduced basis methods, such as Proper Orthogonal Decomposition (POD) methods. The use of *a posteriori* error estimates also allows a certification of the obtained reduced model and provides an adaptative approach in the construction of the RB basis.

Chapter 4

Numerical experiments with a greedy reduced basis approach for the resolution of affine-parametrized two-group neutron diffusion problems on mock-up codes

The aim of this chapter is to illustrate the behavior of the proposed reduced basis method, detailed in the last chapter, on examples arising from neutronics applications. We consider the two-group neutron diffusion equations, defined in (1.41), and under the same assumptions presented in Section 1.4.2. We focus our study on three test cases. In Section 4.1, we propose a first application of the reduced basis method to a two-dimensional simplistic core made of four material regions, endowed with non-physical properties and cross-sections. It enables a fast computational analysis of the RB method via prefactor-free *a posteriori* error estimates. In Section 4.2, a more realistic rectangular core example, namely the *Minicore*, challenges the RB method and uses practical *a posteriori* error estimates. Finally, in Section 4.3, we highlight the importance of considering the adjoint problem in the construction of the reduced basis with quick calculations on a three-dimensional Pressurized Water Reactor (PWR) benchmark.

For the two first examples, calculations and implementations are carried out on a mock-up code, written in Python 3.6. The high-fidelity discretization along the space variable is generated using the methods from the open-source FEniCS Project. For the third and last example, we use a POD code written in MATLAB [57], based on the finite element library *deal.II* [8], and the GMSH mesh generator [58].

In the following numerical tests, the domain $\mathcal{R}$ is chosen as $[0, L]^2$ for some $L > 0$. We introduce a partition $(\mathcal{R}_k)_{k=1}^K$ of the domain $\mathcal{R}$ and the parameter functions in the definition of Problem (1.41) are assumed to be piecewise constant on each $\mathcal{P}_k$ for $1 \leq k \leq K$. The parameter μ is thus a K-dimensional vector of either scalars or vectors (containing macro-parameters such as the material, the burn up, the fuel temperature, or the boron concentration for example), which allows to set the values of the coefficients $D^1, \Sigma^{11}, \Sigma^{12}, D^2, \Sigma^{21}, \Sigma^{22}, \chi^1, \chi^2, \Sigma_f^1, \Sigma_f^2$ on the domain $\mathcal{R}_k$, for each $1 \leq k \leq K$. Note that for all $\mu \in \mathcal{P}$, the matrices A_μ and B_μ admit an affine decomposition as in (3.9) and (3.10).

The eigenvalue solver, for both high-fidelity and reduced-order models, is selected as

the inverse power method given in Algorithm 2. In our experiments, we consider relative error tolerances set to $\tau_u = 10^{-6}$ and $\tau_\lambda = 10^{-7}$.

In order to ensure a positive high-fidelity flux, we assume the positivity conditions

$$\sum_{j=1}^{\mathcal{N}} u_\mu^j \geq 0, \quad \sum_{j=1}^{\mathcal{N}} u_\mu^{*,j} \geq 0.$$

Then, to define properly the reduced flux, we choose

$$\langle u_\mu, B_\mu u_{\mu,N} \rangle \geq 0,$$
$$\langle u_\mu^*, B_\mu u_{\mu,N}^* \rangle \geq 0,$$

with vectors u_μ, u_μ^*, $u_{\mu,N}$, $u_{\mu,N}^*$ normalized with respect to the matrix B_μ, such that

$$\langle u_\mu, B_\mu u_\mu \rangle = \langle u_\mu^*, B_\mu u_\mu^* \rangle = 1,$$
$$\langle u_{\mu,N}, B_\mu u_{\mu,N} \rangle = \langle u_{\mu,N}^*, B_\mu u_{\mu,N}^* \rangle = 1.$$

Let us introduce the following notations

$$e_N^k(\mu) = |k_\mu - k_{\mu,N}|, \quad e_N^{k,rel}(\mu) = \frac{|k_\mu - k_{\mu,N}|}{|k_\mu|},$$
$$e_N^u(\mu) = \|u_\mu - u_{\mu,N}\|_{\ell^2}, \quad e_N^{u,rel}(\mu) = \frac{\|u_\mu - u_{\mu,N}\|_{\ell^2}}{\|u_\mu\|_{\ell^2}}, \quad e_{N,L^2}^{u,rel}(\mu) = \frac{\|u_\mu - u_{\mu,N}\|_{L^2}}{\|u_\mu\|_{L^2}},$$
$$e_N^{u^*}(\mu) = \|u_\mu^* - u_{\mu,N}^*\|_{\ell^2},$$

where the ℓ^2 norm is the Euclidean norm, and L^2 refers to the L^2 functional norm applied to the functions in the space $V_\mathcal{N}$ built from the vectors in $\mathbb{R}^\mathcal{N}$ through (1.48). Moreover, we denote by t_{HF} and t_{RB} the mean computational times for one run (for a given parameter) of the high-fidelity and reduced solvers respectively.

4.1 The toy problem

The reduced basis method is first applied to a simple test case where $L = 60$ (we use here reduced units) modeled with $\mathcal{N} = 2 \times 841$ degrees of freedom along $K = 4$ subdomains. Figure 4.1 shows the mesh used for the test case as well as the decomposition of $\mathcal{R}$ into four subdomains. Here, we set $B_\mu = I$ for all $\mu \in \mathcal{P}$.

The training and test sets $\mathcal{P}_{\text{train}}$ and $\mathcal{P}_{\text{test}}$ are constructed using the following random sampling scheme: in each subdomain $\mathcal{R}_k$, for $1 \leq k \leq K$, the values of the coefficients are independently distributed according to the following laws:

- $\Sigma_{s,0}^{i\to j}$: uniform law on $[0, 0.15]$, $1 \leq i, j \leq 2$;
- Σ_t^1 and Σ_t^2: uniform law on $[2(\Sigma_{s,0}^{1\to 2} + \Sigma_{s,0}^{2\to 1}), 0.7]$;
- $D^i = \dfrac{1}{3\Sigma_t^1}$, $i = 1, 2$;
- $\chi_i \nu \Sigma_f^j = \delta_{ij}$, $1 \leq i, j \leq 2$.

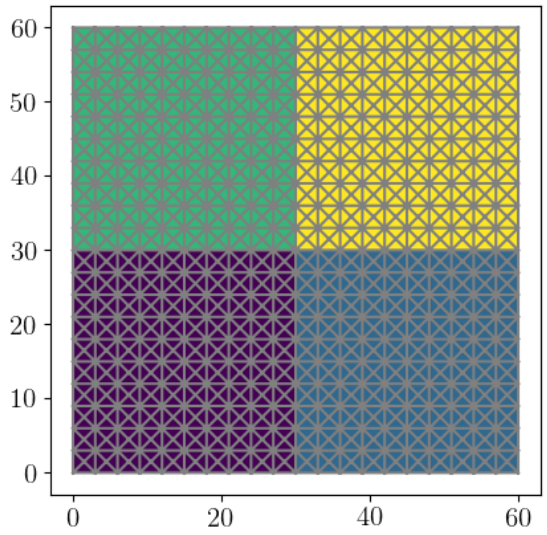

Figure 4.1: Domain of calculation for the two-group toy example with its associated mesh

The coefficients are chosen so that the coercivity of Problems (1.44) and (1.45) are ensured. The parametric spaces $\mathcal{P}_{\text{train}}$ and $\mathcal{P}_{\text{test}}$ are selected following the random sampling procedure described above so that

$$\begin{cases} \#\mathcal{P}_{\text{train}} = 300 \\ \#\mathcal{P}_{\text{test}} = 50 \\ \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset. \end{cases}$$

In the offline stage, the greedy algorithm is performed using the *a posteriori* estimator

$$\Delta_N(\mu) = \eta_N^k(\mu)$$

defined in (2.29) for all $\mu \in \mathcal{P}$ (in other words, we choose here $\overline{C}_N^k(\mu) = 1$ for all μ, following the notation (2.31)).

4.1.1 Convergence analysis and computational cost of the RB method

The left part of Figure 4.2 depicts the fast convergence of the reduced basis method with respect to the size of the reduced space. The relative errors on the eigenfunctions $e_N^{u,rel}(\mu)$ and $e_{N,L^2}^{u,rel}(\mu)$ follow the same trend. The relative error $e_N^{k,rel}(\mu)$ between the high-fidelity solution and the reduced basis solution on the multiplication factor k_μ reaches the order of 10^{-5} for $N = 100$. Moreover, this error decreases by 4 orders of magnitude from $N = 10$ to $N = 100$. As expected, the error on the eigenvalue decreases twice faster than the error on the eigenvector. Moreover, we checked that the value of the a posterior error estimator $\eta_N^k(\mu)$ stays below 10^{-12} for the selected parameters, as expected.

In terms of computational time, the right part of Figure 4.2 shows that, in the chosen setting, while the high-fidelity solution is computed in about 5.8s, the reduced solution is computed within up to 0.09s, which is overall 60 up to 115 times faster than the high-fidelity solver to obtain a relative error of order 10^{-4} to 10^{-5} on the eigenvalue.

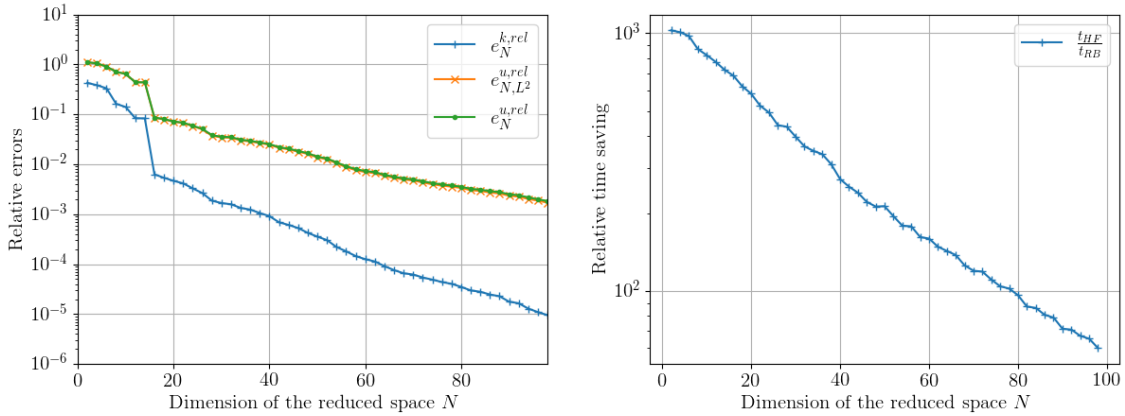

Figure 4.2: (Left) Mean relative errors over $\mathcal{P}_{\text{test}}$; (Right) Relative time saving factor $\dfrac{t_{\text{HF}}}{t_{\text{RB}}}$ as a function of the dimension of the reduced space N.

4.1.2 Certification of the RB method via prefactor-free *a posteriori* error estimates

It is also interesting to look at the behavior of the implemented *a posteriori* error estimators. The relation between the error $e_N^k(\mu)$ and the estimator $\Delta_N(\mu) = \eta_N^k(\mu)$ we used here can be first analyzed by looking at the prefactor $C_N^k(\mu)$, defined in (2.30). The value of $C_N^k(\mu)$ on the test set $\mathcal{P}_{\text{test}}$ is presented in Figure 4.3. In that particular case, we fall into the framework developed in Section 2.4.2. Indeed, when we compute the perturbation magnitude ε_μ as

$$\varepsilon_\mu = \frac{\left\| \frac{A_\mu - A_\mu^T}{2} \right\|}{\left\| \frac{A_\mu + A_\mu^T}{2} \right\|}, \tag{4.1}$$

we observe that ε_μ varies between 3×10^{-7} and 3×10^{-6} for $\mu \in \mathcal{P}_{\text{test}}$. Therefore, we expect $C_N^{k,\text{sym}}(\mu)$ defined in (2.35) to be a a good approximation of $C_N^k(\mu)$. Unfortunately, this is not always the case as we observe on the left plot of Figure 4.3. Actually, in the cases where the prefactors differ a lot, we observe that condition (2.42) in Proposition (2.4.8) is not satisfied, which explains why the perturbative expansions may not be sharp.

Figure 4.4 compares the behavior of the simple *a posteriori* error estimators $\|R_N(\mu)\|$, $\|R_N^*(\mu)\|$ and $\eta_N^k(\mu)$ defined in (2.29), with the corresponding errors $e_N^u(\mu)$, $e_N^{u^*}(\mu)$, and $e_N^k(\mu)$ over the dimension of the reduced space. The plots of the true errors and the corresponding estimators are parallel for $N \geq 20$, illustrating the similar convergence rate for the computed *a posteriori* estimators as the associated real errors. Actually, the quantity $\eta_N^k(\mu)$ seems to be a reliable and efficient *a posteriori* estimate of the true error up to roughly a constant multiplicative factor over a large range of parameter values, as Figure 4.3 illustrates.

In terms of absolute value, for $N = 100$, the estimator $\eta_N^k(\mu)$ for the multiplication factor is about 10^{-2} while the true error is approximately 10^{-4}: this illustrates the importance of introducing prefactors $\overline{C}_N^k(\mu)$, $\overline{C}_N^u(\mu)$ and $\overline{C}_N^{u^*}(\mu)$ to estimate the true errors, see (2.31), and in particular to stop the greedy procedure once the real error is below a given threshold (see Section 3.3.3). This will be discussed in the next test case below.

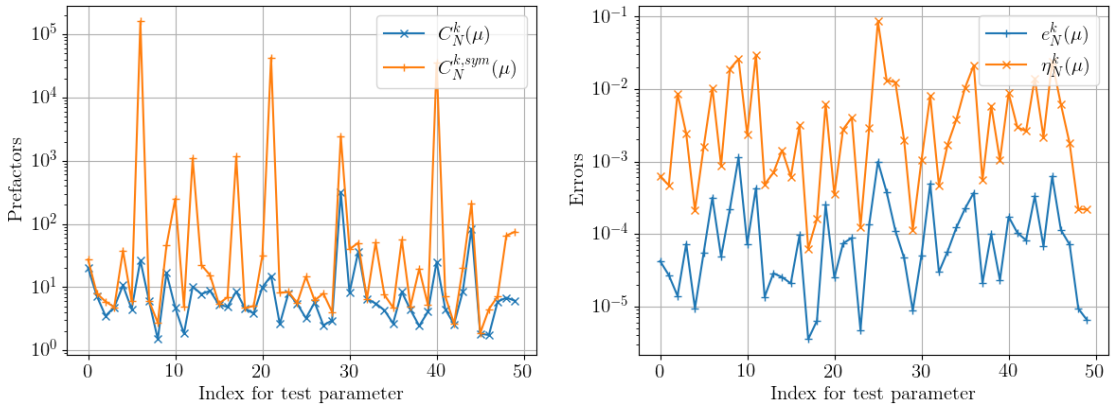

Figure 4.3: (Left) Variations of the prefactor $C_N^k(\mu)$ and $C_N^{k,\mathrm{sym}}(\mu)$ over $\mathcal{P}_{\mathrm{test}}$ for $N = 100$. (Right) Parametric variations of the real eigenvalue error $e_N^k(\mu)$ (in blue) and the associated *a posteriori* error estimator $\eta_N^k(\mu)$ (in orange) over $\mathcal{P}_{\mathrm{test}}$, for $N = 100$.

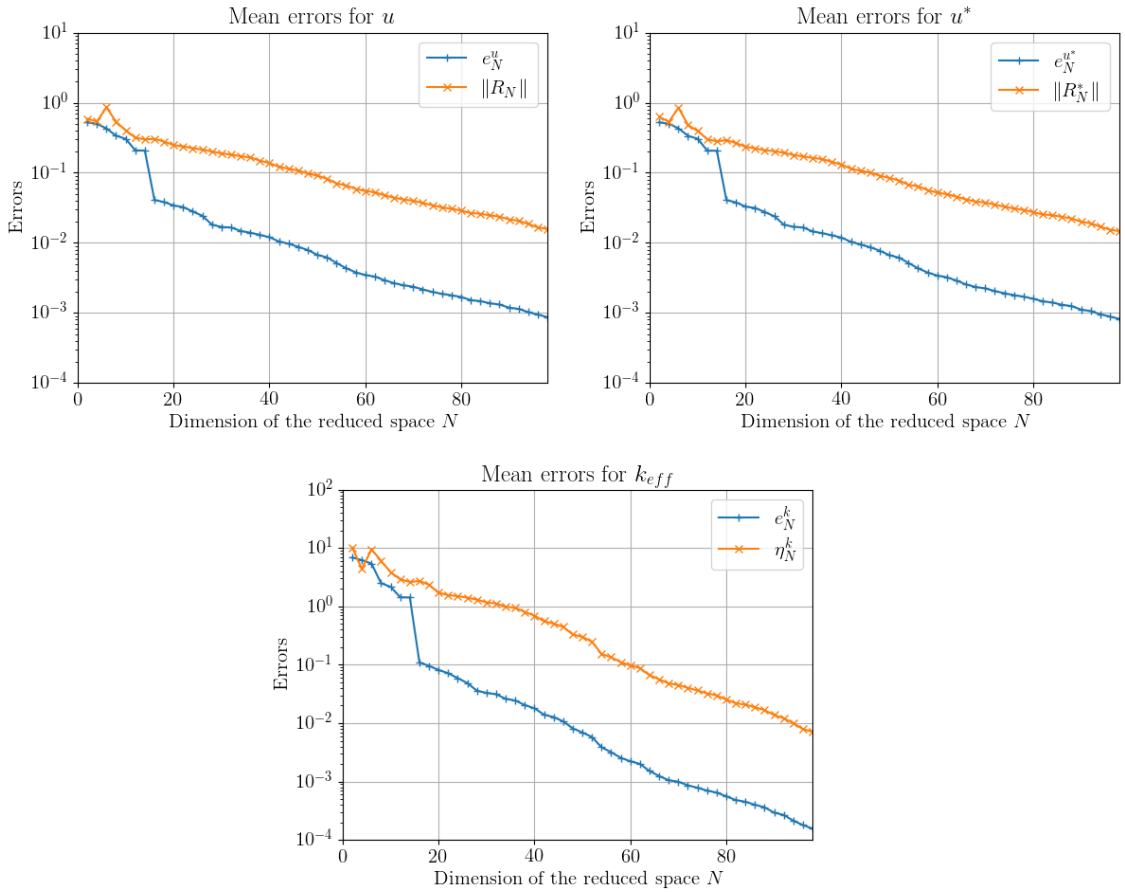

Figure 4.4: Mean values for errors and associated *a posteriori* error estimators over $\mathcal{P}_{\mathrm{test}}$. (Left) e_N^u and $\|R_N\|$; (Middle) $e_N^{u^*}$ and $\|R_N^*\|$; (Right) e_N^k and η_N^k.

4.2 The *Minicore* problem

We now provide a second, more challenging, test case called *Minicore*. The core is modeled as a square of side length $L = 107.52$ cm. As Figure 4.5 shows, it is constructed out of $K =$

25 assemblies (1 fuel assembly composed of a mix of uranium dioxyde and Gadolinium oxyde denoted UGD12 + 8 fuel assemblies composed of uranium dioxyde labeled UO2 + 16 radial reflector assemblies named REFR), each being 21.504 cm long. It is discretized into $\mathcal{N} = 2602$ degrees of freedom. Here, $B_\mu \neq I$, and the Dirichlet boundary condition in Problem (1.41) is replaced by a Robin-type vacuum boundary condition

$$D^i(r,\mu)\nabla\phi^i(r,\mu).\vec{n} + \frac{1}{2}\phi^i(r,\mu) = 0 \quad \text{on } \partial\mathcal{R}, \quad 1 \leq i \leq 2,$$

where $\vec{n}$ is the outward unit normal vector to $\partial\mathcal{R}$.

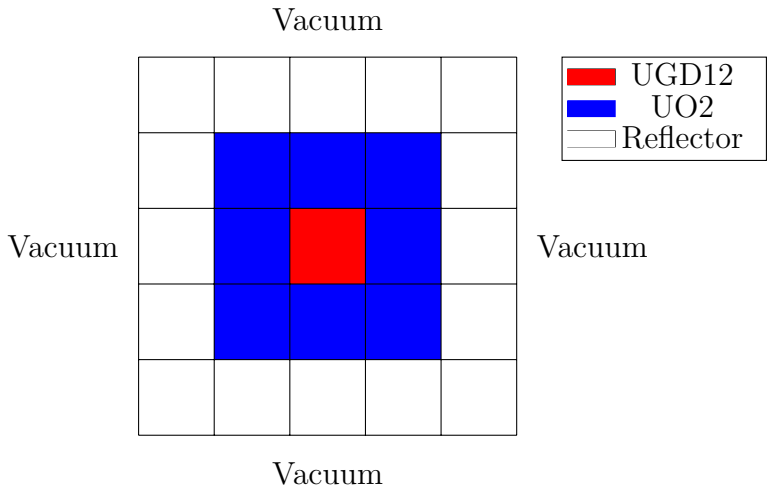

Figure 4.5: Median cross-sectional view of the *Minicore*

In this test case, the parameter μ contains five values which determine all the physical parameters entering (1.41). More precisely, by recalling the partition $(\mathcal{R})_{k=1}^K$ of the domain $\mathcal{R}$, the parameter set $\mathcal{P}$ is the $5K$ dimensional vector space

$$\mathcal{P} = \left\{\mu = (\mu_1, \ldots, \mu_K)\,,\, \forall 1 \leq k \leq K,\, \mu_k \in \mathbb{R}^5\right\},$$

such that μ_k contains the following information attached to the subdomain $\mathcal{R}$:

- the nature of the material in $\mathcal{R}_k$;
- the burnup value, in MWd/ton;
- the fuel temperature, in K;
- the boron concentration, in particle per million (ppm);
- the moderator density.

The parametric sets $\mathcal{P}_{\text{train}}$ and $\mathcal{P}_{\text{test}}$ are randomly generated in $\mathcal{P}$ such that

$$\begin{cases} \#\mathcal{P}_{\text{train}} = 1000 \\ \#\mathcal{P}_{\text{test}} = 50 \\ \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset. \end{cases}$$

Regarding the offline stage, in order to avoid any stability issue, a POD procedure over a reduced space of dimension 10 (generated from 5 direct plus 5 adjoint eigenvectors snapshots) is used to initialize the greedy procedure (see Section 3.3.1). Then, the greedy procedure is performed using the *a posteriori* estimator $\|R_N\| + \|R_N^*\|$, as the quantity of interest here is the two-group flux (ϕ^1, ϕ^2) as well as its adjoint $(\phi^{*,1}, \phi^{*,2})$.

4.2.1 Convergence analysis and computational cost of the RB method

The left part of Figure 4.6 depicts mean relative errors $e_N^{k,\text{rel}}$, $e_{N,L^2}^{u,\text{rel}}$, and $e_N^{u,\text{rel}}$ as a function of the dimension of the reduced basis. The relative error on the multiplication factor is of the order of 10^{-5} for $N = 80$. Typically, as the left part of Figure 4.7 shows, for a certain $\mu_0 \in \mathcal{P}$ and for $N = 100$, the maximum point-wise error on the associated first-group flux does not exceed 3.2×10^{-4}; as for the second group, the right part of Figure 4.7 shows that the flux error is locally gathered in an area of low flux, quite far from the hot spot.

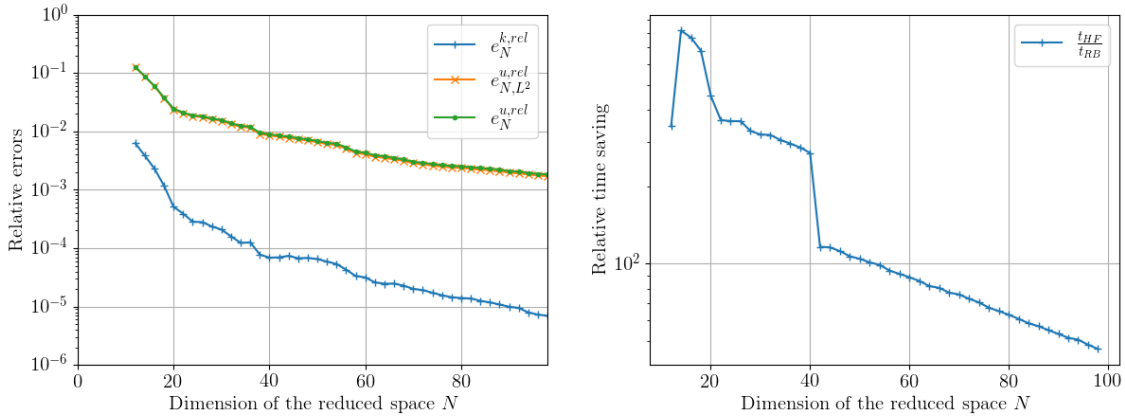

Figure 4.6: (Left) Mean relative errors over $\mathcal{P}_{\text{test}}$; (Right) Relative time saving factor $\dfrac{t_{\text{HF}}}{t_{\text{RB}}}$ as a function of the dimension of the reduced space N.

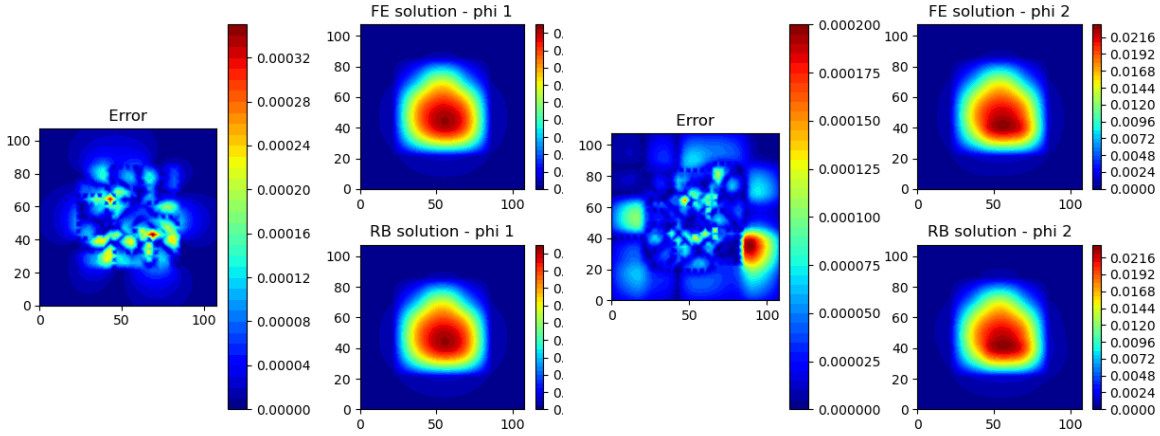

Figure 4.7: (Left) Plots of the first energy group of high-fidelity (upper right) and reduced (lower right) solutions u_{μ_0} and $u_{\mu_0,N}$, and their error (left) $|u_{\mu_0} - u_{\mu_0,N}|$, for $N = 100$ and for $\mu_0 \in \mathcal{P}_{\text{test}}$.
(Right) Plots of the second energy group of high-fidelity (upper right) and reduced (lower right) solutions u_{μ_0} and $u_{\mu_0,N}$, and their error (left) $|u_{\mu_0} - u_{\mu_0,N}|$, for $N = 100$ and for $\mu_0 \in \mathcal{P}_{\text{test}}$

Importantly, the reduced method enables the solution to be computed faster than the high-fidelity approach, which typically takes about 4.56 s to be computed for the present test case. The right part of Figure 4.6 illustrates that the relative saving time

factor is a decreasing function of the dimension of the reduced space N, and exhibits a large computational gain compared to the high-fidelity solver. It is observed that for a relative error on k_{eff} ranging from 10^{-4} to 10^{-6}, the reduced solution can be obtained with a computational time from 50 up to 300 times smaller than the high-fidelity solution.

Finally, we gather in Table 4.1 the measured computational times for several quantities of interest and main stages obtained using Python. Overall, the reduced basis method is very useful when the number p of solutions that must be computed is very large, such as in an optimization process. Roughly, if t_{offline} denotes the computational time of the *offline* stage, t_{HF} the *high-fidelity* solver computational time, and t_{RB} the reduced solver computational time, the reduced basis method becomes relevant when

$$t_{\text{offline}} + p \times t_{\text{RB}} < p \times t_{\text{HF}},$$

that is

$$p > \frac{t_{\text{offline}}}{t_{\text{HF}} - t_{\text{RB}}}.$$

For this test case, this corresponds to $p > 1743$ parameter values.

	Mean computational time
Offline stage (t_{offline})	$\approx$ 11 hours
Assembling residual norm (*offline* part)	49.19 s
Assembling residual norm (*online* part)	5.03 s
Solving the *high-fidelity* problem (t_{HF})	14.71 s
Solving the reduced problem (t_{RB})	0.44 s

Table 4.1: Mean computational times for the Efficient Greedy reduced basis method applied to the 2D two-group *Minicore* in Python, for $N = 100$

4.2.2 Certification of the RB method via practical *a posteriori* error estimates

We now study the certification of the method performed by the *a posteriori* error estimator. Figure 4.8 shows that, although the residuals display similar values as those for the real eigenvector errors, for the eigenvalue, the order of magnitude of the *a posteriori* estimator is roughly 10 times larger than the real error, for $N \geq 30$. Despite the fairly good parametric variations of the estimate, illustrated in Figure 4.9, the gap between real error and estimator must be corrected in order to implement a relevant stopping criterion in the greedy algorithm. This points out a certain variation of the prefactor $C_N^k(\mu)$ over the dimension of the reduced space N. In order to bring a correction to the model, the practical efficiency of the estimator proposed in Section 2.4.3 is computed. The right plot of Figure 4.9 shows that the efficiency $\mathcal{E}_N^k$ defined in (2.43) levels off for $N = 100$ at the order of magnitude of 10^{-1}, and does not depend too much on the parameter μ. Therefore, we propose to apply the procedure outlined in Section 2.4.3 to build a posteriori error estimators of the form (2.46), with constants $\overline{C}_N^k$, $\overline{C}_N^u$ and $\overline{C}_N^{u^*}$ approximated by (2.45). This requires to choose a set $\mathcal{P}_{\text{pref}}$, that we randomly chose in $\mathcal{P}$ such that

$$\begin{cases} \#\mathcal{P}_{\text{pref}} = 10 \\ \mathcal{P}_{\text{pref}} \cap \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset. \end{cases}$$

As a result of this procedure, Figure 4.10 shows that the order of magnitude of the modified estimator corresponds to the one of the real error, showing that the new *a posteriori* estimator can be used as an optimal stopping indicator.

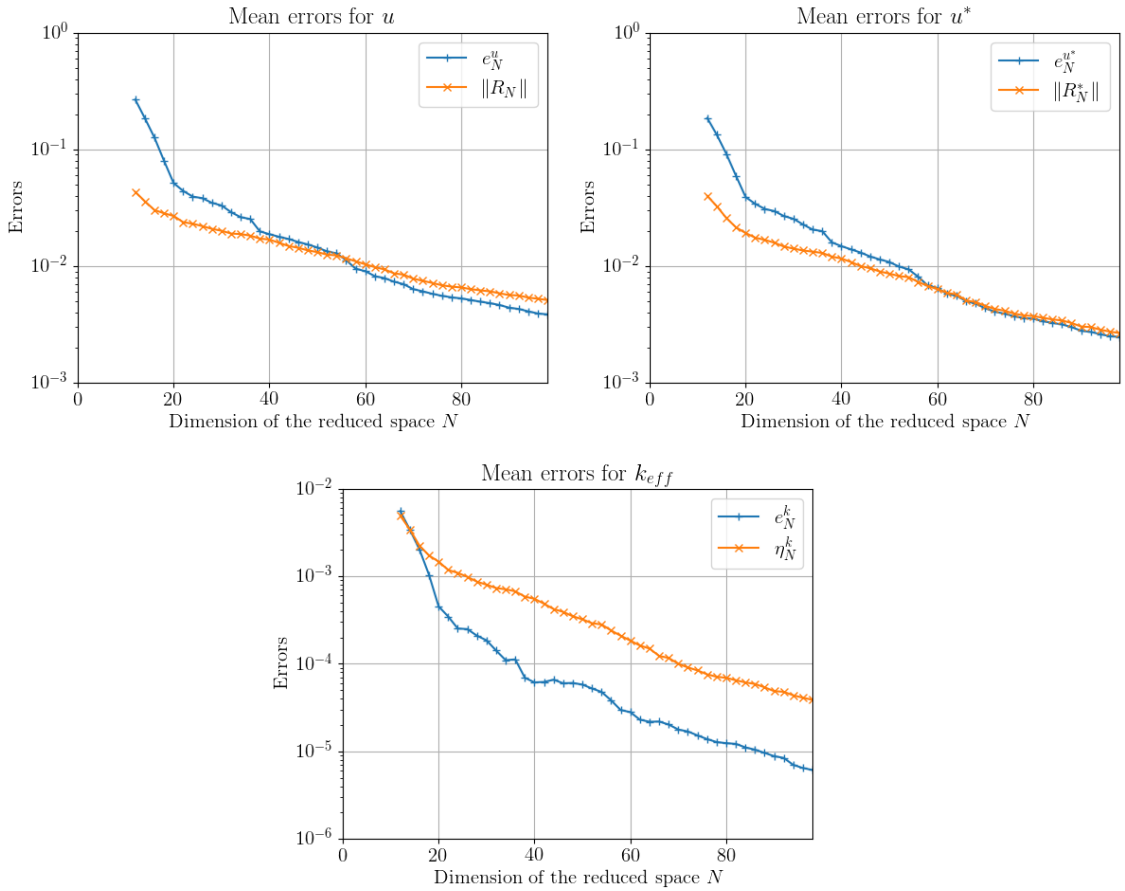

Figure 4.8: Mean values for errors and associated *a posteriori* error estimators over $\mathcal{P}_{\text{test}}$. (Left) e_N^u and $\|R_N\|$; (Middle) $e_N^{u^*}$ and $\|R_N^*\|$; (Right) e_N^k and η_N^k.

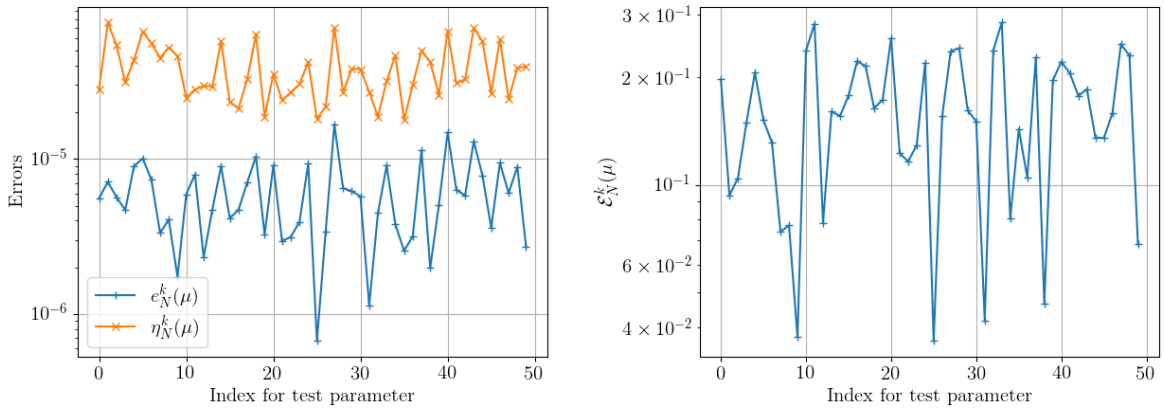

Figure 4.9: (Left) Parametric variations of the real eigenvalue error $e_N^k(\mu)$ (in blue) and its associated *a posteriori* error estimator $\eta_N^k(\mu)$ (in orange) over $\mathcal{P}_{\text{test}}$, for $N = 100$; (Right) Parametric variations of the practical efficiency $\mathcal{E}_N^k(\mu)$ over $\mathcal{P}_{\text{test}}$, for $N = 100$.

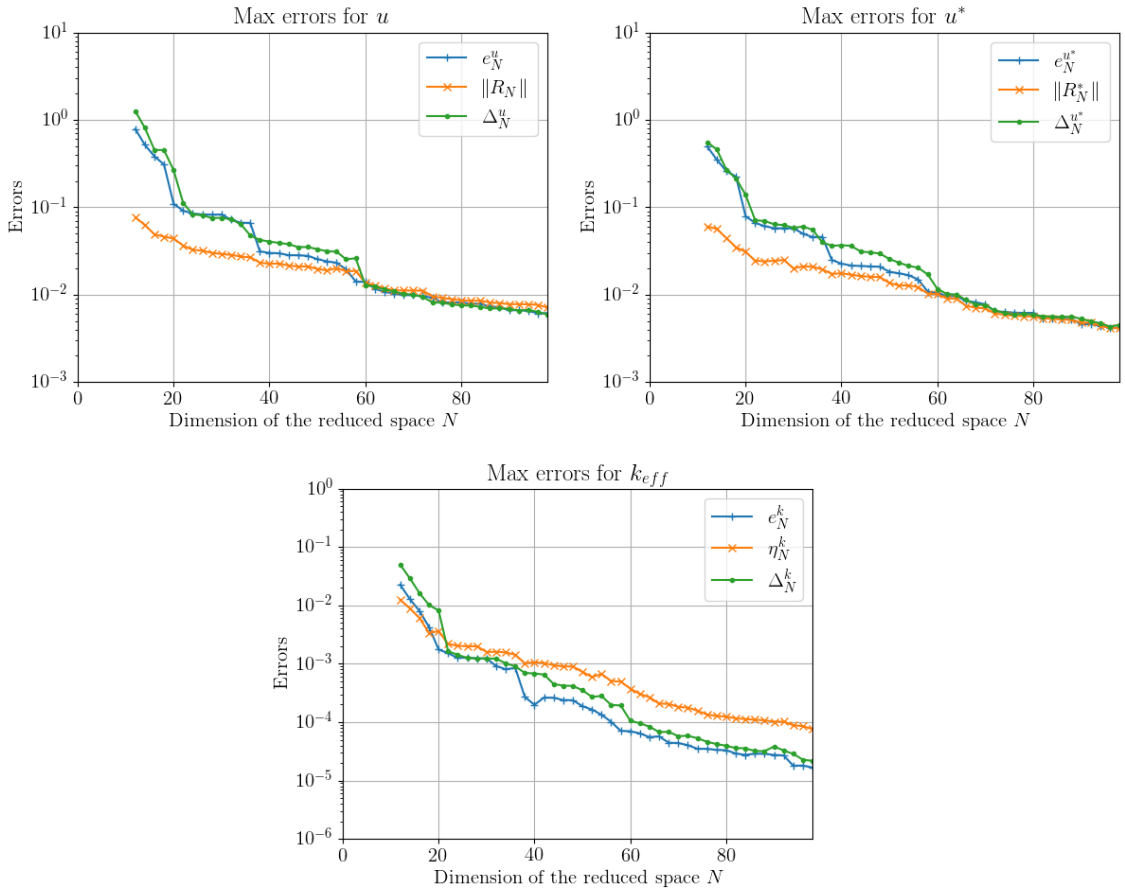

Figure 4.10: Maximum values for errors and associated *a posteriori* error estimators over $\mathcal{P}_{\text{test}}$. Upper left: e_N^u, $\|R_N\|$ and Δ_N^u; upper right: $e_N^{u^*}$, $\|R_N^*\|$ and $\Delta_N^{u^*}$; lower: e_N^k, η_N^k and Δ_N^k.

4.3 A 3D PWR core with homogenized fuel assemblies

Many thanks are addressed to Prof. J. Ragusa for providing this test case as well as the associated codes.

We consider a quarter of a PWR core in the three dimensions, that corresponds to benchmark BSS-11 published in the Argonne (ANL) Benchmark Problem Book [81]. The geometry of the core is depicted in Figure 4.11.

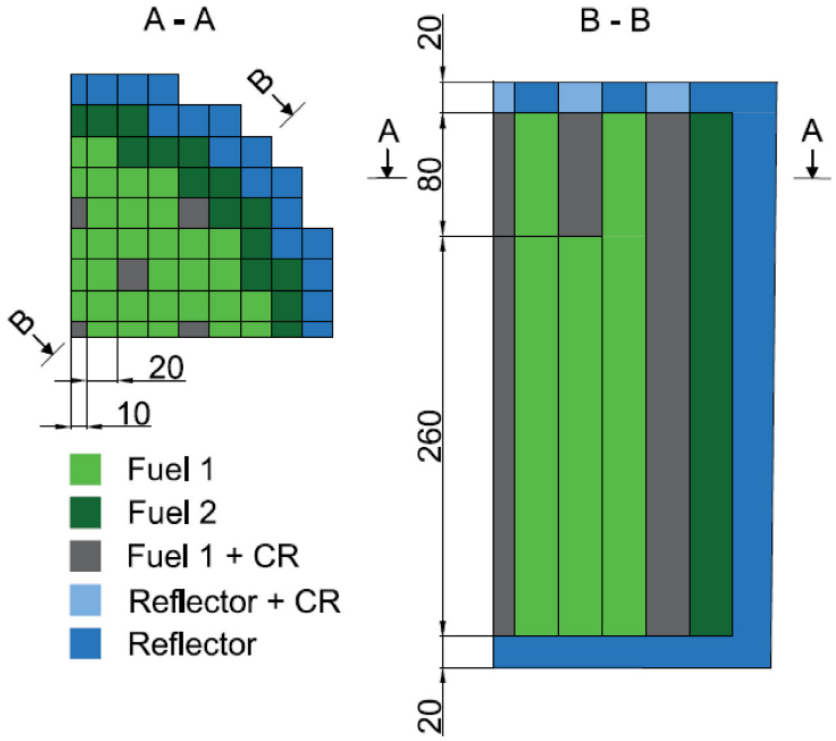

Figure 4.11: Cross-sectional views of BSS-11.
Source: [57]

The core is made of five material regions and it is endowed with reflective boundary conditions along the symmetry planes, and a Marshak-type vacuum boundary condition elsewhere at the border of the core. The high-fidelity discretization yields $\mathcal{N} = 2 \times 36632$ degrees of freedom, and it is performed by the RESOLVED (Reduced Eigenvalue SOLVEr for Diffusion) code [57]. The parametrization of the core lies in a $\pm$ 20% variation of the cross-sections values given in [81].

The training space $\mathcal{P}_{\text{train}} = \{\mu_1, \ldots, \mu_N\}$ consists of $N = 10$ parameter sets, generated via Latin Hypercube Sampling (LHS). We set here $\mathcal{P}_{\text{test}} = \mathcal{P}_{\text{train}}$ as sanity check. The goal here is to compare two ROM approaches: a first approach, that was notably tested in [57], named the "Direct" method, which consists in the reduced space

$$\mathcal{V}_1 = \text{Span}\{u_{\mu_1}, \ldots, u_{\mu_N}\},$$

and a second method, which takes over our strategy to consider both direct and adjoint problems in the construction of the reduced space, and therefore named "Direct+Adjoint", which then consists in the reduced space

$$\mathcal{V}_2 = \text{Span}\{u_{\mu_1}, u^*_{\mu_1} \ldots, u_{\mu_N}, u^*_{\mu_N}\}.$$

Each reduced space comes from a Singular Value Decomposition (SVD) of the associated family of eigenvectors. We gather in Table 4.2 the different errors on the effective multiplication factor for both "Direct" and "Direct+Adjoint" reduced-basis methods.

e_N^k (in pcm)	$\mathcal{V}_1$	$\mathcal{V}_2$
μ_1	9463.78	2.66454×10^{-10}
μ_2	7.10543×10^{-10}	4.21885×10^{-10}
μ_3	6.66134×10^{-10}	5.77316×10^{-10}
μ_4	18946.9	3.35132×10^{-10}
μ_5	6126.34	5.1561×10^{-10}
μ_6	3.51501×10^{-5}	4.00065×10^{-5}
μ_7	19194.3	8.17124×10^{-8}
μ_8	1.75082×10^{-8}	7.21645×10^{-9}
μ_9	9773.44	4.59022×10^{-7}
μ_{10}	2.66454×10^{-10}	1.77636×10^{-10}

Table 4.2: Eigenvalue errors for both "Direct" ($\mathcal{V}_1$) and "Direct+Adjoint" ($\mathcal{V}_2$) reduced order models

It is obvious that the "Direct+Adjoint" approach yields better results, as the eigenvalue error remains below 10^{-9} overall, while the naive "Direct" approach yields spurious eigenvalues for half of the parameters, with gigantic errors of the order of magnitude of 10^{-2}, or even 10^{-1}, when errors close to zero are expected. This example is the proof that the *a priori* error analysis introduced at the beginning of Chapter 2 is a key element in our study.

***

The three test cases presented in this chapter illustrate the ability of the newly proposed RB method to provide an inexpensive, reliable and certified reduced-order model for the neutron diffusion equations. The two first test cases show that the proposed RB model is able to give the k-effective of the HF model at the order of the pcm (1 pcm $= 10^{-5}$), within a computational time of the order of the millisecond, while it confirmed that the heuristic approach in the development of reliable *a posteriori* error estimates yields acceptable error surrogates for this problem. At last, the POD approach for the PWR benchmark illustrates the necessity of including the contribution of the adjoint problem in the RB model, in order to get a reliable reduced-order model.

Chapter 5

Towards an application of reduced basis methods to core computation in APOLLO3®

In this chapter, we are interested in the parameterized neutron diffusion equation, when it is solved multiple times for different values of the parameters, e.g. in optimization problems. This is often called a "multi-query context". Let us focus on the multigroup approximation over an energy range $[E_{\min}, E_{\max}] = [E_G, E_{G-1}] \cup \ldots \cup [E_1, E_0]$, where G stands for the given number of neutron energy groups. Given a parameter μ, the steady-state neutron diffusion equation [48, Chapter 7] (see Section 1.4.2) seeks the multigroup neutron scalar flux $\phi_\mu = \left(\phi_\mu^1, \ldots, \phi_\mu^G\right)$ associated with the multiplication factor $k_{\text{eff},\mu}$ (the largest eigenvalue in modulus) inside the nuclear reactor core $\mathcal{R}$ such that

$$\mathbb{L}^g_{\text{diff},\mu}\phi^g_\mu - \mathbb{H}^g_{\text{diff},\mu}\phi_\mu = \frac{1}{k_{\text{eff},\mu}}\mathbb{F}^g_{\text{diff},\mu}\phi_\mu, \quad \forall g = \{1, \ldots, G\}, \quad \text{in } \mathcal{R}, \tag{5.1}$$

and vacuum boundary conditions (as the ones used in Section 4.2) on $\partial\mathcal{R}$ where $\mathcal{R}$ is a bounded and open subset of $\mathbb{R}^3$. The advection operator $\mathbb{L}^g_{\text{diff},\mu}$, the scattering operator $\mathbb{H}^g_{\text{diff},\mu}$ and the fission operator $\mathbb{F}^g_{\text{diff},\mu}$ are defined by

- $\mathbb{L}^g_{\text{diff},\mu}\phi^g_\mu = -\text{div}\left(D^g_\mu \nabla\phi^g_\mu\right) + \Sigma^g_{t,\mu}\phi^g_\mu$;
- $\mathbb{H}^g_{\text{diff},\mu}\phi_\mu = \displaystyle\sum_{g'=1}^{G} \mathbb{H}^{g'\to g}_\mu \phi^{g'}_\mu$, where $\mathbb{H}^{g'\to g}_\mu \phi^{g'}_\mu := \Sigma^{g'\to g}_{s,0,\mu}\phi^{g'}_\mu$;
- $\mathbb{F}^g_{\text{diff},\mu}\phi_\mu = \displaystyle\sum_{g'=1}^{G} \mathbb{F}^{g',g}_\mu \phi^{g'}_\mu$, where $\mathbb{F}^{g',g}_\mu \phi^{g'}_\mu := \chi^g_\mu (\nu\Sigma_f)^{g'}_\mu \phi^{g'}_\mu$;

where D^g_μ, $\Sigma^g_{t,\mu}$, χ^g_μ, ν^g_μ and $\Sigma^g_{f_\mu}$ are respectively, for the group g, the diffusion coefficient, the total cross-section, the total spectrum, the average number of neutrons emitted per fission, the fission cross-section, and $\Sigma^{g'\to g}_{s,0,\mu}$ is the Legendre moment of order 0 of the scattering cross-section from group g' to group g. We introduce a partition $(\mathcal{R}_m)_{m=1}^M$ of the domain $\mathcal{R}$ with $M \in \mathbb{N}^*$ so that for all $1 \le m \le M$, $\mathcal{R}_m$ is a domain with Lipschitz, piecewise regular boundaries. For $g, g' \in [1, \ldots, G]$, the coefficients D^g_μ, $\Sigma^g_{t,\mu}$, $\Sigma^{g'\to g}_{s,0,\mu}$, χ^g_μ, $(\nu\Sigma_f)^g_\mu$ are assumed to be piecewise regular on each domain $\mathcal{R}_m$ for $1 \le m \le M$.

Several reduced-order models have been proposed in this context [57, 87, 24]. In this work, we propose a reduced basis (RB) approach, see [102] for a general introduction

and [32, 107] for applications in neutronics. The method relies on an approximation of the manifold of solutions using a Proper Orthogonal Decomposition (POD) approach. As in the RB methodology described in Chapter 3, the method is composed of two stages. In the *offline* stage, we build a reduced space which approximates the manifold. In the *online* stage, for any given new set of parameters, we solve a reduced problem on the reduced space within a much smaller computational time than the required time to solve the high-fidelity problem (5.1). Here, we focus on the development in the project APOLLO3® [96], a shared platform among CEA, FRAMATOME and EDF, which includes different deterministic solvers for the neutron transport equation. Particularly, we are interested in the MINARET solver [79] in the diffusion approximation, discretized with discontinuous finite elements.

5.1 The MINARET solver and the high-fidelity core computation

MINARET [79] is a deterministic solver for reactor physics calculations developed in the project APOLLO3® code [109]. MINARET can solve either the multigroup neutron transport or diffusion problem from Problems (1.19) and (1.40). The numerical scheme to compute the multiplication factor k_{eff} is based on the inverse power method (see, e.g., [71]). MINARET uses the S_N discrete ordinate method to deal with the angular variable, and Discontinuous Galerkin Finite Elements to solve spatially the neutron transport equation [103]. It applies the Symmetric Interior Penalty Galerkin method (SIPG) [46, Chapter 4] for the discretization of the neutron diffusion equation (1.40). In all cases, the solver uses cylindrical meshes devised by extrusion of a 2D triangular mesh.

The high-fidelity (HF) problem which stems from the discretization of Problem (5.1) by the MINARET solver reads as

$$\begin{aligned} &\text{Find } (u_\mu, k_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R} \text{ such that} \\ &A_\mu u_\mu = \frac{1}{k_\mu} B_\mu u_\mu, \end{aligned} \tag{5.2}$$

where k_μ is the largest eigenvalue in modulus, $\mu \in \mathcal{P}$ stands for the parametric dependence of the problem with $\mathcal{P}$ a compact set of $\mathbb{R}^d$, $d \geq 1$, A_μ is an invertible non-symmetric matrix, namely the discretized diffusion operator, or disappearance matrix, B_μ is a non-negative non-symmetric matrix, namely the discretized fission operator, or production matrix, and $\mathcal{N}$ is the total number of degrees of freedom of the considered high-fidelity discretization. Typically, for multigroup neutron diffusion calculations, if we denote by $\mathcal{N}_\mathcal{R}$ the total number of spatial degrees of freedom, we have $\mathcal{N} = G \times \mathcal{N}_\mathcal{R}$. In this context and in the numerical applications detailed below, the high-fidelity discretization is such that $\mathcal{N} > 10^5$.

Note that the multiplication factor k_μ is also solution to the following adjoint problem

$$\begin{aligned} &\text{Find } (u^*_\mu, k_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R} \text{ such that} \\ &A^T_\mu u^*_\mu = \frac{1}{k_\mu} B^T_\mu u^*_\mu, \end{aligned} \tag{5.3}$$

where k_μ is the largest eigenvalue in modulus.

The goal of the *offline* is to find a linear space of dimension $n << \mathcal{N}$, denoted by $\mathcal{V}_n$, such that any solution in the manifold

$$\mathcal{M} = \{(u_\mu, k_\mu); \mu \in \mathcal{P}\}$$

can be well-approximated in the space $\mathcal{V}_n$.

5.2 The *offline* stage

To build such a reduced space, we use the information contained in a training space $\mathcal{P}_{\text{train}} = \{\mu_1, \dots, \mu_{n_s}\}$ of n_s parameters. The classical *a priori* error analysis exhibits an upper bound on the eigenvalue error which depends on the error on the left and right eigenvectors [11, 19] (see Section 2.1). Following this insight, the reduced space $\mathcal{V}_n$ is built such that

$$\mathcal{V}_n \subseteq \text{Span}\left(u_\mu, u_\mu^*;\, \mu \in \mathcal{P}_{\text{train}}\right). \tag{5.4}$$

In order to give the best n-rank approximation of the manifold $\mathcal{M}$, we first compute a Singular Value Decomposition (SVD) to the so-called *matrix of snapshots* composed of right eigenvectors

$$S = \left(u_{\mu_1} | \cdots | u_{\mu_{n_s}}\right) \in \mathbb{R}^{\mathcal{N} \times n_s}, \tag{5.5}$$

which writes

$$\begin{aligned} &S = U \Sigma Z^T, \\ &U = (\xi_1 | \dots | \xi_{\mathcal{N}}) \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}, \\ &\Sigma = \text{diag}\left(\sigma_1, \dots, \sigma_{\min(n_s, \mathcal{N})}\right), \\ &Z = (\psi_1 | \dots | \psi_{n_s}) \in \mathbb{R}^{n_s \times n_s}, \end{aligned} \tag{5.6}$$

where the $\sigma_i \in \mathbb{R}_+$ are the singular values of S, sorted in decreasing order, and U and Z are two orthogonal matrices. The reduced space associated with the right eigenvectors comes from a Proper Orthogonal Decomposition (POD) and it is obtained by

$$V^{\text{right}} = (\xi_1 | \dots | \xi_{n_1}), \tag{5.7}$$

where $1 \leq n_1 \leq n_s$, which minimizes the 2-norm error between each snapshot and its orthogonal projection onto the subset of dimension n_1 spanned by the columns of V^{right}. The integer n_1 comes from a truncation of the SVD with respect to a given tolerance criterion related to the singular values. We then proceed similarly to approximate the adjoint manifold $\{(u_\mu^*, k_\mu); \mu \in \mathcal{P}\}$. We perform a SVD to the matrix of snapshots S^* of left eigenvectors, and the POD gives the adjoint reduced space

$$V^{\text{left}} = \left(\xi_1^* | \dots | \xi_{n_2}^*\right), \tag{5.8}$$

where $1 \leq n_2 \leq n_s$. The resulting reduced space $\mathcal{V}_n$ is defined as the sum of spaces V^{right} and V^{left}, using an orthonormalization procedure to obtain a basis, the integer n being the dimension of $\mathcal{V}_n$ ($n \leq n_1 + n_2$).

Remark 5.2.1. *Note here that the ultimate goal is actually to implement a greedy reduced basis method, such as detailed in Chapter 3, for the offline stage. However, at the actual stage of this work, we only rely on POD, as an efficient implementation of the greedy procedure has not been investigated yet.*

5.3 The *online* stage

5.3.1 Assembling the reduced matrices and the reduced problem

Let V_n be a matrix containing an orthonormal basis of the reduced space $\mathcal{V}_n$ as columns. A Galerkin projection of Problem (5.2) is obtained by constructing, for any $\mu \in \mathcal{P}$, the reduced $n \times n$ matrices

$$A_{\mu,n} = V_n^T A_\mu V_n, \tag{5.9}$$

$$B_{\mu,n} = V_n^T B_\mu V_n. \tag{5.10}$$

Assembling such matrices is not trivial, since the high-fidelity sparse matrices A_μ and B_μ are not fully assembled. Indeed,

$$A_\mu = \begin{pmatrix} A_\mu^{1,1} & A_\mu^{1,2} & \cdots & A_\mu^{1,G} \\ A_\mu^{2,1} & A_\mu^{2,2} & \cdots & A_\mu^{2,G} \\ \vdots & \vdots & \ddots & \vdots \\ A_\mu^{G,1} & A_\mu^{G,2} & \cdots & A_\mu^{G,G} \end{pmatrix}, \qquad B_\mu = \begin{pmatrix} B_\mu^{1,1} & B_\mu^{1,2} & \cdots & B_\mu^{1,G} \\ B_\mu^{2,1} & B_\mu^{2,2} & \cdots & B_\mu^{2,G} \\ \vdots & \vdots & \ddots & \vdots \\ B_\mu^{G,1} & B_\mu^{G,2} & \cdots & B_\mu^{G,G} \end{pmatrix} \tag{5.11}$$

where, for $g' \neq g$, the block matrices $A_\mu^{g,g'}$ and $B_\mu^{g,g'}$ are sparse $\mathcal{N}_\mathcal{R} \times \mathcal{N}_\mathcal{R}$ matrices, for $g, g' = \{1, \ldots, G\}$, and the diagonal blocks $A_\mu^{g,g}$ are directly accessible in memory. Therefore, if we decompose the reduced matrix $V_n = (\xi_1 | \ldots | \xi_n)$ along its G group components such that $V_n = \begin{pmatrix} \xi_1^1 & \cdots & \xi_n^1 \\ \vdots & & \vdots \\ \xi_1^G & \cdots & \xi_n^G \end{pmatrix}$, then we have

$$(A_{\mu,n})_{i,j} := (V_n^T A_\mu V_n)_{i,j} = \sum_{g=1}^{G} (\xi_i^g)^T A_\mu^{g,g} \xi_j^g - \sum_{\substack{g,g'=1 \\ g' \neq g}}^{G} (\xi_i^g)^T \mathbb{H}_\mu^{g' \to g} \xi_j^{g'}, \tag{5.12}$$

$$(B_{\mu,n})_{i,j} := (V_n^T B_\mu V_n)_{i,j} = \sum_{g,g'=1}^{G} (\xi_i^g)^T \mathbb{F}_\mu^{g',g} \xi_j^{g'}. \tag{5.13}$$

In terms of complexity, the computation of the reduced matrices $A_{\mu,n}$ and $B_{\mu,n}$ as in (5.12) and (5.13) respectively, is then $O(\mathcal{N}^2)$. Note that in the approach detailed in Chapter 3, we had a complexity of $O(n^2)$, thanks to the affine decomposition of the high-fidelity matrices.

For a given parameter $\mu \in \mathcal{P}$, the reduced problem is then the following:

$$\begin{aligned} &\text{Find } (c_{\mu,n}, k_{\mu,n}) \in \mathbb{R}^n \times \mathbb{R} \text{ such that} \\ &A_{\mu,n} c_{\mu,n} = \frac{1}{k_{\mu,n}} B_{\mu,n} c_{\mu,n}, \end{aligned} \tag{5.14}$$

where $k_{\mu,n}$ is the largest eigenvalue in modulus. We then obtain $u_{\mu,n} = V_n c_{\mu,n}$, as the approximated right eigenvector written in the high-fidelity space $\mathbb{R}^\mathcal{N}$. The associated adjoint problem writes,

$$\begin{aligned} &\text{Find } (c^*_{\mu,n}, k_{\mu,n}) \in \mathbb{R}^n \times \mathbb{R} \text{ such that} \\ &A^T_{\mu,n} c^*_{\mu,n} = \frac{1}{k_{\mu,n}} B^T_{\mu,n} c^*_{\mu,n}, \end{aligned} \tag{5.15}$$

where $k_{\mu,n}$ is again the largest eigenvalue in modulus. Similarly, we obtain $u^*_{\mu,n} = V_n c^*_{\mu,n}$, as the approximated left eigenvector written in the high-fidelity space $\mathbb{R}^{\mathcal{N}}$. In order to solve the reduced problem, we use the power method described in Algorithm 2 (see Section 3.4) with given relative error tolerances and a maximum number of iterations.

5.3.2 Computing errors and error estimates

In order to quantify the approximation error induced by projecting Problems (5.2) and (5.3) on the reduced V_n of dimension $n \in \mathbb{N}^*$, we first normalize all high-fidelity and reduced multigroup fluxes so that $\|u_\mu\|_2 = \|u^*_\mu\|_2 = \|u_{\mu,n}\|_2 = \|u^*_{\mu,n}\|_2 = 1$. Note that $u_{\mu,n}$ and $u^*_{\mu,n}$ are defined up to a sign, and to preserve positivity of the flux, we choose the convention that

$$\langle u_{\mu,n}, u_\mu \rangle_2 \geq 0,$$
$$\langle u^*_{\mu,n}, u^*_\mu \rangle_2 \geq 0.$$

Then, we define the respective following ℓ^2-errors on the eigenvectors and ℓ^2-error on the eigenvalue

$$e^u_{\mu,n} := \|u_\mu - u_{\mu,n}\|_2, \tag{5.16}$$
$$e^{u^*}_{\mu,n} := \|u^*_\mu - u^*_{\mu,n}\|_2, \tag{5.17}$$
$$e^k_{\mu,n} := |k_\mu - k_{\mu,n}|. \tag{5.18}$$

Let us respectively define the residuals on the direct and adjoint flux by

$$R_{\mu,n} := (B_\mu - k_{\mu,n} A_\mu) u_{\mu,n}, \tag{5.19}$$
$$R^*_{\mu,n} := (B^T_\mu - k_{\mu,n} A^T_\mu) u^*_{\mu,n}. \tag{5.20}$$

In the following, we will consider the error estimates $\|R_{\mu,n}\|_2$, $\|R^*_{\mu,n}\|_2$ and $\eta^k_{\mu,n} := \dfrac{\|R_{\mu,n}\| \|R^*_{\mu,n}\|}{\langle c^*_{\mu,n}, A_{\mu,n} c_{\mu,n} \rangle}$ respectively on the reduced direct flux u_μ, the reduced adjoint flux u^*_μ, and the reduced multiplication factor $k_{\mu,n}$ [37]. As in the heuristic approach detailed in Section 2.4.3, we also introduce the *prefactors* C^u_n, $C^{u^*}_n$ and C^k_n defined by

$$C^u_n := \max_{\mu \in \mathcal{P}_{\text{pref}}} \frac{e^u_{\mu,n}}{\|R_{\mu,n}\|_2}, \tag{5.21}$$
$$C^{u^*}_n := \max_{\mu \in \mathcal{P}_{\text{pref}}} \frac{e^{u^*}_{\mu,n}}{\|R^*_{\mu,n}\|_2}, \tag{5.22}$$
$$C^k_n := \max_{\mu \in \mathcal{P}_{\text{pref}}} \frac{e^k_{\mu,n}}{\eta^k_{\mu,n}}, \tag{5.23}$$

where $\mathcal{P}_{\text{pref}} \subset \mathcal{P}$ such that $\mathcal{P}_{\text{pref}} \cap \mathcal{P}_{\text{train}} = \emptyset$.

5.4 Numerical applications to multigroup diffusion core calculation

5.4.1 Convergence analysis of the POD method on benchmark calculations

The POD reduced-basis approach, as implemented in the APOLLO3®code, is first tested on Model 1 Case 1 of Takeda neutronics benchmarks [112]. We refer the reader to Chap-

ter 6 for an associated work, which carried out state estimation techniques, on this test case using a POD reduced basis from power maps [38]. The considered geometry, as shown in Figure 5.1, is a 3D quarter core in the domain $\{(x, y, z) \in \mathbb{R}^3, 0 \leq x \leq 25\,\text{cm}; 0 \leq y \leq 25\,\text{cm}; 0 \leq z \leq 25\,\text{cm}\}$. The MINARET solver is run with $G = 2$ energy groups and $\mathcal{N}_{\mathcal{R}} = 3 \times 10^5$ spatial degrees of freedom. For this high-fidelity solver, we provide a maximum of 500 outer iterations, with relative L^2-error tolerances of 10^{-7} and 10^{-8} on the two-group flux and on the effective multiplication factor, respectively. The reduced solver runs a power iteration method with respectively relative ℓ^2-error tolerances of 10^{-8} and 10^{-9} on the reduced eigenvector and the reduced eigenvalue.

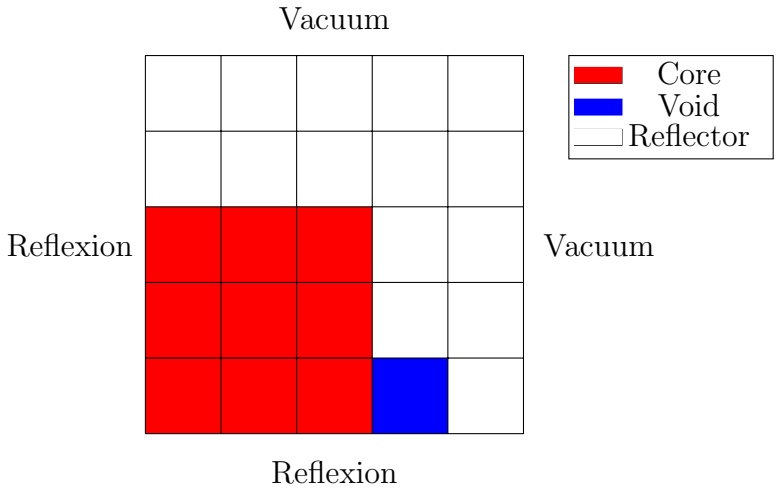

Figure 5.1: Cross-sectional view of the core ($z = 0$ cm)

Here, the parameter μ lies in the 5-dimensional subset $[0.8, 1.2]^5$, and then enables a $\pm 20\%$ variation in the equation coefficients such that, for $\mu = (\mu_1, \ldots, \mu_5) \in [0.8, 1.2]^5$,

$$\begin{aligned} &\left(D^1_\mu, D^2_\mu, \Sigma^1_{a,\mu}, \Sigma^2_{a,\mu}, \Sigma^{1\to 2}_{s,0,\mu}, (\nu\Sigma_f)^1_\mu, (\nu\Sigma_f)^2_\mu, \chi^1_\mu, \chi^2_\mu\right) \\ &= \left(\frac{D^1}{\mu_1}, \frac{D^2}{\mu_2}, \mu_1\Sigma^1_a, \mu_2\Sigma^2_a, \mu_3\Sigma^{1\to 2}_{s,0}, \mu_4(\nu\Sigma_f)^1, \mu_5(\nu\Sigma_f)^2, \chi^1, \chi^2\right), \end{aligned}$$

where, for $g, g' \in \{1, 2\}$, $\Sigma^g_a = (\Sigma^g_t - \Sigma^{g\to g}_{s,0})$, and the values for the coefficients D^g, Σ^g_a, $\Sigma^{g\to g'}_{s,0}$, $(\nu\Sigma_f)^g$ and χ^g are given in Appendix 3 of [112].

We generate a training set $\mathcal{P}_{\text{train}}$ of $n_s = 100$ parameters with a Latin Hypercube Sampling (LHS) over $[0.8, 1.2]^5$. We then compute the SVD of the $2n_s$ *snapshot* matrix as defined in (5.5). The singular values are shown in Figure 5.2. The fast decrease of the singular values illustrates the ability of the training set to approximate the manifold of high-fidelity solutions with a reduced basis of small dimension. Here, for example, the 10 first singular values range from 10^4 to 10^{-1}.

The SVD truncation at the order n then provides a reduced space, and the reduced basis method is tested on the parameter $\mu_{\text{test}} = (1, 1, 1, 1, 1)$ which does not belong to the training set in order to determine to what extent the reduced solver is able to compute a good approximation of the two-group flux and effective multiplicative factor of the Takeda benchmark. The relative errors are depicted in Figure 5.3. For $n = 5$, the reduced solver already returns the same k_{eff} at the order of the pcm (per cent mille) (10^{-5}), and then the error levels off at the order of magnitude of 10^{-7}, as the order of convergence is limited by the convergence criterion of the high-fidelity solver. The solution for the test parameter

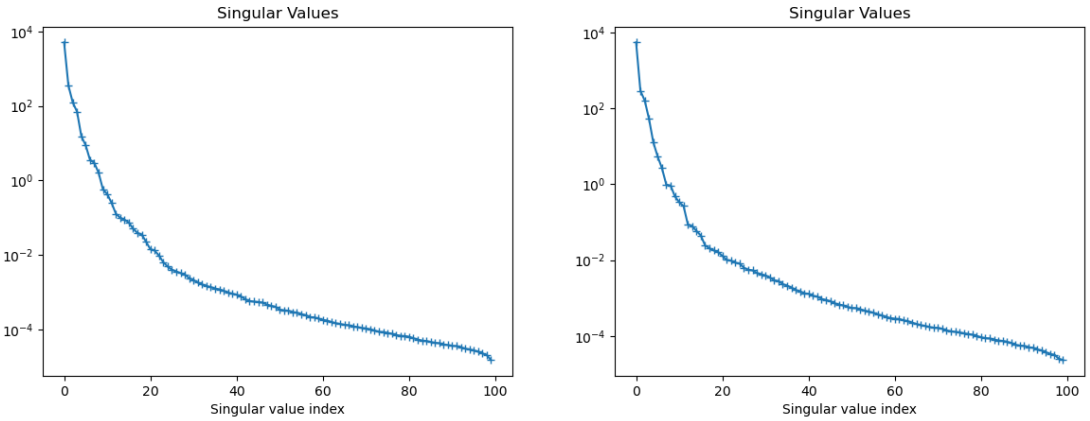

Figure 5.2: Singular values from the SVD of Takeda snapshots, for $n_s = 100$. Left: direct eigenvectors; Right: adjoint eigenvectors.

is thus particularly well represented by the training space, which explains that the error on the k_{eff} already reaches the order of the pcm, for $n = 5$.

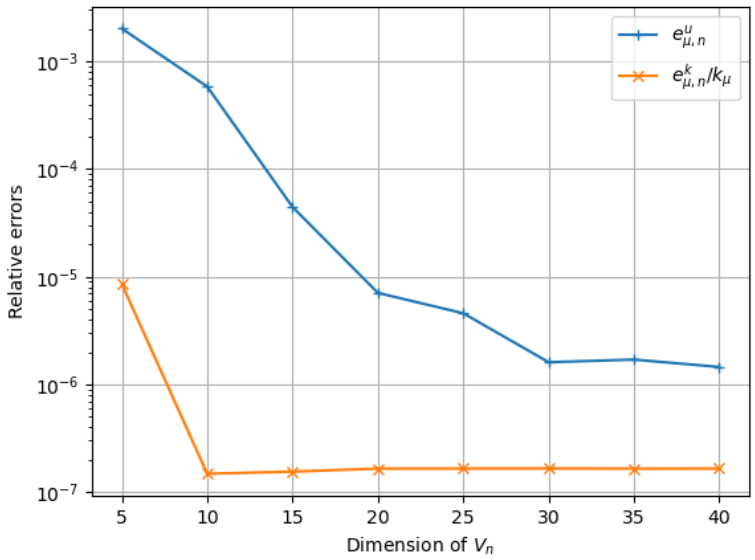

Figure 5.3: Relative errors on the two-group flux and the k_{eff} with respect to the dimension n of the reduced space $\mathcal{V}_n$, for $\mu = \mu_{\text{test}}$

5.4.2 Computational time reduction on a burnup parametrized nuclear core

We now test the POD-RB method on a small nuclear core, namely the *MiniCore* problem (see Section 4.2). The nuclear core geometry is shown in Figure 4.5. It is a 3D nuclear core and the domain is $\{(x, y, z) \in \mathbb{R}^3, 0 \leq x \leq 107.52\,\text{cm}; 0 \leq y \leq 107.52\,\text{cm}; 0 \leq z \leq 468.72\,\text{cm}\}$. The MINARET solver runs with $G = 2$ energy groups and $\mathcal{N}_{\mathcal{R}} = 108800$ degrees of freedom. For this high-fidelity solver, we provide a maximum of 1000 outer iterations, with relative L^2-error tolerances of 10^{-7} and 10^{-8} on the two-group flux and

on the effective multiplication factor, respectively. An example of high-fidelity solution is depicted in Figure 5.4.

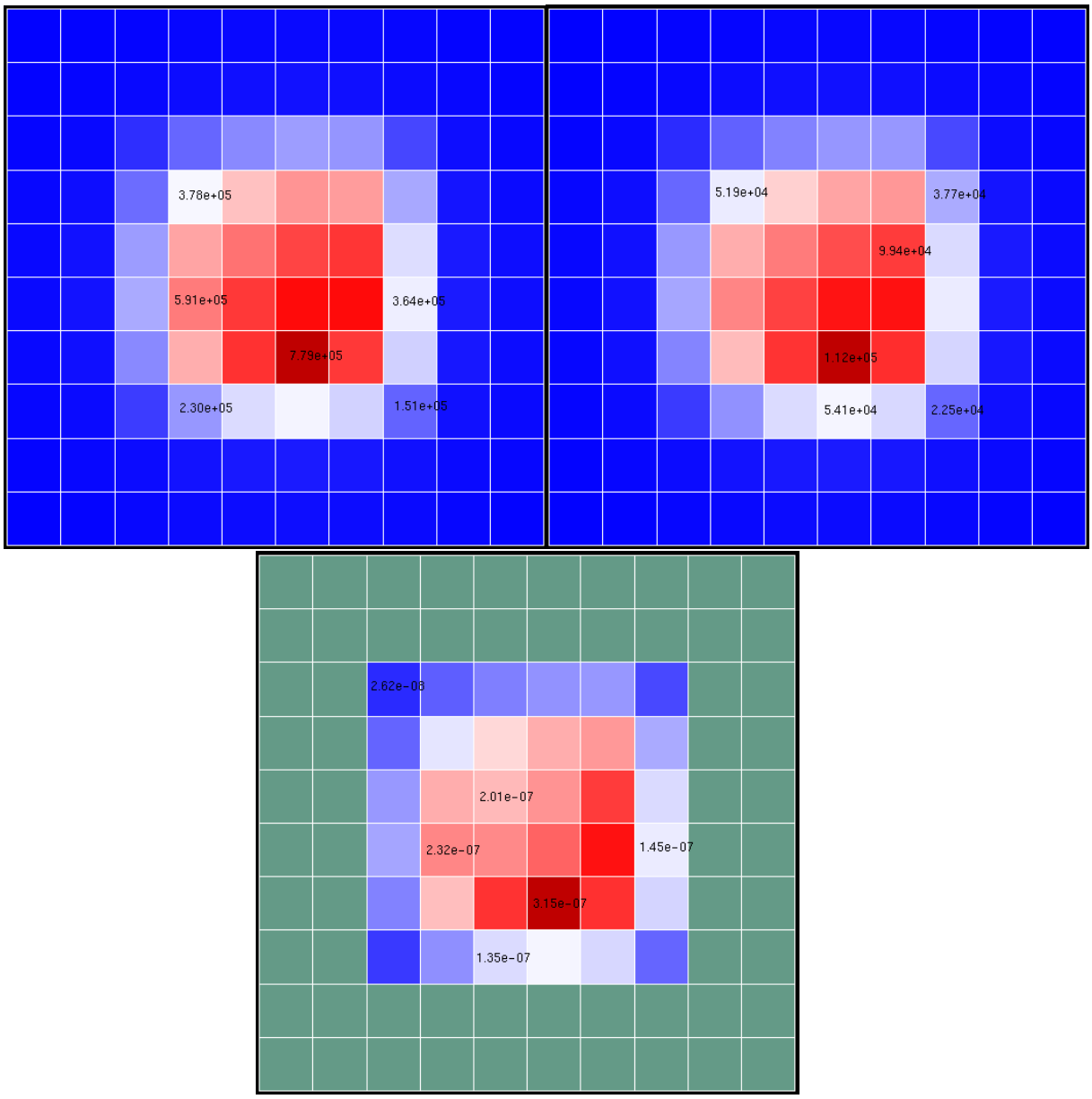

Figure 5.4: Median cross-sectional view of a high-fidelity core calculation with the MINARET solver ($z = 234.36$ cm). Upper left: first-group flux; upper right: second-group flux; lower: power map.

The reduced solver runs a power iteration method with respectively relative ℓ^2-error tolerances of 10^{-8} and 10^{-9} on the reduced eigenvector and the reduced eigenvalue.

The problem is parametrized by the burnup value for the 9 fuel assemblies (one UGD12 and eight UO2). Here, we generate $n_s = 100$ parameters with a Latin Hypercube Sampling (LHS) over the 9-dimensional space

$$\mathcal{P}_{\text{train}} \subset \left\{\mu = (\mu_1, \ldots, \mu_9) \in \mathbb{R}^9;\ \mu_1 \in [0, 72000];\ \mu_2, \ldots, \mu_9 \in [0, 30000]\right\},$$

where μ_1 is the burnup value of the UGD12 assembly and $\mu_2, \ldots, \mu_9$ are the burnup values of the UO2 assemblies, in MWd/ton. Figure 5.5 shows an example of SVD with such a training set with $n_s = 100$. As in the previous test case, a reduced order model can be

obtained for the considered snapshot family as the 25 first singular values range from the order of magnitude of 10^4 to 10^1.

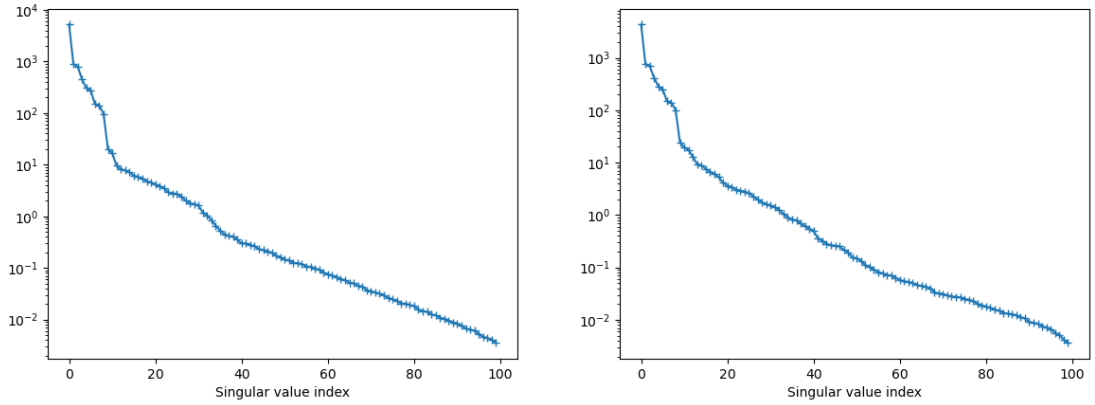

Figure 5.5: Singular values from the SVD of *MiniCore* snapshots, for $n_s = 100$. Left: direct eigenvectors; Right: adjoint eigenvectors.

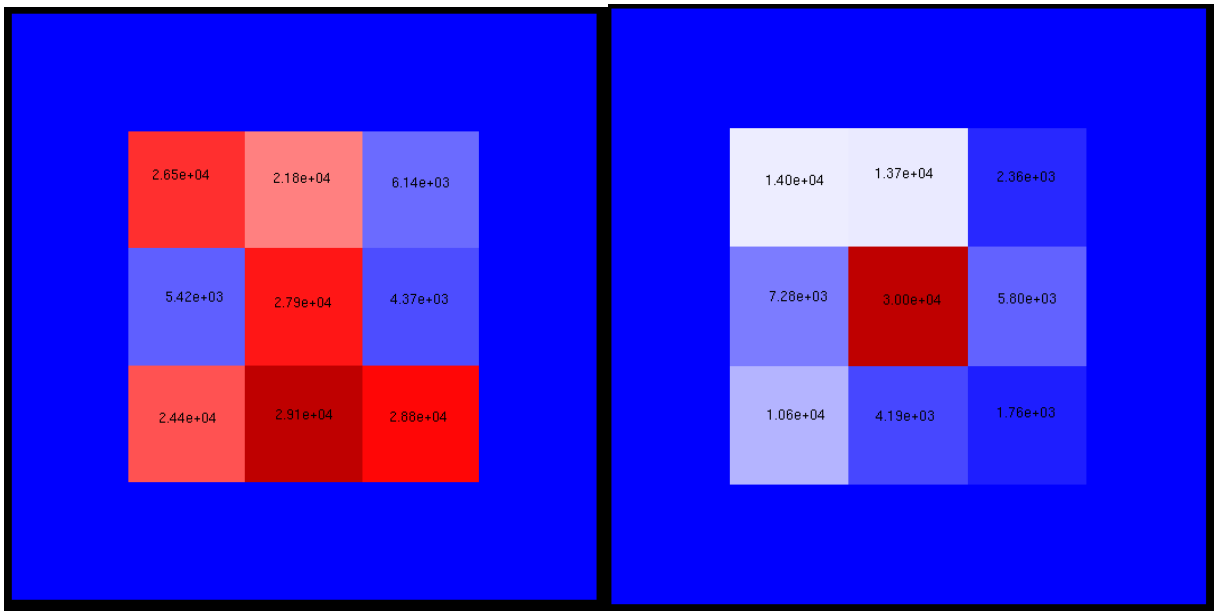

Figure 5.6: An example of burnup map for the *MiniCore*. Values in MWd/tU. Left: $\mu \in \mathcal{P}_{\text{train}}$; Right: $\mu \in \mathcal{P}_{\text{test}}$.

The reduced basis is tested on 10 burnup maps chosen along a LHS over the test space

$$\mathcal{P}_{\text{test}} \subset \left\{ \mu = (\mu_1, \ldots, \mu_9) \in \mathbb{R}^9;\ \mu_1 = 30000;\ \mu_2, \ldots, \mu_9 \in [0, 15000] \right\}.$$

Figure 5.6 illustrates an example of a burnup map for both training and test sets. Figure 5.7 illustrates the convergence of the RB method, as well as the ability for the *a posteriori* error estimates to quantify the approximation. We can see that for $n > 45$, the errors on the direct and adjoint flux and on the k_{eff} are respectively below 10^{-3} and 10^{-5}. Regarding the estimates, their convergence, although they are not at the same rate as those of the real errors, are relevant to their potential use in the construction of such an approximation space. In terms of computational cost, the cost of the *offline* stage

here highly depends on the high-fidelity MINARET solver's cost, as we need to compute high-fidelity solutions for all $\mu \in \mathcal{P}_{\text{train}}$. For the *MiniCore*, one high-fidelity calculation of the k_{eff} runs in nearly 59.15 s, whereas computing the reduced eigenvalue $k_{\mu,n}$ requires a computational time of the order of the millisecond, as Figure 5.8 shows. Note that the procedure that consists of the two SVDs and the orthogonalization of the basis runs in 138 s. Computing the residual norm *online* takes about 0.03 s. Nevertheless, in order to assemble the reduced problem, one needs to compute the reduced matrices, which is particularly expensive in the current naive implementation, as Figure 5.9 shows, as the associated time increases exponentially with respect to the dimension n of the reduced space. Indeed, the complexity is here $O(\mathcal{N}^2)$. Any affine exploitation, as detailed in Chapter 3, would enable a much faster assembling of the reduced matrices, with complexity of $O(n^2)$. However, the structure and the design of the code does not allow an immediate use of this approach. To overcome this cost *online*, we could instead use an interpolation method, such as GEIM [90, 30].

In order to get more reliable *a posteriori* error estimates, we define the set $\mathcal{P}_{\text{pref}}$ such that

$$\mathcal{P}_{\text{pref}} \subset \left\{\mu = (\mu_1, \ldots, \mu_9) \in \mathbb{R}^9;\ \mu_1 \in [0, 72000];\ \mu_2, \ldots, \mu_9 \in [0, 30000]\right\},$$
$$\#\mathcal{P}_{\text{pref}} = 5, \quad \text{and} \quad \mathcal{P}_{\text{pref}} \cap \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset.$$

We then consider the *a posteriori* error estimates

$$\Delta^u_{\mu,n} := C^u_n \|R_{\mu,n}\|_2, \tag{5.24}$$
$$\Delta^{u^*}_{\mu,n} := C^{u^*}_n \|R^*_{\mu,n}\|_2, \tag{5.25}$$
$$\Delta^k_{\mu,n} := C^k_n \eta^k_{\mu,n}, \tag{5.26}$$

where the constants C^u_n, $C^{u^*}_n$ and C^k_n are defined as in (5.21), (5.22) and (5.23) respectively. Figure 5.10 shows that the estimates defined right below are more reliable as they remain of the same order of magnitude as the real errors, independently of the value of n.

***

The two test cases that were developed highlight the possibility of a reduced basis method implementation in the APOLLO3®code, in terms of accuracy and computational time reduction. Note that *a posteriori* error estimators in the reduced basis context may be applied in a greedy approach in the *offline* stage [25, 106, 61], as done in Chapter 3, so that it minimizes calls to the high-fidelity solver, or in an *online* certification of the reduced model. To do so, we should investigate on how to compute the reduced matrices by breaking, as much as possible, their parameter dependency. We could, for example, consider a General Empirical Interpolation Method (GEIM) [90, 30]. This will be the subject of future works.

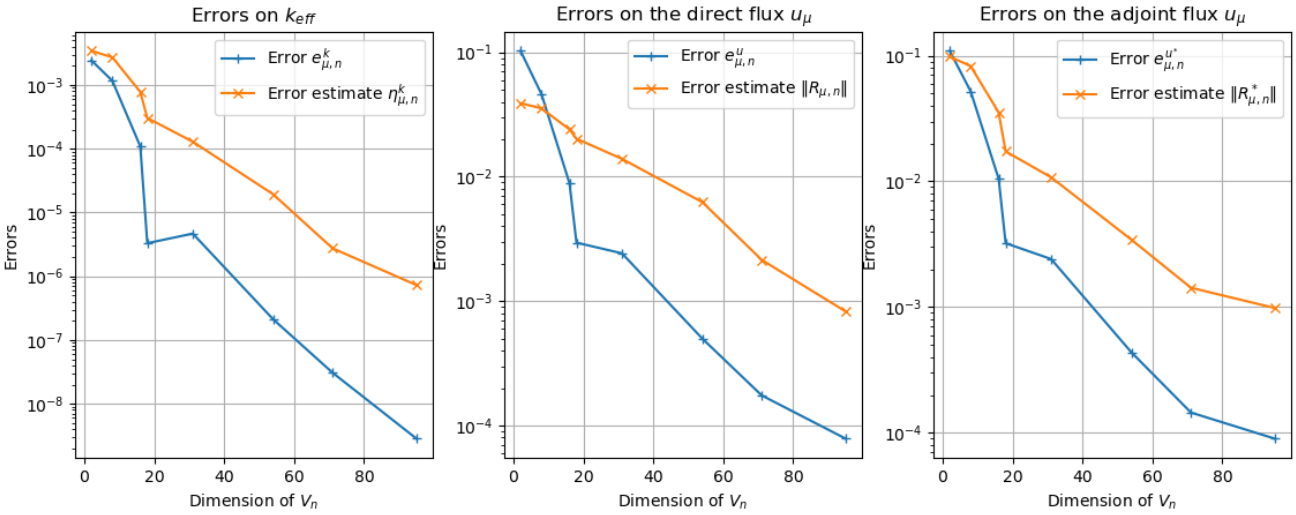

Figure 5.7: Mean errors and their associated error estimates with respect to the dimension n of the reduced space $\mathcal{V}_n$, over $\mathcal{P}_{\text{test}}$. From left to right: error $e^k_{\mu,n}$ and $\eta^k_{\mu,n}$; error $e^u_{\mu,n}$ and residual norm $\|R_{\mu,n}\|_2$; error $e^{u^*}_{\mu,n}$ and residual norm $\|R^*_{\mu,n}\|_2$.

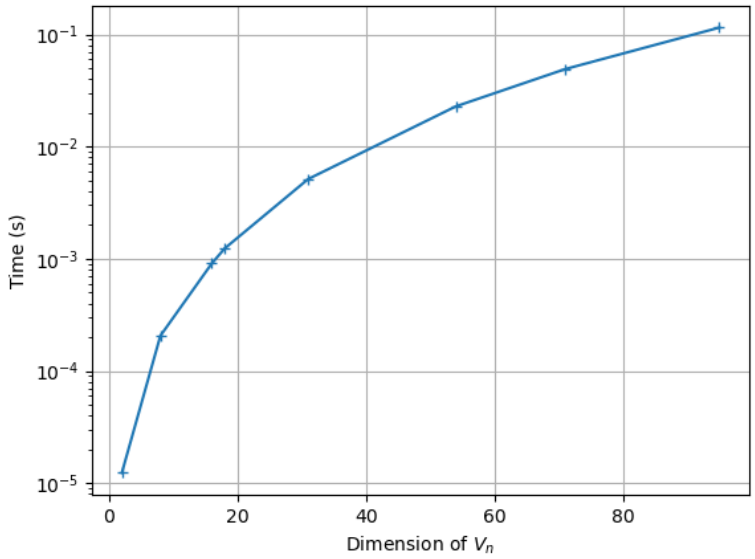

Figure 5.8: Mean computational time for the power method of the reduced solver, over $\mathcal{P}_{\text{test}}$, as a function of the dimension n of the reduced space

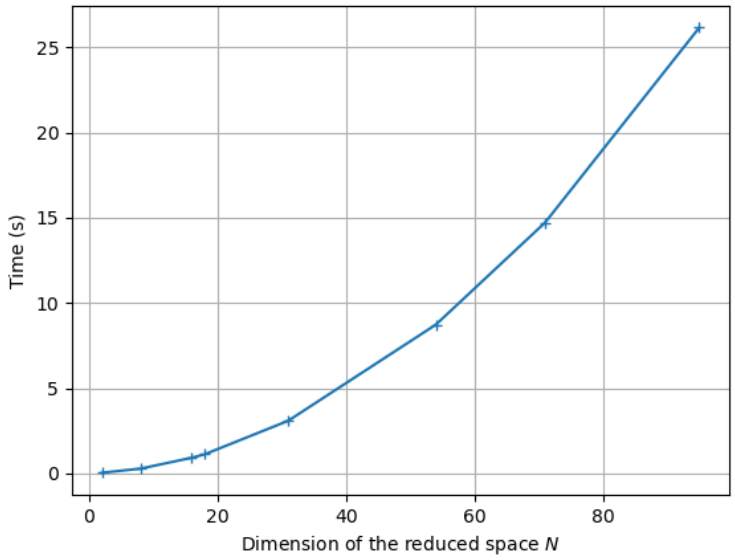

Figure 5.9: Mean computational time for the assembling of the reduced matrices $A_{\mu,n}$ and $B_{\mu,n}$, over $\mathcal{P}_{\text{test}}$, as a function of the dimension n of the reduced space

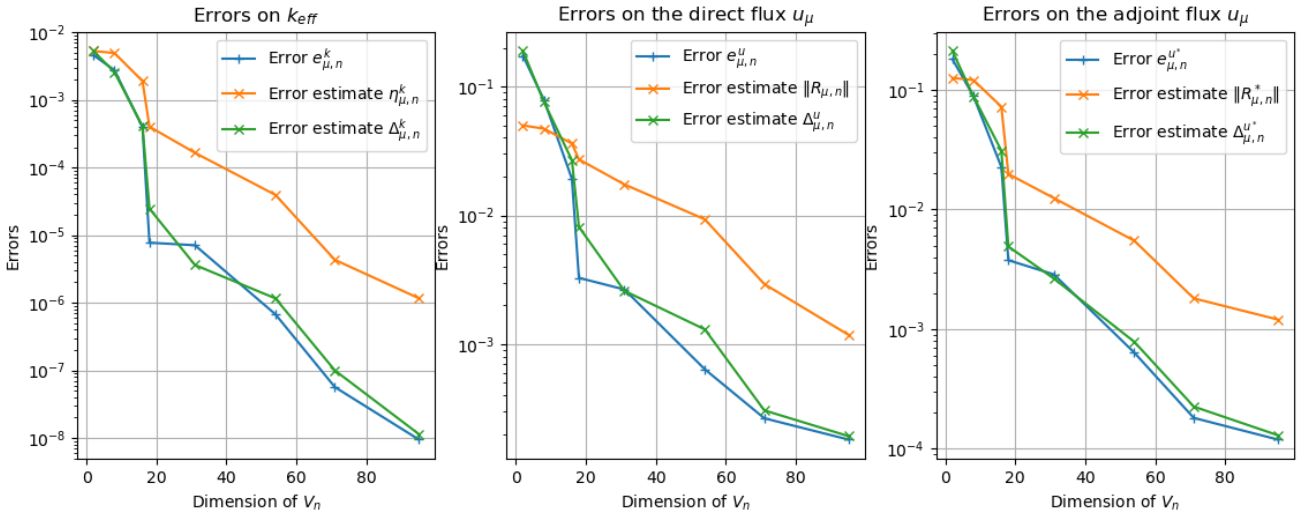

Figure 5.10: Maximum errors and their associated error estimates with respect to the dimension n of the reduced space $\mathcal{V}_n$, over $\mathcal{P}_{\text{test}}$. From left to right: error $e^k_{\mu,n}$, $\eta^k_{\mu,n}$ and $\Delta^k_{\mu,n}$; error $e^u_{\mu,n}$, residual norm $\|R_{\mu,n}\|_2$ and $\Delta^u_{\mu,n}$; error $e^{u^*}_{\mu,n}$, residual norm $\|R^*_{\mu,n}\|_2$ and $\Delta^{u^*}_{\mu,n}$.

Chapter 6

Impact of physical model error on state estimation for neutronics applications

This chapter comes at the end of this manuscript to offer another illustrative example of Model Order Reduction application to neutronics. It suggests a distinct ROM from the one detailed from Chapters 2 to 5, since it is based on experimental measurements, and aims at reconstructing the power map inside a nuclear reactor.

It is a published proceeding from the MOCO project of the CEMRACS 2021 research session dedicated to data assimilation and reduced modeling for high dimensional problems. Its reference in the manuscript is:

[38] Y. Conjungo Taumhas, D. Labeurthre, F. Madiot, O. Mula, and T. Taddei, Impact of physical model error on state estimation for neutronics applications, ESAIM: Proceedings and Surveys, vol. 73 (2023), pp. 158–172.

Abstract In this paper, we consider the inverse problem of state estimation of nuclear power fields in a power plant from a limited number of observations of the neutron flux. For this, we use the Parametrized Background Data Weak approach. The method combines the observations with a parametrized PDE model for the behavior of the neutron flux. Since, in general, even the most sophisticated models cannot perfectly capture reality, an inevitable model error is made. We investigate the impact of the model error in the power reconstruction when we use a diffusion model for the neutron flux, and we assume that the true physics are governed by a neutron transport model.

Introduction

In the field of nuclear engineering, numerical methods play a crucial role at several stages: they are involved in important assessments and decisions related to design, safety, energy efficiency, and reactor loading plans. In this paper, we focus on the task of providing real time information about the spatial distribution of the nuclear power generated by a nuclear reactor from a limited number of measurement observations. We combine this data with physical models in order to provide a complete spatial reconstruction of the power field. This task is a state estimation problem, and we work with the Parametrized Background Data Weak (PBDW), originally introduced in [91]. The method has the appealing feature of providing very fast reconstructions by leveraging techniques from model order reduction of parametric Partial Differential Equations (PDEs). We refer to [17, 45, 33, 34] for theoretical analysis of the method, optimal recovery results and nonlinear extensions. A recent overview may be found in [98].

The main ideas of the above state estimation methodology have been applied to the field of nuclear physics for applications connected to neutronics (see [7, 6, 5]). We could also cite other works such as [108, 63, 31, 57, 87] which study the forward reduced modeling problem for neutronics (compared to these works, note that there is a salient difference in the nature of the task that we consider, which is inverse state estimation). In this paper, we again consider neutronics but our goal is to study the impact of inaccuracies in the physical model that is involved in the reconstruction algorithm, and which is often assumed to perfectly describe reality. This assumption goes beyond the present application on neutronics but studying it for this particular topic has the advantage that we have two very well identified models with different levels of accuracy, thereby allowing to examine synthetically what one can expect when working in a real application scenario.

In neutronics, the most accurate physical model is the so-called neutron transport equation which describes the evolution of the neutronic population in a reactor core by expressing it in the form of a balance between produced and lost neutrons [71]. This model is often approximated at the reactor core scale by a neutron diffusion model to save computing time. This is why in this work, we explore the impact of model inaccuracies by applying a reconstruction based on a diffusion model for the neutron flux, and then assuming that the true physical system is governed by a neutron transport model.

The paper is organized as follows. Section 6.1 is devoted to presenting inverse state estimation problems and the PBDW method. Section 6.2 details the application of the methodology to the reconstruction of nuclear power. Section 6.3 provides some numerical results.

6.1 Inverse State Estimation with PBDW

In this section, we introduce the problem of state estimation, and the Parametrized Background Data Weak method which combines measurement observations and reduced models from parametric PDEs. We refer the reader to [98] for an overview of inverse problem algorithms using these elements.

Let $\mathcal{R}$ be a fixed given domain of $\mathbb{R}^d$ with dimension $d \geq 1$, and let V be a Hilbert space defined over $\mathcal{R}$. In our application, $\mathcal{R}$ will be defined as the nuclear reactor domain. The space V is endowed with an inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. The choice of V must be relevant for the problem under consideration: typical options are L^2, H^1; for pointwise measurements, a Reproducing Kernel Hilbert Space should be considered.

Our goal is to recover an unknown function $u \in V$ from m measurement observations

$$y_i = \ell_i(u), \quad i = 1, \dots, m, \tag{6.1}$$

where the ℓ_i are linearly independent linear forms from the dual V'. Note that we have assumed that experimental observations are perfect; however, the methodology could be extended to deal with noisy measurements (see, e.g., [111, 64, 52]). In practical applications, each ℓ_i models a sensor device which is used to collect the measurement data $\ell_i(u)$. In the applications that we present in our numerical tests, the observations come from sensors for the neutron flux which are placed in the reactor.

We denote by $\omega_i \in V$ the Riesz representers of the ℓ_i. They are defined via the variational equation

$$\langle \omega_i, v \rangle = \ell_i(v), \quad \forall v \in V.$$

Since the ℓ_i are linearly independent in V', so are the ω_i in V and they span an m-dimensional space

$$W_m = \text{span}\{\omega_1, \dots, \omega_m\} \subset V.$$

When there is no measurement noise, knowing the observations $y_i = \ell_i(u)$ is equivalent to knowing the orthogonal projection

$$\omega = P_{W_m} u. \tag{6.2}$$

In this setting, the task of recovering u from the measurement observation ω can be viewed as building a recovery algorithm

$$A : W_m \mapsto V$$

such that $A(P_{W_m} u)$ is a good approximation of u in the sense that $\|u - A(P_{W_m} u)\|$ is small.

Recovering u from the measurements $P_{W_m} u$ is a very ill-posed problem since V is generally a space of very high or infinite dimension so, in general, there are infinitely many $v \in V$ such that $P_{W_m} v = \omega$. It is thus necessary to add some a priori information on u in order to recover the state up to a guaranteed accuracy. In the following, we work in the setting where u is a solution to some parameter-dependent PDE of the general form

$$\mathcal{P}(u, \mu) = 0,$$

where $\mathcal{P}$ is a differential operator and μ is a vector of parameters that describe some physical property and belong to a given set $\mathcal{D} \subset \mathbb{R}^p$. For every $\mu \in \mathcal{D}$, we assume that the PDE has a unique solution $u = u(\mu) \in V$. Therefore, our prior on u is that it belongs to the so-called *solution manifold*

$$\mathcal{M} \coloneqq \{u(\mu) \in V \ : \ \mu \in \mathcal{D}\}. \tag{6.3}$$

In practical applications, the PDE model $\mathcal{P}$ might not be known exactly or might be too expensive to evaluate: we should thus rely on a surrogate approximate model to perform state estimation.

Performance Benchmarks: The quality of a recovery mapping A is quantified in two ways:

- If the sole prior information is that u belongs to the manifold $\mathcal{M}$, the performance is usually measured by the worst case reconstruction error

$$E_{\mathrm{wc}}(A, \mathcal{M}) = \sup_{u \in \mathcal{M}} \|u - A(P_{W_m} u)\| . \tag{6.4}$$

- In some cases u is described by a probability distribution p on V supported on $\mathcal{M}$. This distribution is itself induced by a probability distribution on $\mathcal{D}$ that is assumed to be known. When no information about the distribution is available, usually the uniform distribution is taken. In this Bayesian-type setting, the performance is usually measured in an average sense through the mean-square error

$$E^2_{\mathrm{ms}}(A, \mathcal{M}) = \mathbb{E}\left(\|u - A(P_{W_m} u)\|^2\right) = \int_V \|u - A(P_{W_m} u)\|^2 dp(u) , \tag{6.5}$$

  and it naturally follows that $E_{\mathrm{ms}}(A, \mathcal{M}) \leq E_{\mathrm{wc}}(A, \mathcal{M})$.

PBDW algorithm: In this work, we resort to the Parametrized-Background Data-Weak algorithm (PBDW, [91]) to estimate the state u. Other choices would of course be possible but the PBDW algorithm is relevant for the following reasons:

- **Simplicity and Speed:** It is easily implementable and it provides reconstructions in near-real time.
- **Optimality:** It has strong connections with optimal linear reconstruction algorithms as has been studied in [17, 33].
- **Extensions:** If required, the algorithm can easily be extended to enhance its reconstruction performance (see [34, 55]). In particular, it is shown in [34] that piecewise PBDW reconstruction strategy can deliver near-optimal performance. The PBDW algorithm can also be easily adapted to accommodate noisy measurements (see [111, 64]) and some easy-to implement extension to mitigate the model error exist (in the following however, we assume the PDE model is perfect for the sake of simplicity).

Since the geometry of $\mathcal{M}$ is generally complex, optimization tasks posed on $\mathcal{M}$ are difficult (lack of convexity, high evaluation costs for different parameters). Therefore, instead of working with $\mathcal{M}$, PBDW works with a linear (or affine) space V_n of reduced dimension n which is expected to approximate the solution manifold well in the sense that the approximation error of the manifold

$$\delta_n^{(\mathrm{wc})} \coloneqq \sup_{u \in \mathcal{M}} \mathrm{dist}(u, V_n) , \quad \text{or} \quad \delta_n^{(\mathrm{ms})} \coloneqq \mathbb{E}\left(\mathrm{dist}(u, V_n)^2\right)^{1/2} \tag{6.6}$$

decays rapidly if we increase the dimension n. It has been proven in [35] that it is possible to find such hierarchies of spaces $(V_n)_{n \geq 1}$ for certain manifolds coming from classes of elliptic and parabolic problems, and numerous strategies have been proposed to build

the spaces in practice (see, e.g., [25, 105] for reduced basis techniques and [35, 36] for polynomial approximations in the μ variable).

Assuming that we are given a reduced model V_n with $1 \leq n \leq m$, the PBDW algorithm

$$A_{m,n}^{(\mathrm{pbdw})} : W_m \to V$$

gives for any $\omega \in W_m$ a solution of

$$A_{m,n}^{(\mathrm{pbdw})}(\omega) \in \underset{u \in \omega + W(\mathcal{R})^\perp}{\arg\min} \operatorname{dist}(u, V_n). \tag{6.7}$$

The minimizer is unique as soon as $n \leq m$ and $\beta(V_n, W_m) > 0$, which is an assumption to which we adhere in the following. The quantity β is defined as follows. For any pair of closed subspaces (E, F) of V, $\beta(E, F)$ is defined as

$$\beta(E,F) := \inf_{e \in E} \sup_{f \in F} \frac{\langle e, f \rangle}{\|e\| \, \|f\|} = \inf_{e \in E} \frac{\|P_F e\|}{\|e\|} \in [0,1]. \tag{6.8}$$

We can prove that $A_{m,n}^{(\mathrm{pbdw})}$ is a bounded linear map from W_m to $V_n \oplus (W_m \cap V_n^\perp)$.

In practice, solving problem (6.7) boils down to solving a linear least squares minimization problem whose cost is essentially of order n^2+m, and we can compute $\beta(V_n, W_m)$ by finding the smallest eigenvalue of an $n \times n$ matrix. We refer, e.g., to [98, Appendix A, B] for details on how to compute these elements in practice. It follows that, since in general m is not very large, if the dimension n of the reduced model is moderate, the reconstruction with (6.7) can take place in close to real-time.

For any $u \in V$, the reconstruction error is bounded by

$$\|u - A_{m,n}^{(\mathrm{pbdw})}(\omega)\| \leq \beta^{-1}(V_n, W_m) \|u - P_{V_n \oplus (W_m \cap V_n^\perp)} u\| \leq \beta^{-1}(V_n, W_m) \|u - P_{V_n} u\|, \tag{6.9}$$

where we have omitted the dependency of the spaces on $\mathcal{R}$ in order not to overload the notation, and we will keep omitting this dependency until the end of this section. Depending on whether V_n is built to address the worst case or mean square error, the reconstruction performance over the whole manifold $\mathcal{M}$ is bounded by

$$e_{m,n}^{(\mathrm{wc,\ pbdw})} := E_{\mathrm{wc}}(A_{m,n}^{(\mathrm{pbdw})}, \mathcal{M}) \leq \beta^{-1}(V_n, W_m) \max_{u \in \mathcal{M}} \operatorname{dist}(u, V_n \oplus (V_n^\perp \cap W_m)) \leq \beta^{-1}(V_n, W_m) \, \delta_n^{(\mathrm{wc})}, \tag{6.10}$$

or

$$\begin{aligned} e_{m,n}^{(\mathrm{ms,\ pbdw})} := E_{\mathrm{ms}}(A_{m,n}^{(\mathrm{pbdw})}, \mathcal{M}) &\leq \beta^{-1}(V_n, W_m) \mathbb{E}\left(\operatorname{dist}(u, V_n \oplus (V_n^\perp \cap W_m))^2\right)^{1/2} \\ &\leq \beta^{-1}(V_n, W_m) \, \delta_n^{(\mathrm{ms})}. \end{aligned} \tag{6.11}$$

Note that $\beta(V_n, W_m)$ can thus be understood as a stability constant. It can also be interpreted as the cosine of the angle between V_n and W_m. The error bounds involve the distance of u to the space $V_n \oplus (V_n^\perp \cap W_m)$ which provides slightly more accuracy than the reduced model V_n alone. This term is the reason why it is sometimes said that the method can correct model error to some extent. In the following, to ease the reading we will write errors only with the second type of bounds (6.11) that do not involve the correction part on $V_n^\perp \cap W_m$.

An important observation is that for a fixed measurement space W_m (which is the setting in our numerical tests), the error functions

$$n \mapsto e_{m,n}^{(\mathrm{wc,\ pbdw})}, \quad \text{and} \quad n \mapsto e_{m,n}^{(\mathrm{ms,\ pbdw})}$$

reach a minimal value for a certain dimension n_{wc}^* and n_{ms}^* as the dimension n varies from 1 to m. This behavior is due to the trade-off between:

- the improvement of the approximation properties of V_n as n grows ($\delta_n^{(\mathrm{wc})}$ and $\delta_n^{(\mathrm{ms})} \to 0$ as n grows)
- the degradation of the stability of the algorithm, given here by the decrease of $\beta(V_n, W_m)$ to 0 as $n \to m$. When $n > m$, $\beta(V_n, W_m) = 0$.

As a result, the best reconstruction performance with PBDW is given by

$$e_{m,n^*_{\mathrm{wc}}}^{(\mathrm{wc,\ pbdw})} = \min_{1\le n\le m} e_{m,n}^{(\mathrm{wc,\ pbdw})}, \quad \text{or} \quad e_{m,n^*_{\mathrm{ms}}}^{(\mathrm{ms,\ pbdw})} = \min_{1\le n\le m} e_{m,n}^{(\mathrm{ms,\ pbdw})}.$$

Noise and Model Error: To account for measurement noise and model bias in the above analysis, let us assume that we get noisy observations $\tilde{\omega} = \omega + \eta$ with $||\eta|| \le \varepsilon_{noise}$. Suppose also that the true state u does not lie in $\mathcal{M}$ but satisfies $\mathrm{dist}(u, \mathcal{M}) \le \varepsilon_{model}$. We can prove that the error bound (6.9) should be modified into

$$\|u - A_{m,n}^{(\mathrm{pbdw})}(\tilde{\omega})\| \le \beta^{-1}(V_n, W_m)(\|u - P_{V_n} u\| + \varepsilon_{noise} + \varepsilon_{model}).$$

Thus (6.10) and (6.11) become

$$e_{m,n}^{(\mathrm{wc,\ pbdw})} := E_{\mathrm{wc}}(A_{m,n}^{(\mathrm{pbdw})}, \mathcal{M}) \le \beta^{-1}(V_n, W_m)\,(\delta_n^{(\mathrm{wc})} + \varepsilon_{noise} + \varepsilon_{model}), \tag{6.12}$$

and

$$e_{m,n}^{(\mathrm{ms,\ pbdw})} := E_{\mathrm{ms}}(A_{m,n}^{(\mathrm{pbdw})}, \mathcal{M}) \le \beta^{-1}(V_n, W_m)\,(\delta_n^{(\mathrm{ms})} + \varepsilon_{noise} + \varepsilon_{model}). \tag{6.13}$$

Note that the estimation accuracy benefits from decreasing the model error, and the noise. Since both errors have the same additive effect on the reconstruction accuracy, model error could be understood as measurement error and vice-versa. However, since the underlying physical reasons leading to model and measurement error are entirely different, it is preferable to clearly keep both concepts separately. Note further that the computational complexity of the method is not affected by these errors. This is in contrast to Bayesian methods for which small noise levels induce computational difficulties due to the concentration of the posterior distribution.

Sensor modeling error: Another error that can occur comes from our choice of the observation functions ω_i which are built to mimic the response of the sensor devices. Suppose that we work with imperfect functions $\tilde{\omega}_i$ that deviate from the exact one ω_i with $\|\omega_i - \tilde{\omega}_i\| \le \rho$ for some $\rho > 0$. Then noiseless observations can be written as

$$y_i = \ell_i(u) = \langle \omega_i, u\rangle = \langle \tilde{\omega}_i, u\rangle + \langle \omega_i - \tilde{\omega}_i, u\rangle\,.$$

The right hand side tells us that by working with the inexact $\tilde{\omega}_i$, we are introducing a term of noise which is $\langle \omega_i - \tilde{\omega}_i, u\rangle$. The noise has level $\rho\|u\|$. It follows that working with an inexact representation of the sensor response can be understood as introducing additional noise to the observations.

6.2 Application to the reconstruction of nuclear power

In this work, we apply the above general framework to reconstruct the nuclear power P generated in a nuclear reactor core defined on a convex domain $\mathcal{R}$. The power P is a

real-valued function in $\mathcal{R}$, $P : \mathcal{R} \to \mathbb{R}_+$, and in the following we reconstruct it by viewing it as a function in the space

$$V = L^2(\mathcal{R}).$$

The nuclear power P we want to rebuild always comes from the neutron transport model. However, the spaces used to reconstruct P will be divided into two cases. One space is made up of solutions of the transport model while the other is made up of solutions of the diffusion model as discussed in the following sections.

6.2.1 The neutron transport model

We assume that the reactor is in a stationary state where the neutron population ψ, usually called the angular flux, depends on (r, ω, E), namely the spatial position $r \in \mathcal{R} \subset \mathbb{R}^d$, the direction of propagation $\omega \in \mathbb{S}_d$ where $\mathbb{S}_d$ is the unit sphere of $\mathbb{R}^d$, and the kinetic energy $E \in \mathbb{R}^+$. We work with a multi-group approach where we consider a discrete set of energies $E_G < \cdots < E_0$, and we denote

$$\psi(r, \omega, [E_g, E_{g-1}]) \coloneqq \psi^g(r, \omega), \quad \forall (r, \omega) \in \mathcal{R} \times \mathbb{S}_d, \quad \forall g \in \{1, \dots, G\}.$$

With this notation, the neutron transport equation is a generalized eigenvalue problem in which we search for a multigroup flux $\psi = (\psi^g)_{g=1}^G$, and a generalized eigenvalue $\lambda \in \mathbb{C}^*$ (see [43])

$$\begin{cases} L^g\psi^g(r,\omega) &= H^g\psi(r,\omega) + \lambda F^g\psi(r,\omega) \quad \text{in } \mathcal{R} \times \mathbb{S}_2, \quad \forall g \in \{1, \dots, G\} \\ \psi(r,\omega) &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{on } \partial\Gamma_- \coloneqq \{(r,\omega) \in \partial\mathcal{R} \times \mathbb{S}_d \ : \ n(r) \cdot \omega < 0\}, \end{cases} \tag{6.14}$$

where

$$L^g\psi^g(r,\omega) \coloneqq (\omega \cdot \nabla + \Sigma_t^g(r))\psi^g(r,\omega) \text{ is the advection operator,}$$

$$H^g\psi(r,\omega) \coloneqq \sum_{g'=1}^{G} \int_{\mathbb{S}_2} \Sigma_s^{g' \to g}(r, \omega' \cdot \omega)\psi^{g'}(r,\omega')d\omega' \text{ is the scattering operator,}$$

$$F^g\psi(r,\omega) \coloneqq \frac{\chi^g(r)}{4\pi} \sum_{g'=1}^{G} (\nu\Sigma_f)^{g'}(r) \int_{\mathbb{S}_2} \psi^{g'}(r,\omega)d\omega \text{ is the fission operator.}$$

In the listed terms, $\Sigma_t^g(r)$ denotes the total cross-section and $\Sigma_s^{g' \to g}(r, \omega' \cdot \omega)$ is the scattering cross-section from energy group g' and direction ω' to energy group g and direction ω, $\Sigma_f^g(r)$ is the fission cross-section, $\nu^g(r)$ is the average number of neutrons emitted per fission and $\chi^g(r)$ is the fission spectrum. We suppose that all the coefficients are measurable bounded functions of their arguments.

Under certain conditions (which we assume to be satisfied in the following), the eigenvalue $\lambda_{\min}$ with the smallest modulus is simple, real and strictly positive. We refer to [2, Theorem 2.2] for the sketch of the proof detailed in [12, Theorem 2.1.1, p 92]. The associated eigenfunction ψ belongs to the Hilbert space $W^2(\mathcal{R})^G$ where $W^2(\mathcal{R} \times \mathbb{S}_2) = \{\psi \in L^2(\mathcal{R} \times \mathbb{S}_2) \text{ s.t. } \omega \cdot \nabla\psi \in L^2(\mathcal{R} \times \mathbb{S}_2)\}$, is also real and positive at almost every $(x, \omega) \in \mathcal{R} \times \mathbb{S}_2$. With this model, once the neutron flux is computed by solving (6.14) numerically, the nuclear power is given by

$$P(r) \coloneqq \sum_{g'=1}^{G} (\kappa\Sigma_f)^{g'} \int_{\mathbb{S}_2} \psi^{g'}(r,\omega)\,\mathrm{d}\omega, \quad \forall r \in \mathcal{R} \text{ a.e,}$$

where $\kappa^g \in L^\infty(\mathcal{R})$ is the released energy per fission and since $\psi \in (W^2(\mathcal{R} \times \mathbb{S}_2))^G$, we have that $P \in V$.

6.2.2 The neutron diffusion equations

In this work, the neutron flux ϕ is modeled with the two-group neutron diffusion equation with null flux boundary conditions. So ϕ has two energy groups $\phi = (\phi^1, \phi^2)$. Index 1 denotes the high energy group and 2 the thermal energy one. The flux is the solution to the following eigenvalue problem (see [71])

$$\text{Find } (\lambda, \phi) \in \mathbb{C} \times (H^1(\mathcal{R}) \times H^1(\mathcal{R})) \text{ such that,}$$

$$\begin{cases} -\nabla\left(D^1 \nabla \phi^1\right) + \Sigma_a^1 \phi^1 - \Sigma_{s,0}^{2\to 1} \phi^2 = \lambda\left(\chi^1 (\nu\Sigma_f)^1 \phi^1 + \chi^1 (\nu\Sigma_f)^2 \phi^2\right), & \text{in } \mathcal{R}, \\ -\nabla\left(D^2 \nabla \phi^2\right) + \Sigma_a^2 \phi^2 - \Sigma_{s,0}^{1\to 2} \phi^1 = \lambda\left(\chi^2 (\nu\Sigma_f)^1 \phi^1 + \chi^2 (\nu\Sigma_f)^2 \phi^2\right), & \text{in } \mathcal{R}, \end{cases} \tag{6.15}$$

with

$$D^g \nabla \phi^g \cdot n(r) + \frac{1}{2}\phi^g = 0 \quad \text{ on } \partial\mathcal{R}, \text{ for } g = 1, 2.$$

The coefficients involved are the following:

- D^g is the diffusion coefficient of group g with $g \in \{1, 2\}$.
- Σ_a^g is the macroscopic absorption cross section of group g.
- $\Sigma_{s,0}^{g'\to g}$ is the macroscopic scattering cross section of anisotropy order 0 from group g' to g.
- χ^g is the fission spectrum of group g.

We assume that they are either constant of piecewise constant in $\mathcal{R}$ so we can view them as functions from $L^\infty(\mathcal{R})$.

The generated power is

$$P \coloneqq (\kappa\Sigma_f)^1 \phi^1 + (\kappa\Sigma_f)^2 \phi^2, \tag{6.16}$$

and since ϕ^1 and $\phi^2 \in H^1(\mathcal{R})$, we have $P \in V$.

We next make some comments on the coefficients and recall well-posedness results of the eigenvalue problem (6.15). First of all, the first five coefficients (D^g, Σ_a^g, $\Sigma_{s,0}^{1\to 2}$, $\Sigma_{s,0}^{2\to 1}$ and $(\nu\Sigma_f)^g$) might depend on the spatial variable. In the following, we assume that they are either constant or piecewise constant so that our set of parameters is

$$\mu = \{D^1, D^2, \Sigma_a^1, \Sigma_a^2, \Sigma_{s,0}^{1\to 2}, (\nu\Sigma_f)^1, (\nu\Sigma_f)^2, \chi^1, \chi^2\}. \tag{6.17}$$

By abuse of notation, in (6.17) we have written D^g to denote the set of values that this coefficient might take in space and similarly for the other parameters.

Under some mild conditions on the parameters μ, the eigenvalue $\lambda_{\min}$ with the smallest modulus is simple, real and strictly positive (see [43, Chapter XXI]). The associated eigenfunction ϕ is also real and positive at almost every point $x \in \mathcal{R}$ and it is what is classically called the flux. In neutronics, it is customary to work with the inverse of $\lambda_{\min}$, which is called the multiplication factor

$$k_{\text{eff}} \coloneqq 1/\lambda_{\min}. \tag{6.18}$$

Therefore k_{eff} is not a parameter in our setting because, for each value of the parameters μ, k_{eff} is determined by the solution to the eigenvalue problem.

If the parameters of our diffusion model range in, say,

$$D^1 \in [D^1_{\min}, D^1_{\max}],\ D^2 \in [D^2_{\min}, D^2_{\max}], \ldots, \chi^2 \in [\chi^2_{\min}, \chi^2_{\max}],$$

then

$$\mathcal{D} \coloneqq [D^1_{\min}, D^1_{\max}] \times \cdots \times [\chi^2_{\min}, \chi^2_{\max}], \tag{6.19}$$

and the set of all possible states of the power is given by

$$\mathcal{M}_{\text{diff}} = \{P(\mu)\ :\ \mu \in \mathcal{D}\}\ \subset V, \tag{6.20}$$

which is the manifold of solutions of our problem.

6.3 Numerical Examples

6.3.1 Description of the test case and the numerical solver

The test-case: We consider Model 1 Case 1 of the well-known Takeda neutronics benchmark [112] to build our test case. The geometry of the core is three-dimensional and the domain is $\{(x, y, z) \in \mathbb{R}^3, 0 \leq x \leq 25\,\text{cm}; 0 \leq y \leq 25\,\text{cm}; 0 \leq z \leq 25\,\text{cm}\}$. This test is defined with $G = 2$ energy groups and isotropic scattering and we set $\kappa^g = 1$ MeV for $g = 1, 2$. The reactor core geometry is depicted in Figure 6.1. In the following, we implicitly refer to the cross-sections and the other coefficients of this test case. Our goal

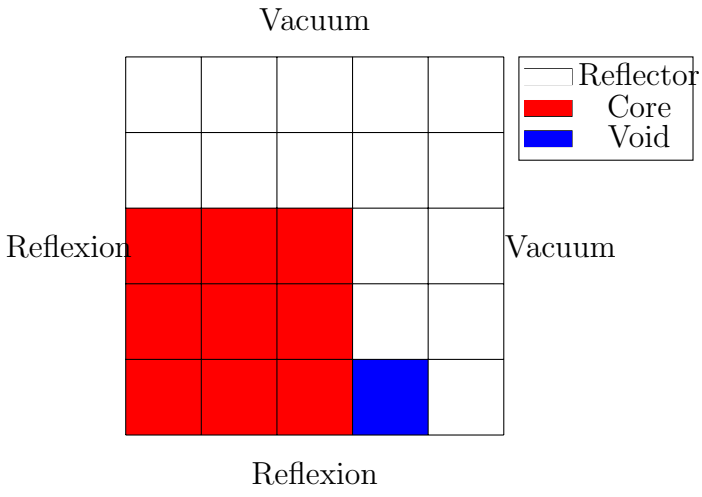

Figure 6.1: Cross-sectional view of the core ($z = 0$ cm).

is to reconstruct in real time the spatial power field of the reactor. We assume that the neutron transport equation perfectly describes reality, and the set of all possible states is given by the manifold

$$\mathcal{M}_{\text{tr}} = \{P^{\text{tr}}(\mu)\ :\ \mu \in \mathcal{D}\}\ \subset V.$$

The set of solutions of the neutron diffusion equation is

$$\mathcal{M}_{\text{diff}} = \{P^{\text{diff}}(\mu)\ :\ \mu \in \mathcal{D}\}\ \subset V.$$

It is an imperfect description of the true states given by $\mathcal{M}_{\text{tr}}$.

The parameter set μ from equation (6.17) is generated by the mapping

$$\mu : [0.8, 1]^5 \subset \mathbb{R}^5 \to \mathbb{R}^9$$
$$\alpha \mapsto \mu(\alpha) \left(\frac{D^1}{\alpha_1}, \frac{D^2}{\alpha_2}, \alpha_1 \Sigma_a^1, \alpha_2 \Sigma_a^2, \alpha_3 \Sigma_{s,0}^{1\to 2}, \alpha_4 (\nu\Sigma_f)^1, \alpha_5 (\nu\Sigma_f)^2, \chi^1, \chi^2 \right).$$

We can thus view the parameter set either as the 5 dimensional tensorized subset $[0.8, 1]^5$ where α ranges, or as a 5-dimensional surface manifold from $\mathbb{R}^9$ where the 9 coefficients μ of the neutronic model live.

We work with $m = 54$ measurements observations that are placed uniformly in the reactor. They are defined as local averages over small subdomains $\mathcal{R}_i \subset \mathcal{R}$

$$\omega_i(x) = \frac{1}{|\mathcal{R}_i|} \mathbb{1}_{\mathcal{R}_i}(x), \quad \forall x \in \mathcal{R},\, i = 1, \dots, m. \tag{6.21}$$

We compare two cases:

1. **Perfect physical model:** We apply PBDW using reduced models from the transport manifold which represents the true reality in our experiments.

2. **Imperfect physical model:** We assume that a perfect model is out of reach and we use the diffusion manifold. The reconstruction will thus be affected by a model bias.

The solver: To generate the snapshots and the reduced models, we have worked with MINARET [79], a deterministic solver for reactor physics calculations developed in the framework of the APOLLO3® code [109] (see Section 5.1). For our simulations, we work with a level-symmetric formula of order $N = 8$ for the S_N quadrature, and the spatial approximation uses discontinuous $\mathbb{P}_1$ finite elements of a uniform mesh. The physical output power map is post-processed on an approximation space of dimension $N_h = 540$ (N_h degrees of freedom).

6.3.2 Case 1: Reconstruction with a perfect physical model

Here we assume that we have access to a perfect description of the physics, and we work with the neutron transport manifold $\mathcal{M}_{\text{tr}}$.

In order to create a reduced space V_n of small dimension $n \ll N_h$, we apply a Proper Orthogonal Decomposition (POD) based on the training set

$$\mathcal{P}_{\text{training}} = \{P^{\text{tr}}(\mu(\alpha)), \alpha \in \{0.8, 0.9, 1\}^5\} \subset \mathcal{M}_{\text{tr}}$$

of power maps obtained from solutions of the transport neutron equations, also called snapshots.

We measure the relative approximation error $\widetilde{\delta}_n^{(\text{wc})}$ as defined in Equation (6.6). For this, we define a collection of power maps of reference

$$\mathcal{P}_{\text{test}} = \{P^{\text{tr}}(\mu(\alpha))\ ,\ \alpha \in \{0.85, 0.95\}^5\}. \tag{6.22}$$

Figure 6.2 shows that the training space is well approximated with a few POD modes. For $n \geq 30$, the relative error between one power map and its projection onto V_n is smaller than 10^{-6}.

We next study the ability to reconstruct the power field with measurement observations, and the PBDW method, as Figure 6.3 shows in the 3D space. For this, we compute for $1 \leq n \leq m$:

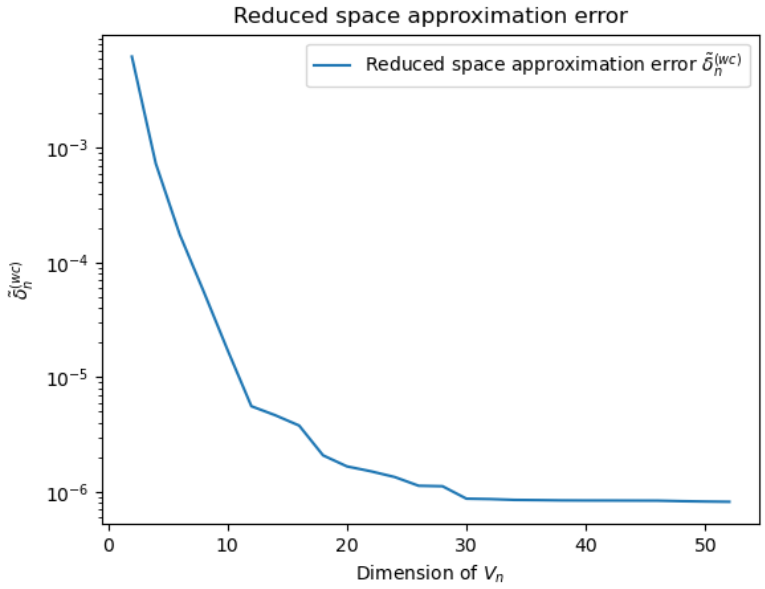

Figure 6.2: Relative approximation error $\widetilde{\delta}_n^{(\mathrm{wc})}$ of the transport manifold $\mathcal{M}_{\mathrm{tr}}$ with respect to the dimension n of the reduced space. Here the reduced space is a POD computed using the same manifold $\mathcal{M}_{\mathrm{tr}}$.

- The relative reconstruction error given by $\widetilde{e}_{m,n}^{(\mathrm{wc,\ pbdw})} = \max_{u \in \widetilde{\mathcal{M}}_{\mathrm{tr}}} \dfrac{\|u - A_{m,n}^{(\mathrm{pbdw})}(\omega)\|}{\|u\|}$,
- The upper bound of the reconstruction error given by $\beta^{-1}(V_n, W_m)\widetilde{\delta}_n^{(\mathrm{wc})}$, as given in Equation (6.11).

Figure 6.4 shows that the upper bound is about two orders of magnitude above the actual reconstruction error. This gap is expected to decrease if we use more functions in the test set. The second observation is that the reconstruction accuracy reaches a minimum for a dimension $n^* \approx 25$. If we work with the optimal dimension n^*, an important result is that we can recover the power field from measurement observations at almost the same accuracy ($\approx 10^{-6}$, see Figure 6.2) as the one given by the orthogonal projection onto V_n (to see this, compare the errors at n^* in Figures 6.2 and 6.4).

The behavior of the reconstruction error with the dimension n is connected to a loss of stability illustrated in Figure 6.5. It warns about a compromise to find between the approximation error of the manifold and the stability in order to optimize the accuracy of the power map reconstruction. One strategy to mitigate stability problems is to find locations for the sensor measurements that span spaces W_m maximizing the value of $\beta(V_n, W_m)$ (see, e.g., [18]).

6.3.3 Case 2: Reconstruction of the power map from diffusion snapshots

We now consider the diffusion neutron equations as the best available model while the true states are given by the neutron transport model. They are therefore members of $\mathcal{M}_{\mathrm{tr}}$.

Similarly as done in Section 6.3.2, we apply a POD over a collection

$$\mathcal{P}_{\mathrm{training}} = \{P^{\mathrm{diff}}(\mu(\alpha)) \ : \ \alpha \in \{0.8, 0.9, 1\}^5\} \subset \mathcal{M}_{\mathrm{diff}},$$

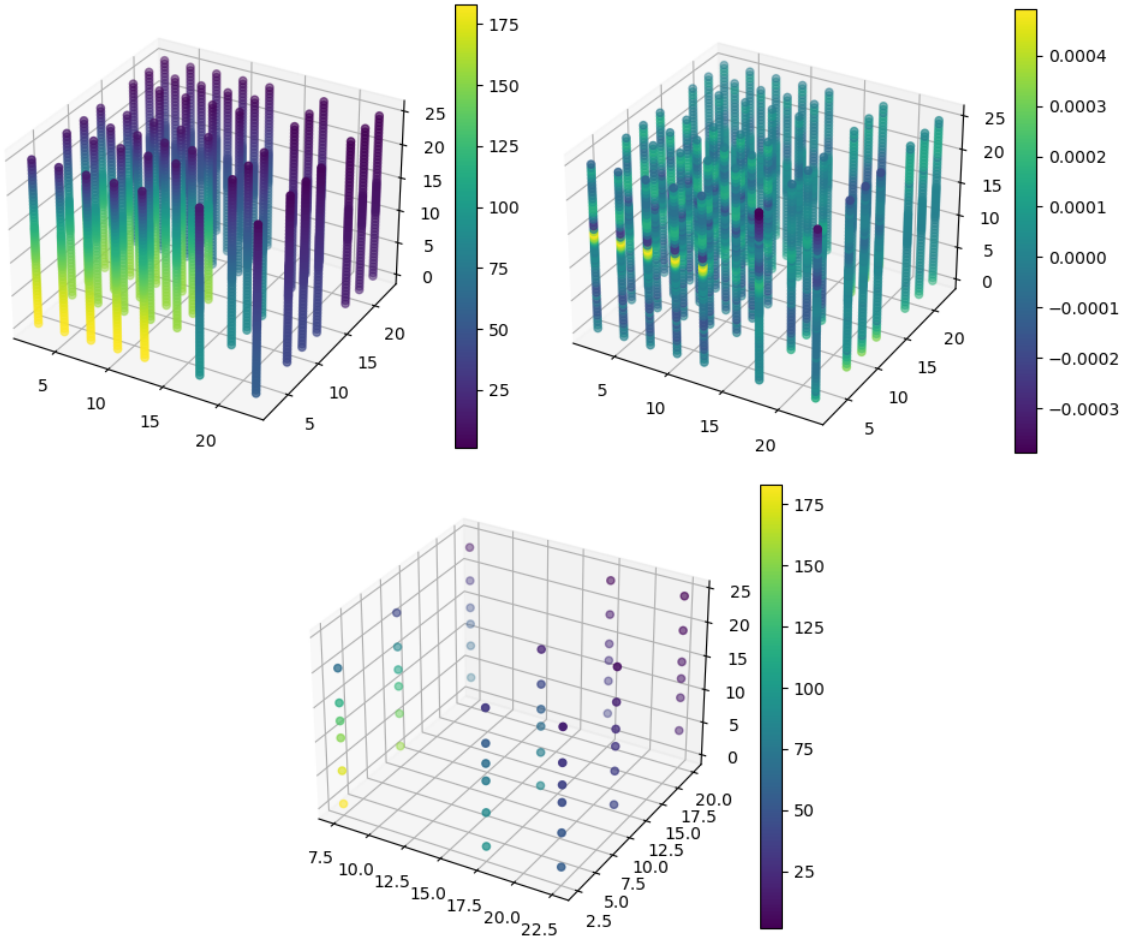

Figure 6.3: 3D representation of the power map $P^{\mathrm{tr}}(\mu(\alpha))$ with $\alpha = \{0.85\}^5$ (upper left), the algebraic reconstruction error by PBDW (upper right) and the $m = 54$ measurements

to create a reduced space V_n of dimension $n \ll N_h$. The main difference lies in the fact that the snapshots are obtained from the neutron diffusion equations, as we consider that the transport model cannot be computed in this section.

Figure 6.6 shows that the approximation error of the transport manifold $\mathcal{M}_{\mathrm{tr}}$ is less accurate than in the previous case due to the bias between the two models. Typically, for $n = 50$, we approximate the manifold at the accuracy of 6×10^{-3}, whereas the approximation with the transport model was about 10^3 times better (compare Figure 6.6 and Figure 6.4). Therefore, the reconstruction error will have a similar order of magnitude to those observed for the approximation error.

Similarly, we compute for $1 \leq n \leq m$:

- The relative reconstruction error given by $\widetilde{e}_{m,n}^{(\mathrm{wc,\ pbdw})} = \max_{u \in \widetilde{\mathcal{M}}_{\mathrm{tr}}} \dfrac{\|u - A_{m,n}^{(\mathrm{pbdw})}(\omega)\|}{\|u\|}$,
- The upper bound of the reconstruction error given by $\beta^{-1}(V_n, W_m)\widetilde{\delta}_n^{(\mathrm{wc})}$, as given in Equation (6.11).

As done before, the PBDW reconstruction procedure is then performed by extracting measurements over the collection power maps of reference defined in (6.22). Figure 6.7

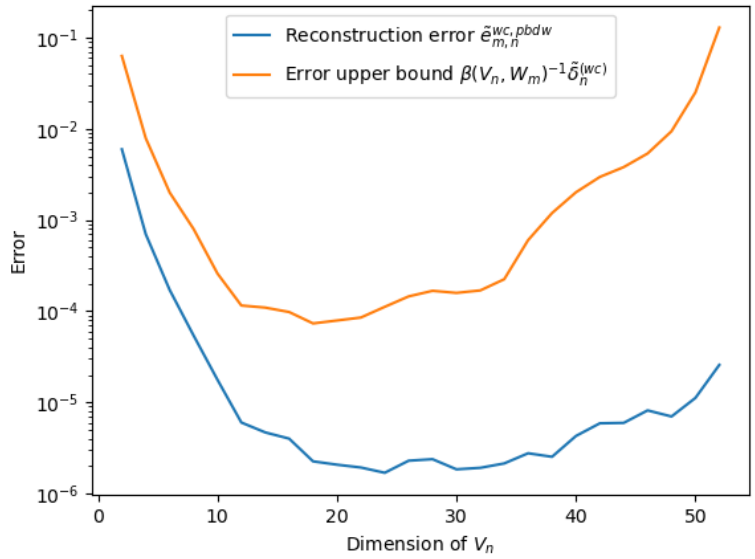

Figure 6.4: Relative reconstruction error $\widetilde{e}_{m,n}^{(\mathrm{wc,\ pbdw})}$ (in blue) and error estimate $\beta(V_n, W_m)^{-1}\widetilde{\delta}_n^{(\mathrm{wc})}$ (in yellow) with respect to the dimension n of the reduced space

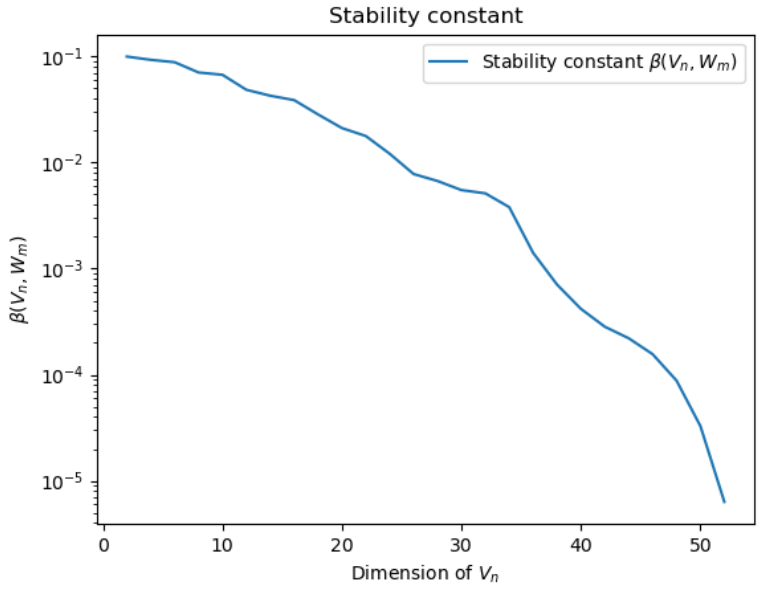

Figure 6.5: Stability constant $\beta(V_n, W_m)$ with respect to the dimension n of the reduced space ($m = 54$)

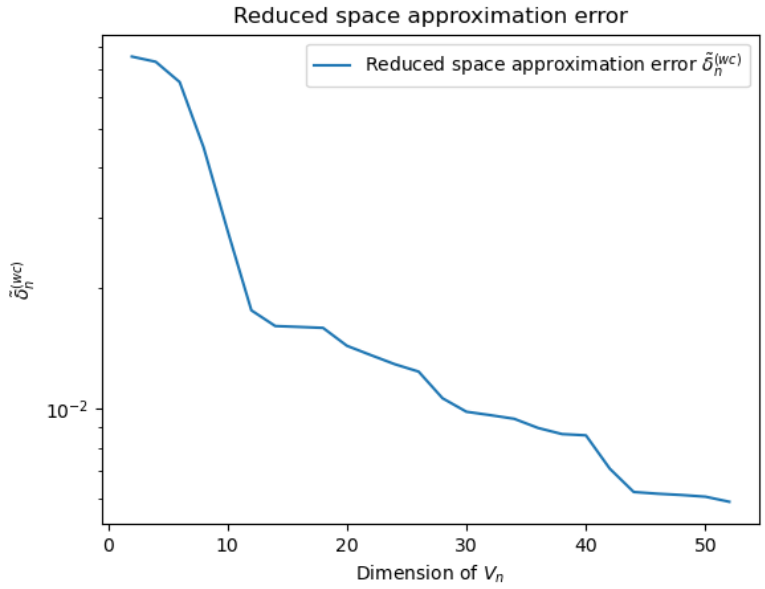

Figure 6.6: Relative approximation error $\widetilde{\delta}_n^{(\mathrm{wc})}$ of the transport manifold $\mathcal{M}_{\mathrm{tr}}$ with respect to the dimension n of the reduced space. Here the reduced space is a POD computed using the diffusion manifold $\mathcal{M}_{\mathrm{diff}}$.

illustrates that the minimum for the reconstruction error reaches about 1.5×10^{-2} for $n^* \approx 35$. The gap between the reconstruction error and its estimate here is bigger as the stability plays a secondary role. Hence, the reconstruction error is only led by the model bias.

Figures 6.8 shows that the stability constant presents the same behavior as in the case of V_n built with transport snapshots.

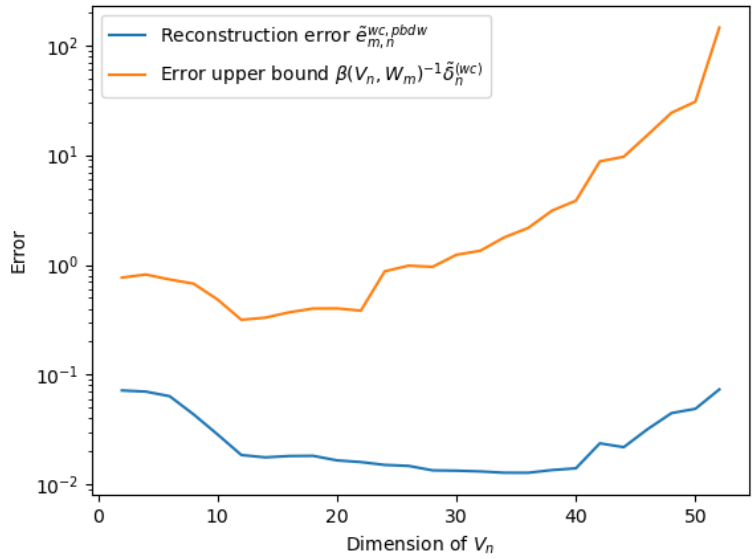

Figure 6.7: Relative reconstruction error $\widetilde{e}^{(\mathrm{wc,\ pbdw})}_{m,n}$ (in blue) and error estimate $\beta(V_n, W_m)^{-1}\widetilde{\delta}^{(\mathrm{wc})}_n$ (in yellow) with respect to the dimension n of the reduced space

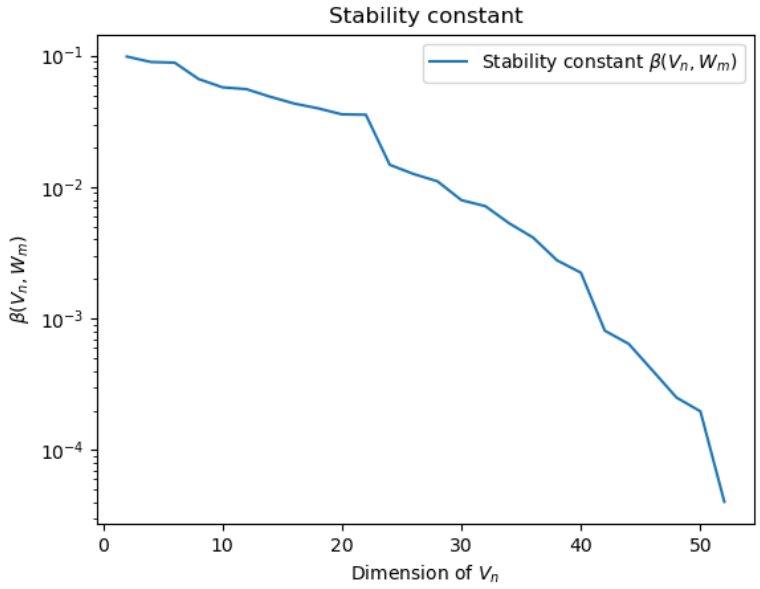

Figure 6.8: Stability constant $\beta(V_n, W_m)$ with respect to the dimension n of the reduced space ($m = 54$)

Acknowledgements

The authors gratefully acknowledge O. Lafitte for fruitful discussions.

***

From the numerical experiments, it follows that the PBDW algorithm can reconstruct data very efficiently when the physical model is perfect. An interesting fact is that there are optimal values n^* for the dimension in the reduced models V_n used in the PBDW algorithm which make the reconstruction with measurement observations comparable with the approximation accuracy by projection on V_n (see, e.g., Figures 6.2 and 6.4).

In presence of model error, the conclusions are analogous. However, the approximation accuracy by direct projection is degraded by the presence of the model error as the comparison between Figure 6.2 and Figure 6.6 shows. This degradation may be reduced if some snapshots are computed with the transport model. The selection of these snapshots may be based on a posteriori estimators devised specifically for the model error.

In principle, the PBDW is expected to be able to correct to some extent the model error due to fact that reconstructions lie in $V_n \oplus (W \cap V_n^\perp)$ and not only in V_n. There, if the model is biased and yields a reduced model V_n which is not perfectly appropriate, the component $(W \cap V_n^\perp)$ is expected to help to correct this inaccuracy. However, our results tend to indicate that this correction component has a very limited effect in our case. This may be due to the poor approximation properties of the observation space W, which is, in our case, spanned by functions that are very localized in space (see equation (6.21) for the definition of the ω_i). This behavior could be improved by working with parametrized families of spaces such as Reproducing Kernel Hilbert spaces (see [94]). In that case, we could try to find an appropriate space for which the ω_i would better enhance the final reconstruction quality. Another option would be to consider purely data-driven corrections on top of the PBDW reconstruction, making use of supervised learning techniques and feed forward neural networks. These ideas will be the starting point of future works in mitigating the effect of model error in state estimation.

Conclusion

Throughout this study, we managed to raise techniques and approaches from the field of Model Order Reduction and *a posteriori* analysis to answer several problems that arise from core calculations in neutronics. We quickly tackled the complexity behind the modeling of neutron dynamics inside a nuclear core, as it exists many discretization techniques for associated mathematical problems; different physical models, that involve model bias, are even considered in order to compute core calculations, as it notably resides in the link between transport and diffusion models. Furthermore, the non-self-adjointness of the differential operator appearing in the equations challenges the implementation of inexpensive reduced models for such high-fidelity problems, especially in the case of multi-query optimization problems.

An *a priori* error analysis and developments of *a posteriori* error estimates in the case of non-symmetric eigenvalue problems, as it is the case for criticality problems in neutronics, enabled a state-of-the-art approach for the construction of an associated reduce-order model. Several approaches were presented in order to ensure that the error estimates were reliable, efficient and allow an inexpensive, efficient at most, implementation of the reduced basis method. To do so, one crucial point was to examine and try to understand the information contained in the prefactors that appear in the error bounds for the eigenvalue and eigenvectors of the problems.

A key argument in the design of an inexpensive reduced basis method for criticality problems, as well as for any non-symmetric eigenvalue problem that shares similar mathematical properties, was to first consider the case where the high-fidelity problem offers an affine decomposition of the matrices along the parametric dependency. In that case, we implemented an efficient and inexpensive reduced basis method based on the greedy algorithm along with an *offline/online* procedure, which used the robust *a posteriori* error estimates that we developed. A few test cases provided showed that the considered reduced basis method exhibits satisfying convergence rates and computational cost, while they drew attention to the significance of the adjoint problem in the construction of the reduced-order model.

The urge of engineers and researchers in the nuclear field to have reliable optimization codes for multiple applications, such as loading pattern problems, naturally motivated the work of implementation of a reduced-order model in the code APOLLO3®, developed at CEA. Furthermore, reduced-order models are key elements in many studies in the field of neutronics, such as in data assimilation problems, as it is the subject of the last chapter of this manuscript.

While a first implementation of a reduced basis method must provide some of the deterministic solvers of APOLLO3®with an associated reduced-order model, it remains difficult to rely on an immediate efficient implementation of the reduced basis method based on the greedy algorithm, described in this manuscript, in the state-of-the-art and already existing deterministic neutron transport codes, due to the complexity behind assembling, in practice, the reduced matrices of the problem.

We may try to take advantage of any *offline/online* affine decomposition in the matrices to assemble, even though it is not guaranteed in most cases. Otherwise, in order to get an efficient computation of the Galerkin projection of the matrices, we may call some interpolation methods, such as the Generalized Empirical Interpolation Method (GEIM), but in that case, note that an additional bias in the calculations would need to be studied.

At last, a direct, even naive, application of the reduced-order model to optimization algorithms is still expected to provide results which compete with actual orders of computational times in the nuclear industry.

Bibliography

[1] C. Ahrens and G. Beylkin, *Rotationally invariant quadratures for the sphere*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 465 (2009), pp. 3103–3125.

[2] G. Allaire and G. Bal, *Homogenization of the criticality spectral equation in neutron transport*, ESAIM: M2AN, 33 (1999), pp. 721–746.

[3] G. Allaire, X. Blanc, B. Després, and F. Golse, *Transport et diffusion*, Ecole Polytechnique, 2019.

[4] E. Allen, *A finite element approach for treating the energy variable in the numerical solution of the neutron transport equation*, Transport Theory and Statistical Physics, 15 (1986), pp. 449–478.

[5] J.-P. Argaud, B. Bouriquet, F. de Caso, H. Gong, Y. Maday, and O. Mula, *Sensor placement in nuclear reactors based on the generalized empirical interpolation method*, Journal of Computational Physics, 363 (2018), pp. 354 – 370.

[6] J. P. Argaud, B. Bouriquet, H. Gong, Y. Maday, and O. Mula, *Monitoring flux and power in nuclear reactors with data assimilation and reduced models*, in International Conference on Mathematics and Computational Methods Applied to Nuclear Science & Engineering, 2017.

[7] J. P. Argaud, B. Bouriquet, H. Gong, Y. Maday, and O. Mula, *Stabilization of (G)EIM in presence of measurement noise: Application to nuclear reactor physics*, in Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2016: Selected Papers from the ICOSAHOM conference, June 27-July 1, 2016, Rio de Janeiro, Brazil, M. L. Bittencourt, N. A. Dumont, and J. S. Hesthaven, eds., Cham, 2017, Springer International Publishing, pp. 133–145.

[8] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells, *The deal.II library, version 8.5*, Journal of Numerical Mathematics, 25 (2017), pp. 137–145.

[9] M. Asadzadeh, *L^2-error estimates for the discrete ordinates method for three-dimensional neutron transport*, Transport Theory and Statistical Physics, 17 (1988), pp. 1–24.

[10] P. Auscher and P. Tchamitchian, *Square roots of elliptic second order divergence operators on strongly lipschitz domains: L^2 theory*, Journal d'Analyse Mathématique, 90 (2003), pp. 1–12.

[11] I. Babuška and J. Osborn, *Eigenvalue problems*, Handbook of numerical analysis, 2 (1991), pp. 641–787.

[12] G. Bal, *Couplage d'équations et homogénéisation en transport neutronique*, PhD thesis, Paris 6, 1997.

[13] A.-M. Baudron and J.-J. Lautard, *Simplified P_N transport core calculations in the APOLLO3® system*, in Proceedings of M&C 2011, Rio de Janeiro, Brazil, 2011.

[14] R. Becker, R. Koch, M. Modest, and H.-J. Bauer, *A finite element treatment of the angular dependency of the even-parity equation of radiative transfer*, Journal of Heat Transfer, 132 (2010), pp. 023404/1–13.

[15] G. I. Bell and S. Glasstone, *Nuclear Reactor Theory*, 10 1970.

[16] G. Berkooz, P. Holmes, and J. L. Lumley, *The proper orthogonal decomposition in the analysis of turbulent flows*, Annual review of fluid mechanics, 25 (1993), pp. 539–575.

[17] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Data assimilation in reduced modeling*, SIAM/ASA Journal on Uncertainty Quantification, 5 (2017), pp. 1–29.

[18] P. Binev, A. Cohen, O. Mula, and J. Nichols, *Greedy algorithms for optimal measurements selection in state estimation using reduced models*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 1101–1126.

[19] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numerica, 19 (2010), pp. 1–120.

[20] D. Boffi, A. Halim, and G. Priyadarshi, *Reduced basis approximation of parametric eigenvalue problems in presence of clusters and intersections*, arXiv preprint arXiv:2302.00898, (2023).

[21] S. Boyaval, C. Le Bris, T. Lelièvre, Y. Maday, N. C. Nguyen, and A. T. Patera, *Reduced basis techniques for stochastic problems*, Archives of Computational methods in Engineering, 17 (2010), pp. 435–454.

[22] T. Braconnier and N. J. Higham, *Computing the field of values and pseudospectra using the Lanczos method with continuation*, BIT Numerical Mathematics, 36 (1996), pp. 422–440.

[23] R. Bronson, G. B. Costa, and J. T. Saccoman, *Linear Algebra: Algorithms, Applications, and Techniques*, Academic Press, third ed., 2014.

[24] A. Buchan, C. Pain, F. Fang, and I. Navon, *A POD reduced-order model for eigenvalue problems with application to reactor physics*, International Journal for Numerical Methods in Engineering, 95 (2013), pp. 1011–1032.

[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: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 595–603.

[26] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 2228–2254.

[27] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: a unified framework*, Numerische Mathematik, 140 (2018), pp. 1033–1079.

[28] B. G. Carlson, *Solution of the transport equation by S_N approximations*, tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 1955.

[29] C. Carstensen and J. Gedicke, *Guaranteed lower bounds for eigenvalues*, Mathematics of Computation, 83 (2014), pp. 2605–2629.

[30] F. Casenave, *Reduced order methods applied to aeroacoustic problems solved by integral equations*, PhD thesis, Université Paris-Est, 2013.

[31] W. Chen, D. Yang, J. Zhang, C. Zhang, H. Gong, B. Xia, B. Quan, and L. Wang, *Study of non-intrusive model order reduction of neutron transport problems*, Annals of Nuclear Energy, 162 (2021), p. 108495.

[32] A. Cherezov, R. Sanchez, and H. G. Joo, *A reduced-basis element method for pin-by-pin reactor core calculations in diffusion and SP3 approximations*, Annals of Nuclear Energy, 116 (2018), pp. 195–209.

[33] A. Cohen, W. Dahmen, R. DeVore, J. Fadili, O. Mula, and J. Nichols, *Optimal reduced model algorithms for data-based state estimation*, SIAM Journal on Numerical Analysis, 58 (2020), pp. 3355–3381.

[34] A. Cohen, W. Dahmen, O. Mula, and J. Nichols, *Nonlinear reduced models for state and parameter estimation*, SIAM/ASA Journal on Uncertainty Quantification, 10 (2022), pp. 227–267.

[35] A. Cohen and R. DeVore, *Approximation of high-dimensional parametric PDEs*, Acta Numerica, 24 (2015), pp. 1–159.

[36] A. Cohen, R. DeVore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs*, Analysis and Applications, 09 (2011), pp. 11–47.

[37] Y. Conjungo Taumhas, G. Dusson, V. Ehrlacher, T. Lelièvre, and F. Madiot, *Reduced basis method for non-symmetric eigenvalue problems: application to the multigroup neutron diffusion equations*, arXiv preprint arXiv:2307.05978, (2023).

[38] Y. Conjungo Taumhas, D. Labeurthre, F. Madiot, O. Mula, and T. Taddei, *Impact of physical model error on state estimation for neutronics applications*, ESAIM: Proceedings and Surveys, 73 (2023), pp. 158–172.

[39] M. Coste-Delclaux, *Modélisation du phénomène d'autoprotection dans le code de transport multigroupe APOLLO2*, PhD thesis, Paris, CNAM, 2006.

[40] M. Coste-Delclaux, C. Diop, A. Nicolas, and B. Bonin, *Neutronique*, CEA Saclay; Groupe Moniteur, 2013.

[41] E. Dari, R. Duran, and C. Padra, *A posteriori error estimates for non-conforming approximation of eigenvalue problems*, Applied Numerical Mathematics, 62 (2012), pp. 580–591.

[42] R. Dautray and J.-L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Masson, 1984.

[43] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 6 Evolution Problems II*, Springer Science & Business Media, 2012.

[44] M. D. DeChaine and M. A. Feltus, *Nuclear fuel management optimization using genetic algorithms*, Nuclear Technology, 111 (1995), pp. 109–114.

[45] R. DeVore, G. Petrova, and P. Wojtaszczyk, *Data assimilation and sampling in Banach spaces*, Calcolo, 54 (2017), pp. 963–1007.

[46] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69, Springer Science & Business Media, 2011.

[47] B. Q. Do and L. P. Nguyen, *Application of a genetic algorithm to the fuel reload optimization for a research reactor*, Applied Mathematics and Computation, 187 (2007), pp. 977–988.

[48] J. J. Duderstadt and L. J. Hamilton, *Nuclear reactor analysis*, John Wiley & Sons, Inc., 1976.

[49] R. G. Durán, L. Gastaldi, and C. Padra, *A posteriori error estimators for mixed approximations of eigenvalue problems*, Mathematical Models and Methods in Applied Sciences, 9 (1999), pp. 1165–1178.

[50] A. Ern and J.-L. Guermond, *Finite Elements I: Approximation and interpolation*, vol. 72, Springer Nature, 2021.

[51] A. Ern and J.-L. Guermond, *Finite Elements III: First-Order and Time-Dependent PDEs*, vol. 74, Springer Nature, 2021.

[52] M. Ettehad and S. Foucart, *Instances of computational optimal recovery: dealing with observation errors*, SIAM/ASA Journal on Uncertainty Quantification, 9 (2021), pp. 1438–1456.

[53] D. Fournier and R. Le Tellier, *An adaptive energy discretization of the neutron transport equation based on a wavelet Galerkin method*, in Discrete Wavelet Transforms-Algorithms and Applications, IntechOpen, 2011.

[54] I. Fumagalli, A. Manzoni, N. Parolini, and M. Verani, *Reduced basis approximation and a posteriori error estimates for parametrized elliptic eigenvalue problems*, ESAIM Math. Model. Numer. Anal., 50 (2016), pp. 1857–1885.

[55] F. Galarce, J. Gerbeau, D. Lombardi, and O. Mula, *Fast reconstruction of 3D blood flows from doppler ultrasound images and reduced models*, Computer Methods in Applied Mechanics and Engineering, 375 (2021), p. 113559.

[56] E. M. Gelbard, *Simplified spherical harmonics equations and their use in shielding problems*, tech. rep., Westinghouse Electric Corp. Bettis Atomic Power Lab., Pittsburgh, 1961.

[57] P. German and J. C. Ragusa, *Reduced-order modeling of parameterized multi-group diffusion k-eigenvalue problems*, Annals of Nuclear Energy, 134 (2019), pp. 144–157.

[58] C. Geuzaine and J.-F. Remacle, *Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities*, International journal for numerical methods in engineering, 79 (2009), pp. 1309–1331.

[59] S. Giani, *An a posteriori error estimator for hp-adaptive discontinuous Galerkin methods for computing band gaps in photonic crystals*, Journal of Computational and Applied Mathematics, 236 (2012), pp. 4810–4826.

[60] S. Giani, L. Grubišić, H. Hakula, and J. S. Ovall, *A posteriori error estimates for elliptic eigenvalue problems using auxiliary subspace techniques*, Journal of Scientific Computing, 88 (2021), pp. 1–25.

[61] S. Giani, L. Grubišić, and A. Międlar, *Robust error estimates for approximations of non-self-adjoint eigenvalue problems*, Numerische Mathematik, (2016).

[62] L. Giret, *Numerical analysis of a non-conforming domain decomposition for the multigroup SPN equations*, PhD thesis, Université Paris-Saclay (ComUE), 2018.

[63] H. Gong, W. Chen, C. Zhang, and G. Chen, *Fast solution of neutron diffusion problem with movement of control rods*, Annals of Nuclear Energy, 149 (2020), p. 107814.

[64] H. Gong, Y. Maday, O. Mula, and T. Taddei, *PBDW method for state estimation: error analysis for noisy data and nonlinear formulation*, arXiv e-prints, (2019), p. arXiv:1906.00810.

[65] C. Grossmann, *Numerical treatment of partial differential equations*, Springer, 2007.

[66] M. Halsall, *Cactus, a characteristics solution to the neutron transport equations in complicated geometries*, tech. rep., UKAEA Atomic Energy Establishment, 1980.

[67] J. S. Hesthaven, G. Rozza, and B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, SpringerBriefs in Mathematics, Springer, 2016.

[68] S. G. Hong and N. Z. Cho, *Crx: a code for rectangular and hexagonal lattices based on the method of characteristics*, Annals of Nuclear Energy, 25 (1998), pp. 547–565.

[69] T. Horger, B. Wohlmuth, and T. Dickopf, *Simultaneous reduced basis approximation of parameterized elliptic eigenvalue problems*, Esaim Math. Model. Numer. Anal., 51 (2017), pp. 443–465.

[70] D. B. P. Huynh, D. J. Knezevic, and A. T. Patera, *A static condensation reduced basis element method: approximation and a posteriori error estimation*, ESAIM: Mathematical Modelling and Numerical Analysis, 47 (2013), pp. 213–251.

[71] A. Hébert, *Applied reactor physics*, Presses inter Polytechnique, 2009.

[72] S. Jia, H. Chen, and H. Xie, *A posteriori error estimator for eigenvalue problems by mixed finite element method*, Science China Mathematics, 56 (2013), pp. 887–900.

[73] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous galerkin method for a scalar hyperbolic equation*, Mathematics of computation, 46 (1986), pp. 1–26.

[74] T. Kato, *On the upper and lower bounds of eigenvalues*, Journal of the Physical Society of Japan, 4 (1949), pp. 334–339.

[75] T. Kato, *Fractional powers of dissipative operators*, Journal of the Mathematical Society of Japan, 13 (1961), pp. 246–274.

[76] T. Kato, *Perturbation theory for linear operators*, vol. 132, Springer Science & Business Media, 2013.

[77] M. G. Krein, *Linear operators leaving invariant a cone in a Banach space*, Amer. Math. Soc. Transl. Ser. I, 10 (1962), pp. 199–325.

[78] D. Labeurthre, *Development and comparison of high-order finite element bases for solving the transport equation on hexagonal meshes*, PhD thesis, Université Grenoble Alpes, 2022.

[79] J.-J. Lautard and J.-Y. Moller, *MINARET, a deterministic neutron transport solver for nuclear core calculations*, in Proceedings of M&C 2011, Rio de Janeiro (Brazil), 2011, American Nuclear Society.

[80] R. Le Tellier, D. Fournier, and J. Ruggieri, *A wavelet-based finite element method for the self-shielding issue in neutron transport*, Nuclear science and engineering, 163 (2009), pp. 34–55.

[81] R. Lee et al., *Argonne code centre: Benchmark problem book*, Report No.: ANL-7416, Supp, 2 (1976), pp. 277–466.

[82] P. Lesaint, *Finite element methods for symmetric hyperbolic equations*, Numerische Mathematik, 21 (1973), pp. 244–255.

[83] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, Publications mathématiques et informatique de Rennes, (1974), pp. 1–40.

[84] R. J. LeVeque, *Finite volume methods for hyperbolic problems*, vol. 31, Cambridge university press, 2002.

[85] E. E. Lewis and W. F. Miller, *Computational methods of neutron transport*, John Wiley and Sons, Inc., New York, NY, 1984.

[86] X. Liu, *A framework of verified eigenvalue bounds for self-adjoint differential operators*, Applied Mathematics and Computation, 267 (2015), pp. 341–355.

[87] S. Lorenzi, *An adjoint proper orthogonal decomposition method for a neutronics reduced order model*, Annals of Nuclear Energy, 114 (2018), pp. 245–258.

[88] L. Lunéville, *Méthode multibande aux ordonnées discrètes: formalisme et résultats*, tech. rep., Note CEA-N-2832, 1998.

[89] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D. V. Rovas, *Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems*, Comptes Rendus de l'Academie des Sciences-Series I-Mathematics, 331 (2000), pp. 153–158.

[90] Y. Maday and O. Mula, *A generalized empirical interpolation method: application of reduced basis techniques to data assimilation*, in Analysis and numerics of partial differential equations, Springer, 2013, pp. 221–235.

[91] Y. Maday, A. T. Patera, J. D. Penn, and M. Yano, *A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics*, International Journal for Numerical Methods in Engineering, 102 (2015), pp. 933–965.

[92] Y. Maday, A. T. Patera, and J. Peraire, *A general formulation for a posteriori bounds for output functionals of partial differential equations; application to the eigenvalue problem*, Comptes Rendus de l'Académie des Sciences-Series I-Mathematics, 328 (1999), pp. 823–828.

[93] Y. Maday, A. T. Patera, and G. Turinici, *A priori convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations*, Journal of Scientific Computing, 17 (2002), pp. 437–446.

[94] Y. Maday and T. Taddei, *Adaptive PBDW approach to state estimation: noisy observations; user-defined update spaces*, SIAM Journal on Scientific Computing, 41 (2019), pp. B669–B693.

[95] P. Mosca, *Conception et développement d'un mailleur énergétique adaptatif pour la génération des bibliothèques multigroupes des codes de transport*, PhD thesis, Université Paris Sud-Paris XI, 2009.

[96] P. Mosca, L. Bourhrara, A. Calloo, A. Gammicchia, F. Goubioud, L. Lei-Mao, F. Madiot, F. Malouch, E. Masiello, F. Moreau, S. Santandrea, D. Sciannandrone, I. Zmijarevic, E. Y. Garcias-Cervantes, G. Valocchi, J. Vidal, F. Damian, A. Brighenti, B. Vezzoni, P. Laurent, and A. Willien, *APOLLO3®: Overview of the new code capabilities for reactor physics analysis*, in Proceeding of International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering, M&C 2023, 2023.

[97] O. Mula, *Some contributions towards the parallel simulation of time dependent neutron transport and the integration of observed data in real time*, PhD thesis, Paris VI, 2014.

[98] O. Mula, *Inverse problems: A deterministic approach using physics-based reduced models*, in Model Order Reduction and Applications: Cetraro, Italy 2021, Springer, 2023, pp. 73–124.

[99] F. Nier and B. Helffer, *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*, vol. 1862, Springer Science & Business Media, 2005.

[100] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM Journal on Numerical Analysis, 28 (1991), pp. 133–140.

[101] J. Pitkäranta and L. R. Scott, *Error estimates for the combined spatial and angular approximations of the transport equation for slab geometry*, SIAM journal on numerical analysis, 20 (1983), pp. 922–950.

[102] A. Quarteroni, A. Manzoni, and F. Negri, *Reduced Basis Methods for Partial Differential Equations: An Introduction*, Springer, Aug. 2015.

[103] W. H. Reed and T. R. Hill, *Triangular mesh methods for the neutron transport equation*, tech. rep., Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[104] P. Ribon and J.-M. Maillard, *Les tables de probabilité (application au traitement des sections efficaces pour la neutronique. première partie, dans l'hypothèse statistique)*, tech. rep., Note CEA-N-2185, 1986.

[105] 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*, Archives of Computational Methods in Engineering, 15 (2007), p. 1.

[106] Y. Saad, *Numerical methods for large eigenvalue problems: second edition*, SIAM, 2011.

[107] A. Sartori, A. Cammi, L. Luzzi, M. Ricotti, and G. Rozza, *Reduced order methods: applications to nuclear reactor core spatial dynamics-15566*, in ICAPP 2015 Proceedings, 2015.

[108] A. Sartori, A. Cammi, L. Luzzi, and G. Rozza, *A reduced basis approach for modeling the movement of nuclear reactor control rods*, Journal of Nuclear Engineering and Radiation Science, 2 (2016).

[109] D. Schneider, F. Dolci, F. Gabriel, J.-M. Palau, M. Guillo, B. Pothet, P. Archier, K. Ammar, F. Auffret, R. Baron, A.-M. Baudron, P. Bellier, L. Bourhrara, L. Buiron, M. Coste-Delclaux, C. De Saint Jean, J.-M. Do, B. Espinosa, E. Jamelot, and I. Zmijarevic, *APOLLO3® : CEA/DEN deterministic multi-purpose code for reactor physics analysis*, in Proceedings of Physor 2016, Sun Valley, Idaho (USA), 2016, American Nuclear Society.

[110] G. W. Stewart, *On the early history of the Singular Value Decomposition*, SIAM Review, 35 (1993), pp. 551–566.

[111] T. Taddei, *An adaptive parametrized-background data-weak approach to variational data assimilation*, ESAIM: Mathematical Modelling and Numerical Analysis, 51 (2017), pp. 1827–1858.

[112] T. Takeda and H. Ikeda, *3-D neutron transport benchmarks*, Journal of Nuclear Science and Technology, 28 (1991), pp. 656–669.

[113] G. F. J. Temple, *The accuracy of Rayleigh's method of calculating the natural frequencies of vibrating systems*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 211 (1952), pp. 204–224.

[114] P. J. Turinsky, *Nuclear fuel management optimization: A work in progress*, Nuclear technology, 151 (2005), pp. 3–8.

[115] H. Victory, Jr, *Convergence properties of discrete-ordinates solutions for neutron transport in three-dimensional media*, SIAM Journal on Numerical Analysis, 17 (1980), pp. 71–83.

[116] W. Yang, H. Wu, Y. Zheng, and L. Cao, *Application of wavelets scaling function expansion method in resonance self-shielding calculation*, Annals of Nuclear Energy, 37 (2010), pp. 653–663.

Résumé étendu en français

Dans le domaine de l'énergie nucléaire, les réacteurs d'irradiation technologique (*Materials Testing Reactors*) sont des réacteurs de recherche visant à réaliser des expériences sur les matériaux ou les éléments combustibles des réacteurs de puissance, à la fin de leur cycle, comme c'est notamment le cas des réacteurs exploités par EDF. Au CEA, c'est le cas du réacteur OSIRIS du site de Saclay (arrêté en 2015), ainsi que du réacteur Jules Horowitz (RJH), toujours en construction au CEA/Cadarache. Ces réacteurs nucléaires ont la capacité de réaliser plusieurs expériences d'irradiation à l'intérieur du cœur, ou à l'échelle du réflecteur de neutrons, et assurent également la production de radio-isotopes à des fins médicales, en particulier le technétium 99m (99mTc). Néanmoins, ces réacteurs présentent de nombreuses hétérogénéités. La gestion fine des éléments combustibles utilisant de l'uranium enrichi (de l'ordre de 20 %), ainsi que le respect des exigences pour les différentes expériences d'irradiation sont des enjeux majeurs pour une utilisation optimale des éléments combustibles dans le réacteur. Le principal défi de ces expériences est de garantir les performances attendues en minimisant la consommation en combustible, tout en respectant les exigences de sûreté.

Plus généralement, l'exploitation des réacteurs de recherche et des EPR (*European Pressurized Reactors*, ou plus récemment *Evolutionary Power Reactors*) consiste en des expériences similaires, qui introduisent notamment le problème de l'optimisation des plans de chargement, et consistent à étudier la criticité à l'intérieur du cœur, ce qui revient à résoudre un problème aux valeurs propres non symétrique. La résolution de ce problème de haute fidélité (*high-fidelity*) s'accompagne d'un certain coût de calcul, qui devient très élevé lorsqu'il s'agit de problèmes d'optimisation, comme le problème d'optimisation du plan de chargement. Un tel problème nécessite en effet d'être résolu par de nombreux calculs de haute fidélité associés à différentes configurations du cœur. Pour effectuer ces calculs, il existe plusieurs codes au CEA, dont APOLLO3®. Ce code de calcul neutronique déterministe offre des fonctionnalités et des méthodes avancées qui permettent de réduire le biais de la modélisation sans pénaliser les temps de calcul, contrairement au schéma de calcul en deux étapes fourni par les codes de deuxième génération, tels que le code de transport APOLLO2 avec le code de calcul de cœur CRONOS2. Les travaux en cours au CEA visent à développer un schéma de calcul "best-estimate" utilisant les fonctionnalités d'APOLLO3®, en support aux études neutroniques des premiers cœurs de démarrage pour le RJH.

Cependant, un tel schéma de calcul avancé dans APOLLO3®ne répondra pas au besoin d'un outil d'exploitation pour l'irradiation du combustible et du cœur, dans le contexte d'un suivi en temps réel, ou dans celui de campagnes d'irradiation basées sur les différents états d'irradiation du combustible. Pour un schéma d'évaluation donné, le principal défi consiste à effectuer des calculs peu coûteux tout en préservant ou en contrôlant les biais de modélisation et les erreurs de calcul. En effet, l'élément clé consiste à effectuer des calculs de cœur peu coûteux, tout en étant capable de les réactualiser en cas d'aléas au cours de l'exploitation. Par exemple, un retrait inattendu d'une expérience peut survenir au cours du cycle. Dans le cas du réacteur OSIRIS, un calcul de cycle de combustible prenait quelques minutes.

Dans ce contexte, une approche type "bases réduites" est proposée. Elle consiste à développer et à mettre en œuvre un modèle d'ordre réduit (ROM) pour les calculs de criticité en neutronique, via le développement et l'utilisation d'estimateurs d'erreurs, basés sur une analyse *a posteriori* pour les problèmes aux valeurs propres généralisés et non symétriques, qui permettent entre autres de quantifier l'erreur d'approximation.

Le plan de ce manuscrit est le suivant.

Dans le Chapitre 1, nous rappelons et décrivons les techniques classiques de discrétisation pour le calcul du cœur et les problèmes de criticité en neutronique.

Le Chapitre 2 vise ensuite à établir des estimateurs d'erreur *a posteriori* implémentables et peu coûteux pour les problèmes aux valeurs propres généralisés non symétriques. Ce chapitre reprend notamment l'état de l'art sur les estimateurs d'erreur développés dans le cas symétrique et le lien existant entre l'estimation d'erreur et le gap spectral pour les problèmes aux valeurs propres. Une analyse d'erreur *a priori* appréhende la construction d'un modèle d'ordre réduit dans le cas non symétrique via une méthode de projection de Galerkin, et indique à quel point il est important de considérer les vecteurs propres à gauche et à droite dans notre approche. Ensuite, la principale difficulté consiste à proposer dans ce contexte des estimateurs d'erreur fiables, efficaces et implémentables. Pour ce faire, nous développons des estimateurs d'erreur basées sur les résidus qui présentent tous des *préfacteurs* multiplicatifs dépendant des paramètres dans les bornes d'erreur. Le déploiement d'une approche heuristique permet l'estimation de ces quantités, sous certaines hypothèses, car elles ne sont pas calculables en pratique, mais contiennent des informations clés dans le comportement de l'erreur.

Ensuite, dans le Chapitre 3, nous détaillons la mise en œuvre d'une implémentation "efficace" d'une méthode de base réduite, utilisant les estimateurs d'erreur *a posteriori* développés dans le chapitre précédent, dans le cas d'une décomposition affine des matrices de haute fidélité en ce qui concerne leur dépendance paramétrique. Basée sur un algorithme "gloûton" (*greedy*, en anglais), elle consiste en une procédure en deux étapes *offline/online* et en une projection de Galerkin du problème de haute fidélité sur un espace réduit bien choisi.

Le Chapitre 4 illustre les performances d'une telle méthode de base réduite sur des codes-maquettes de diffusion de neutrons à deux groupes d'énergie, à travers plusieurs tests numériques. Un premier test sur un petit cœur de réacteur nucléaire non physique met en évidence la nécessité de prendre en compte l'ensemble de la borne supérieure dans l'estimation de l'erreur pour la certification du modèle d'ordre réduit. Le deuxième cas test, qu'on appelle le *Minicore*, montre dans quelle mesure la méthode de base réduite fournit un modèle fiable dans des temps de calcul très courts. Enfin, le troisième cas test justifie numériquement le raisonnement qui soulève la pertinence d'inclure à la fois les vecteurs propres directs et adjoints dans l'estimation de l'erreur.

Ensuite, dans le Chapitre 5, nous présentons une première implémentation d'une méthode de réduction de modèles dans le code APOLLO3®. Bien qu'à ce stade, la complexité liée à l'implémentation de la projection de Galerkin des matrices de haute fidélité relève une difficulté non négligeable pour l'implémentation "efficace" d'une procédure de type *greedy*, une méthode de décomposition orthogonale aux valeurs propres (POD) est proposée dans ce contexte. Celle-ci fournit des résultats prometteurs, notamment avec l'utilisation des estimateurs d'erreur, qui ont été développés et utilisés dans les chapitres précédents, dans la certification du modèle réduit alors construit.

Enfin, dans le Chapitre 6, nous donnons une application industrielle d'un modèle

d'ordre réduit issu de la POD dans le contexte de l'estimation d'état et de l'assimilation de données pour la reconstruction de la nappe de puissance dans le cœur, à partir de mesures expérimentales. À noter que ce dernier chapitre est un *proceeding* publié issu de la session de recherche de l'école d'été CEMRACS 2021.

Chapitre 1.

Ce chapitre a été rédigé en se basant sur les références ci-dessous :

- [40] M. Coste-Delclaux, C. Diop, A. Nicolas, and B. Bonin, Neutronique, CEA Saclay; Groupe Moniteur, 2013.
- [97] O. Mula, Some contributions towards the parallel simulation of time dependent neutron transport and the integration of observed data in real time, PhD thesis, Paris VI, 2014.
- [62] L. Giret, Numerical analysis of a non-conforming domain decomposition for the multigroup SPN equations, PhD thesis, Université Paris-Saclay (ComUE), 2018.
- [78] D. Labeurthre, Development and comparison of high-order finite element bases for solving the transport equation on hexagonal meshes, PhD thesis, Université Grenoble Alpes, 2022.

Nous commençons par un aperçu général des techniques de discrétisation pour l'équation de Boltzmann. Dans la Section 1.1, nous rappelons d'abord l'équation de transport des neutrons dépendant du temps, qui donne un modèle général pour la dynamique de la neutronique dans le cœur d'un réacteur nucléaire. Dans la Section 1.2, nous introduisons le problème de criticité en neutronique, qui découle de l'équation de transport de neutrons stationnaire, et qui est le problème généralisé aux valeurs propres qui nous intéresse dans ce travail. Dans la Section 1.3, nous rappelons quelques techniques standards de discrétisation du problème continu. Enfin, dans la Section 1.4, nous évoquons la motivation de l'approximation du modèle de transport des neutrons, par le modèle de diffusion neutronique.

Dans ce chapitre, nous avons donc présenté les techniques classiques utilisées pour discrétiser un tel problème aux valeurs propres dérivé de la neutronique. La dépendance paramétrique du problème motive l'utilisation d'estimateurs d'erreur *a posteriori* à des fins d'optimisation, par exemple, dans le cas où l'on doit résoudre le problème aux valeurs propres pour un très grand nombre de valeurs de paramètres (cf. le problème d'optimisation du plan de chargement). Dans ce cas particulier, afin d'éviter des coûts de calcul considérables, on peut envisager de résoudre un problème approximatif moins coûteux. Ainsi, ces estimateurs peuvent quantifier l'erreur d'approximation sans nécessairement résoudre le problème de référence. La section suivante est consacrée au développement de ces estimateurs dans le cas d'un problème aux valeurs propres généralisé et non symétrique.

Chapitre 2.

Dans ce chapitre, nous proposons des estimateurs d'erreur *a posteriori* pour les problèmes aux valeurs propres généralisés non symétriques. Dans le cadre des problèmes aux valeurs propres symétriques, l'analyse d'erreur *a posteriori* repose naturellement sur des résidus quantifiés selon une norme d'énergie induite par l'opérateur. Nous devons ici généraliser ces estimateurs *a posteriori* dans le cas des problèmes aux valeurs propres généralisés non symétriques, tels que celui de la diffusion neutronique multigroupe, présenté à la fin du chapitre précédent. Pour les problèmes aux valeurs propres non généralisés, nous montrons l'existence, dans la borne supérieure de l'erreur, d'un *prefacteur* dépendant du paramètre qui doit être pris en compte afin d'obtenir des estimateurs fiables. Le calcul d'une valeur précise et optimale de ce préfacteur n'est pas une tâche facile, contrairement au cas des problèmes aux valeurs propres symétriques où il peut être exprimé au moyen du gap spectral de l'opérateur considéré.

Nous commençons, dans la Section 2.1, par revoir la relation spécifique, soulevée par I. Babuška et J. Osborn [11] dans une analyse *a priori*, entre l'erreur sur valeur propre et les erreurs sur les vecteurs propres de meilleure approximation à gauche et à droite, dans le cas des approximations de Galerkin d'un problème aux valeurs propres généralisé non symétrique. Dans la Section 2.2, nous rappelons quelques estimateurs d'erreur *a posteriori* classiques basés sur les résidus pour les problèmes aux valeurs propres. Nous dérivons ensuite, dans la Section 2.3, certaines bornes d'erreur sur les vecteurs propres à gauche et à droite, ainsi que sur la valeur propre, dans le cas d'un problème aux valeurs propres généralisé et non symétrique. Comme nous avons besoin d'estimateurs d'erreur *a posteriori* calculables et fiables qui tiennent compte des informations contenues dans les préfacteurs, nous proposons, dans la Section 2.4, une méthode heuristique pratique pour estimer ces quantités. Pour ce faire, nous fournissons quelques éléments d'analyse théorique pour illustrer le lien étroit entre l'expression obtenue du préfacteur et son expression bien connue dans le cas des problèmes aux valeurs propres symétriques.

Dans ce chapitre, nous avons élaboré des estimateurs d'erreur *a posteriori* basés sur les résidus pour un problème aux valeurs propres généralisé et non symétrique donné. Ces estimateurs présentent tous des *préfacteurs* dépendant des paramètres qui sont trop coûteux à calculer en pratique pour un grand nombre de paramètres. Par conséquent, nous avons d'abord fourni quelques éléments d'analyse théorique pour illustrer le lien étroit entre l'expression obtenue du *préfacteur* et son expression bien connue dans le cas des problèmes aux valeurs propres symétriques. En particulier, nous avons soulevé des arguments perturbatifs pour donner un développement au premier ordre du *préfacteur* lorsque l'opérateur est une petite perturbation d'un opérateur symétrique, comme c'est le cas avec les problèmes aux valeurs propres dérivés de la neutronique, bien que ce soit plus complexe en pratique, car nous devons nous attaquer à un problème aux valeurs propres généralisé, qui ne rassemble pas toutes les hypothèses que nous avons émises. Nous avons finalement proposé une approche heuristique basée sur la donnée pour estimer le *préfacteur*, en supposant que sa dépendance paramétrique peut être ignorée. Le développement d'estimateurs d'erreur *a posteriori* pour les problèmes aux valeurs propres généralisés non symétriques est une étape cruciale dans le développement d'une méthode de base réduite peu coûteuse. Dans ce contexte, ces estimateurs permettent une construction "efficace" de l'espace d'approximation via une procédure itérative, et certifient également la précision des solutions au problème réduit.

Chapitre 3.

Les méthodes de réduction de modèles telles que les méthodes de bases réduites (RB) [21, 67, 102] sont utiles lorsqu'il s'agit d'accélérer le temps de calcul de solutions approximatives à des problèmes paramétrés. Dans le contexte de la neutronique, ces problèmes sont naturellement issus de l'optimisation du plan de chargement d'un cœur de réacteur nucléaire [44, 47, 114]. Mathématiquement, cela revient à optimiser une fonction "objectif" qui implique la solution d'un problème aux valeurs propres généralisé non symétrique. Le but de ce chapitre est donc d'établir, dans ce contexte, une méthodologie pour la mise en œuvre d'une méthode de base réduite, en deux phases (*offline* et *online*), pour les problèmes aux valeurs propres généralisés non symétriques, dépendant de paramètres. Elle peut être considérée comme une généralisation de [69, 54, 20], où des méthodes de bases réduites pour les problèmes aux valeurs propres symétriques ont été développées. Pour ce faire, nous nous appuyons sur les estimateurs d'erreur *a posteriori* pratiques développés dans le chapitre précédent, qui permettent, dans la phase *offline*, de construire l'espace réduit avec un algorithme *greedy* [25], en rompant la dépendance vis-à-vis du solveur de haute fidélité, contrairement aux procédures type POD (voir [16] pour une introduction générale), par exemple ; dans la phase *online*, ils certifient l'approximation et permettent une analyse de convergence du problème réduit. Ici, nous considérons une situation particulière dans laquelle les opérateurs linéaires présentent une dépendance affine en le paramètre. Cette situation doit être prise en compte pour que la base réduite itérative puisse rivaliser avec la POD. Cela donne un exemple d'implémentation de la méthode où les quantités d'intérêt sont assemblées de manière "efficace", en *online*, en fonction du paramètre.

Dans la Section 3.1, nous présentons le problème réduit (RB) aux valeurs propres comme une approximation de Galerkin du problème aux valeurs propres de haute fidélité (HF). Dans la Section 3.2, nous considérons la décomposition affine des matrices de haute fidélité en le paramètre, et nous expliquons comment cette hypothèse clé permet une implémentation "efficace" de la méthode de base réduite et des estimateurs d'erreur *a posteriori*. La Section 3.3 détaille la phase *offline* de la procédure RB, qui s'appuie sur l'algorithme *greedy*, et montre comment l'espace réduit est minutieusement construit. La Section 3.4 présente la phase *online* de la procédure RB qui consiste en la projection de Galerkin sur l'espace réduit, et l'assemblage du problème réduit et des quantités algébriques d'intérêt le long des paramètres.

Grâce aux estimateurs d'erreur *a posteriori* pratiques développés dans le Chapitre 2, nous avons mis en œuvre une méthodologie d'implémentation de la méthode de base réduite pour la résolution d'un problème aux valeurs propres généralisé non symétrique et dépendant d'un paramètre. Basée sur l'algorithme *greedy*, et sous l'hypothèse que les matrices du problème de haute fidélité présentent une décomposition affine le long du paramètre, elle permet une implémentation "efficace" de la méthode de base réduite, qui fournit un modèle réduit dont le coût de calcul rivalise avec d'autres méthodes de réduction de modèles, telles que les méthodes de décomposition orthogonale aux valeurs propres (POD). L'utilisation d'estimateurs d'erreur *a posteriori* permet également une certification du modèle réduit obtenu et fournit une approche adaptative dans la construction de la base réduite.

Chapitre 4.

Le but de ce chapitre est d'illustrer le comportement de la méthode de base réduite détaillée dans le chapitre précédent, sur des exemples provenant d'applications en neutronique. Nous considérons ici les équations de la diffusion neutronique à deux groupes d'énergie, définies en (1.41), et dotées des mêmes hypothèses que celles présentées en Section (1.4.2) de ce manuscrit. Nous concentrons notre étude sur trois cas tests. Dans la Section 4.1, nous proposons une première application de la méthode de base réduite à un cœur de récateur simpliste en 2D composé de quatre régions de matériaux, doté de propriétés et de sections efficaces non physiques. Cet exemple permet une analyse computationnelle rapide de la méthode de base réduite par le biais d'estimateurs d'erreur *a posteriori*, dans le cas où l'on choisit de ne pas considérer dans ceux-ci le fameux préfacteur. Dans la Section 4.2, un cœur rectangulaire plus réaliste, nommé le *Minicore*, met en avant les prouesses de la méthode de base réduite en termes de temps de calcul et de convergence, grâce à l'implémentation d'estimateurs d'erreurs *a posteriori* pratiques, qui cette fois-ci prennent en compte une estimation heuristique du préfacteur. Enfin, dans la Section 4.3, nous soulignons l'importance de considérer le problème adjoint dans la construction de la base réduite avec des calculs rapides sur un *benchmark* neutronique de réacteur à eau pressurisée (REP) en 3D.

Pour les deux premiers exemples, les calculs et les implémentations sont effectués sur un code maquette, écrit en Python 3.6. La discrétisation de haute fidélité le long de la variable d'espace est générée en utilisant les méthodes du projet open-source FEniCS. Pour le troisième et dernier exemple, nous utilisons un code POD fourni par le professeur Jean Ragusa, écrit en MATLAB [57], basé sur la bibliothèque d'éléments finis *deal.II* [8], et le générateur de maillage GMSH [58].

Les trois cas tests présentés dans ce chapitre mettent en évidence la capacité de la méthode de base réduite à fournir un modèle d'ordre réduit peu coûteux, fiable et certifié pour les équations de la diffusion neutronique. Les deux premiers cas tests montrent que le modèle réduit implémenté est capable de donner le facteur de multiplication (valeur propre) du modèle de haute fidélité à l'ordre du pcm (à 10^{-5} près), dans un temps de calcul de l'ordre de la milliseconde, tout en confirmant que l'approche heuristique dans le développement des estimateurs d'erreur *a posteriori* définit des représentants de l'erreur fiables et acceptables pour ce problème. Enfin, l'approche POD pour le *benchmark* dans le dernier cas test illustre la nécessité d'inclure la contribution du problème adjoint dans le modèle réduit, afin d'obtenir un modèle fiable d'ordre réduit.

Chapitre 5.

Dans ce chapitre, nous nous intéressons à l'équation de diffusion des neutrons paramétrée, lorsqu'elle est résolue plusieurs fois pour différentes valeurs des paramètres, comme c'est le cas par exemple dans le problème d'optimisation du plan de chargement du cœur. Nous nous concentrons sur l'approximation multigroupe sur un intervalle d'énergie $[E_{\min}, E_{\max}] = [E_G, E_{G-1}] \cup \ldots \cup [E_1, E_0]$, où G représente le nombre de groupes d'énergie considéré. Étant donné un paramètre μ, l'équation de la diffusion multigroupe en régime stationnaire [48, Chapitre 7] (voir Section 1.4.2) recherche le flux scalaire de neutrons multigroupe $\phi_\mu = \left(\phi_\mu^1, \ldots, \phi_\mu^G\right)$ associé au facteur de multiplication $k_{\text{eff},\mu}$ (la valeur propre de plus grand

module) à l'intérieur du cœur du réacteur nucléaire.

Dans ce travail, nous proposons une méthode reposant sur une approximation de l'espace de l'ensemble des solutions à l'aide d'une décomposition orthogonale aux valeurs propres (POD). Comme dans la méthodologie base réduite décrite dans le Chapitre 3, la méthode se décrit en deux phases. Dans la phase *offline*, nous construisons un espace réduit qui approche l'espace des solutions du problème de diffusion multigroupe. Dans la phase *online*, pour tout nouvel ensemble de paramètres donné, nous résolvons un problème réduit sur l'espace réduit dans un temps de calcul beaucoup plus court que le temps de calcul lié à la résolution du problème de haute fidélité (5.1). Dans ce chapitre, nous nous concentrons sur le développement de la méthode dans le projet APOLLO3® [96], une plateforme partagée entre le CEA, FRAMATOME et EDF, qui comprend différents solveurs déterministes pour l'équation de transport et de diffusion des neutrons. Nous nous intéressons particulièrement au solveur MINARET [79] dans l'approximation de la diffusion, discrétisé avec des éléments finis discontinus.

Les deux cas tests qui ont été développés mettent en évidence la possibilité et la pertinence d'implémenter une méthode de base réduite dans le code APOLLO3®, en termes de précision et de réduction du temps de calcul. Il convient de noter que les estimateurs d'erreur *a posteriori* dans le contexte des bases réduites peuvent être appliqués dans une approche de type *greedy* lors de la phase *offline* [25, 106, 61], comme cela est fait dans le Chapitre 3, de manière à minimiser les appels au solveur de haute fidélité, ou dans une certification *online* du modèle réduit. Pour ce faire, nous devrions étudier la manière de calculer les matrices réduites en se défaisant, autant que possible, de leur dépendance paramétrique. Nous pourrions, par exemple, envisager d'utiliser une méthode d'interpolation empirique, comme GEIM [90, 30]. Cela fera l'objet de futurs travaux.

Chapitre 6.

Ce chapitre vient à la fin de ce manuscrit pour offrir un autre exemple illustratif de l'application d'une réduction de modèles dans le cadre de la neutronique. Il propose un modèle réduit distinct de celui détaillé dans les Chapitres 2 à 5, puisqu'il se base sur des mesures expérimentales, et vise à reconstruire la nappe de puissance à l'intérieur d'un réacteur nucléaire.

Il s'agit d'une publication sous forme de *proceeding*, issue du projet MOCO de la session de recherche de l'école d'été CEMRACS 2021, consacrée à l'assimilation de données et à la réduction de modèle pour les problèmes de très haute dimension. Sa référence dans le manuscrit est :

[38] Y. Conjungo Taumhas, D. Labeurthre, F. Madiot, O. Mula, and T. Taddei, Impact of physical model error on state estimation for neutronics applications, ESAIM: Proceedings and Surveys, vol. 73 (2023), pp. 158–172.

Dans ce travail, nous examinons le problème inverse d'estimation d'état du champs de puissance nucléaire dans un réacteur d'une centrale électrique à partir d'un nombre limité d'observations du flux de neutrons. Pour ce faire, nous utilisons une méthode nommée PBDW (*Parametrized Background Data Weak*). Cette méthode combine les observations avec un modèle d'EDP paramétrée pour le comportement du flux de neutrons au sein

du réacteur. Comme en général, même les modèles les plus sophistiqués ne peuvent pas reproduire parfaitement la réalité, une biais de modélisation est inévitable. Nous étudions l'impact du biais de modèle dans la reconstruction de la nappe de puissance lorsque nous utilisons un modèle de diffusion pour le flux de neutrons et que nous supposons que la véritable physique est régie par un modèle de transport de neutrons.

En conclusion.

Bien qu'une première implémentation d'une méthode de base réduite soit suffisante pour fournir à certains des solveurs déterministes d'APOLLO3®un modèle d'ordre réduit associé, il reste difficile de compter sur une implémentation efficace (au sens peu coûteuse) immédiate de la méthode de base réduite basée sur l'algorithme *greedy*, décrite dans ce manuscrit, dans les codes de transport de neutrons déterministes de pointe déjà existants, en raison de la complexité de l'assemblage, en pratique, des matrices réduites du problème.

Nous pouvons essayer de tirer parti de toute décomposition affine *offline/online* dans les matrices à assembler, même si cela n'est pas garanti dans la plupart des cas. Autrement, afin d'obtenir un calcul efficace de la projection de Galerkin des matrices de haute fidélité, nous pouvons faire appel à certaines méthodes d'interpolation, telles que la méthode d'interpolation empirique généralisée (GEIM), mais dans ce cas, il convient de noter qu'un biais supplémentaire dans les calculs devra être étudié.

Enfin, une application directe, même naïve, du modèle d'ordre réduit aux algorithmes d'optimisation des plans de chargement devrait encore fournir des résultats qui rivalisent avec les ordres de grandeur des temps de calcul dans l'industrie nucléaire.