HAL open science

Reduced basis method for non-symmetric eigenvalue problems: application to the multigroup neutron diffusion equations

Yonah Conjungo Taumhas, Geneviève Dusson, Virginie Ehrlacher, Tony Lelièvre, François Madiot

To cite this version:

Yonah Conjungo Taumhas, Geneviève Dusson, Virginie Ehrlacher, Tony Lelièvre, François Madiot. Reduced basis method for non-symmetric eigenvalue problems: application to the multigroup neutron diffusion equations. 2023. cea-04156959

HAL Id: cea-04156959

https://cea.hal.science/cea-04156959v1

Preprint submitted on 10 Jul 2023

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

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

Reduced basis method for non-symmetric eigenvalue problems: application to the multigroup neutron diffusion equations

Yonah Conjungo Taumhas[1], Geneviève Dusson[2], Virginie Ehrlacher[3], Tony Lelièvre[3], and François Madiot[1]

[1]Université Paris-Saclay, CEA, Service d'Études des Réacteurs et de Mathématiques Appliquées, 91191, Gif-sur-Yvette, France
[2]Laboratoire de Mathématiques de Besancon, UMR CNRS 6623, Université de Franche-Comté, 16 route de Gray, 25030 Besançon, France
[3]CERMICS, École des Ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne-la-Vallée, Cedex 2, France

July 10, 2023

Abstract

In this article, we propose a reduced basis method for parametrized non-symmetric eigenvalue problems arising in the loading pattern optimization of a nuclear core in neutronics. To this end, we derive *a posteriori* error estimates for the eigenvalue and left and right eigenvectors. The practical computation of these estimators requires the estimation of a constant called prefactor, which we can express as the spectral norm of some operator. We provide some elements of theoretical analysis which illustrate the link between the expression of the prefactor we obtain here and its well-known expression in the case of symmetric eigenvalue problems, either using the notion of numerical range of the operator, or via a perturbative analysis. Lastly, we propose a practical method in order to estimate this prefactor which yields interesting numerical results on actual test cases. We provide detailed numerical simulations on two-dimensional examples including a multigroup neutron diffusion equation.

1 Introduction

In this work, we are interested in developing a numerical method to efficiently compute the solutions of a parametrized non self-adjoint eigenvalue problem for a large number of parameter values. An example of application where this type

of problem occurs and which motivates the present work is the resolution of criticality problems in neutronics. The method we propose relies on a reduced basis technique [20, 21].

Model order reduction methods such as reduced basis techniques [5, 15, 22] are useful to accelerate the computation of approximate solutions to parameterized problems. In the context of neutronics, parametrized problems naturally occur when optimizing the loading pattern of a nuclear core [7, 8, 24]. Mathematically, this amounts to optimizing an objective function which involves the solution to a generalized non-symmetric eigenvalue problem, parameterized by the fuel assemblies distribution. The objective of this work is thus to propose a reduced basis technique in this context. It can be seen as a generalization of [16, 11, 4], where reduced basis methods for symmetric eigenvalue problems have been developed. A main difficulty is to obtain reliable *a posteriori* estimators in order to build the reduced basis and certify the results obtained with the reduced problem. We refer to [12, 14] for *a posteriori* error estimators for non self-adjoint eigenvalue problems in a classical finite element context. Let us also mention the recent work [10] and references therein for efficient *a posteriori* estimators for non-symmetric problems. We refer to [6, 13, 19, 23] for some other applications of model order reduction techniques applied to neutronics.

As mentioned above, the main bottleneck here is to propose efficient *a posteriori* error estimators for a reduced basis approximation of non self-adjoint eigenvalue problems. More precisely, we consider a situation where one is interested in computing the eigenvalue of smallest modulus of a parameterized eigenvalue problem, which is assumed to be simple. The *a posteriori* error estimators read as products of norms of the residuals of the direct and associated adjoint eigenvalue problems times a multiplicative constant, which we call hereafter the *prefactor*. 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.

Three main contributions are proposed in this work. First, we derive an expression of the prefactor in the case of non-symmetric eigenvalue problems as the spectral norm of a composition of some well-chosen operators. Second, we provide some elements of theoretical analysis to illustrate the close link between the obtained expression of the prefactor and its well-known expression in the case of symmetric eigenvalue problems. This link is highlighted in two different ways: first, we give an expression of the prefactor using the distance between the approximate eigenvalue and the numerical range of the non-symmetric operator and observe that the numerical range plays a similar role as the spectrum of the operator in the self-adjoint case; second, we use perturbative arguments to give a second-order development of this prefactor when the operator is a small perturbation of a symmetric operator. As our third contribution, we propose a practical heuristic method to estimate the prefactor in the present reduced basis context and demonstrate the efficiency of the approach on test cases stemming from neutronics applications.

The outline of this article is as follows. In Section 2, we describe the prototypical reference problem of interest as well as the model order reduction

method we use, which relies on a greedy algorithm to build the reduced basis. The greedy procedure requires *a posteriori* estimators which are presented in Section 3. These estimators are basically built as the products of residual norms with prefactors, whose computations are discussed in Section 4. Finally, we provide numerical results on two different examples in Section 5: a toy example of a two-dimensional two-group problem, and a two-dimensional simple model of a minicore.

2 Reduced basis method for non-symmetric generalized eigenvalue problems

The objective of this section is to introduce the mathematical framework and the model order reduction method we consider. In Section 2.1, we describe the reference high-fidelity generalized eigenvalue problem of interest. The reduced-order model is then presented in Section 2.2, and the greedy algorithm used to build the reduced basis is finally explained in Section 2.3.

2.1 Reference high-fidelity problem

Let us present the parametrized generalized eigenvalue problem for which we wish to build a reduced-order problem. Let $\mathcal{N} \in \mathbb{N}^*$ be a positive integer which is assumed to be large in our context. In all the following, $\mathbb{R}^{\mathcal{N}}$ is endowed with the Euclidean scalar product[1] denoted by $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. For all values of the vector of parameters μ belonging to the set of parameter values $\mathcal{P} \subset \mathbb{R}^p$ for some $p \geq 1$, we consider two matrices A_μ and B_μ in $\mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ and the following generalized eigenvalue problem: Find $(u_\mu, \lambda_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{C}$ such that λ_μ is an eigenvalue with minimal modulus:

$$A_\mu u_\mu = \lambda_\mu B_\mu u_\mu, \quad \|u_\mu\| = 1. \tag{2.1}$$

As is classical in the context of reduced basis methods, we refer to problem (2.1) as the *high-fidelity* (HF) problem. We make the following additional assumption which is satisfied in the problems we are eventually interested in for neutronics applications.

Assumption 2.1. *For any parameter $\mu \in \mathcal{P}$, A_μ is invertible and there exists a unique positive eigenvalue λ_μ which realizes the smallest modulus solution to* (2.1)*. Moreover, the eigenvalue λ_μ is simple.*

A consequence of Assumption 2.1 is that u_μ is uniquely defined (up to a sign), λ_μ is real, and that there is a spectral gap between λ_μ and the other eigenvalues solutions to problem (2.1), a property that will also play a role in the *a posteriori* analysis below. The associated adjoint problem then reads: Find $(u^*_\mu, \lambda_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R}^*_+$ such that

$$A^T_\mu u^*_\mu = \lambda_\mu B^T_\mu u^*_\mu, \quad \|u^*_\mu\| = 1. \tag{2.2}$$

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

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 scalar 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 line vector of $\mathbb{R}^{\mathcal{N}}$ such that

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

From Assumption 2.1, the eigenvectors u_μ and u_μ^* can be chosen real. In practice, the solutions to (2.1) and (2.2) are approximated by an inverse power method, which will be properly described in Algorithm 2.

We also define for all $\mu \in \mathcal{P}$ the so-called effective multiplication factor

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

There holds

$$k_\mu = \frac{\langle u_\mu^*, B_\mu u_\mu\rangle}{\langle u_\mu^*, A_\mu u_\mu\rangle}. \tag{2.3}$$

Remark 2.2. *On the one hand, Assumption 2.1 holds for instance if A_μ is invertible and the matrix $A_\mu^{-1}B_\mu$ coming from problem* (2.1) *satisfies the assumptions of the Perron–Frobenius theorem [1]. Note that under the assumption that A_μ is invertible, λ_μ is solution to* (2.1) *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 mentioned earlier and detailed in Section 5.1,* (2.1) *is obtained as an appropriate discretization of a continuous problem where the associated resolvent operator satisfies the assumptions of the Krein–Rutman theorem and thus admits a simple real greatest eigenvalue in modulus denoted k_μ^{ex}. Since $1/k_\mu^{\mathrm{ex}}$ is solution to the continuous problem, the smallest eigenvalue of* (2.1) *in modulus is also expected to be simple and positive for fine enough discretization, i.e. large enough $\mathcal{N}$.*

We now assume in all the rest of the article that Assumption 2.1 is satisfied.

Lemma 2.3. *Under Assumption 2.1, $\langle u_\mu^*, A_\mu u_\mu\rangle \neq 0$.*

We postpone the proof of this lemma after Lemma 3.2.

Remark 2.4. *Without Assumption 2.1, it is possible that $\langle u_\mu^*, A_\mu u_\mu\rangle = 0$. Indeed, a simple example is to take $A_\mu = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ and $B_\mu = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Let $u_\mu = (1, 0)$ and $u_\mu^* = (0, 1)$. Equations* (2.1) *and* (2.2) *are satisfied with $\lambda_\mu = 1$ while $\langle u_\mu^*, A_\mu u_\mu\rangle = 0$.*

We are interested in situations where one has to solve the reference high-fidelity problem (2.1) quickly and for many values of μ. The idea is to build a reduced basis using some solutions of (2.1) (so-called snapshots) computed offline, and to use a Galerkin method to project the problem (2.1) onto this

reduced basis, see Section 2.2. This requires *a posteriori* estimators to wisely select the parameters used to build the reduced basis, as well as to certify the numerical results obtained online on the reduced basis: this is discussed in Section 3.

2.2 Reduced-order model

The aim of this section is to present the reduced-order model obtained from a given reduced basis to get an approximation of (2.1). 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 that any solution of problem (2.1) can be accurately approximated by an element of the space $\mathcal{V}$ (the construction of such a subspace will be discussed in the next section). A reduced-order model for problem (2.1) 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 $1 \leq i, j \leq N$,

$$A^{ij}_{\mu,N} := \langle \xi_i, A_\mu \xi_j \rangle,$$
$$B^{ij}_{\mu,N} := \langle \xi_i, B_\mu \xi_j \rangle.$$

The reduced-order model then consists in solving: 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 to

$$A_{\mu,N} c_{\mu,N} = \lambda_{\mu,N} B_{\mu,N} c_{\mu,N}, \quad u_{\mu,N} = \sum_{i=1}^{N} c^i_{\mu,N} \xi_i, \quad \text{and} \quad \|u_{\mu,N}\| = 1, \tag{2.4}$$

where for all $1 \leq i \leq N$, $c^i_{\mu,N}$ is the ith component of the vector $c_{\mu,N}$. Similarly as in Section 2.1 (see Assumption 2.1), we make the following assumption.

Assumption 2.5. *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 realizes the smallest modulus solution to* (2.4)*. 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 scalar product $\langle \cdot, \cdot \rangle_{\ell^2}$, we can consider the solution to the associated reduced adjoint problem: Find $(c^*_{\mu,N}, \lambda_{\mu,N}) \in \mathbb{R}^N \times \mathbb{R}^*_+$ such that the eigenvalue $\lambda_{\mu,N}$ is the smallest in modulus and

$$A^t_{\mu,N} c^*_{\mu,N} = \lambda_{\mu,N} B^t_{\mu,N} c^*_{\mu,N}, \quad u^*_{\mu,N} = \sum_{i=1}^{N} c^{*,i}_{\mu,N} \xi_i, \quad \text{and} \quad \|u^*_{\mu,N}\| = 1. \tag{2.5}$$

where for all $1 \leq i \leq N$, $c^{*,i}_{\mu,N}$ is the ith component of the vector $c^*_{\mu,N}$ and $A^t_{\mu,N}$ and $B^t_{\mu,N}$ 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), and we define

$$k_{\mu,N} = \frac{\langle c^*_{\mu,N}, B_{\mu,N} c_{\mu,N}\rangle_{\ell^2}}{\langle c^*_{\mu,N}, A_{\mu,N} c_{\mu,N}\rangle_{\ell^2}} = \frac{\langle u^*_{\mu,N}, B_\mu u_{\mu,N}\rangle}{\langle u^*_{\mu,N}, A_\mu u_{\mu,N}\rangle}. \tag{2.6}$$

In practice, we use the inverse power method to solve (2.4) and (2.5). 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 power methods 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 the next section. Note that the power methods applied to (2.4) and (2.5) (resp. to (2.1) and (2.2)) are guaranteed to converge if Assumption 2.5 (resp. Assumption 2.1) is satisfied. In the numerical examples presented in Section 5, we observe that the two power methods (for the direct and adjoint 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 Section 5).

2.3 Choice of the reduced space: greedy algorithm

In practice, the reduced space $\mathcal{V}$ used in the reduced-order model described in the previous section is built following the standard procedure of the reduced basis technique [20]. We first initialize the reduced space $\mathcal{V}_0$ as a very low-dimensional space spanned by the first modes obtained from a Proper Orthogonal Decomposition of a family of vectors composed of a few snapshots of the direct and adjoint problems. A sequence of parameter values $(\mu_n)_{n\geq 1}$ is then selected from a greedy procedure described below, from which nested reduced spaces $(\mathcal{V}_n)_{n\geq 1}$ are built as follows:

$$\forall n \geq 1, \quad \mathcal{V}_n = \mathrm{Span}\left\{u_{\mu_1}, \ldots, u_{\mu_n}, u^*_{\mu_1}, \ldots, u^*_{\mu_n}\right\}. \tag{2.7}$$

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 reduced eigenvalue problems described in the previous section for $\mathcal{V} = \mathcal{V}_n$.

Remark 2.6. *The choice made in* (2.7) *to enrich the sequence of reduced spaces with both the eigenvector of the direct and of the adjoint eigenvalue problem stems from the a priori error analysis of Galerkin approximations of generalized eigenvalue problems (see [2, 3]). Indeed, in an asymptotic regime, the error between the approximate and exact eigenvalue scales like*

$$|\lambda_\mu - \lambda_{\mu,N}| \leq C_\mu \frac{1}{\gamma_{\mu,N}} \varepsilon_N \varepsilon^*_N,$$

where C_μ is a positive constant which only depends on the parameter μ, where

$$\varepsilon_N := \inf_{v_N \in \mathcal{V}} \|u_\mu - v_N\|,$$
$$\varepsilon^*_N := \inf_{v_N \in \mathcal{V}} \left\|u^*_\mu - v_N\right\|,$$

and where $\gamma_{\mu,N}$ is the inf-sup constant of the reduced eigenvalue problem. 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 the direct and adjoint eigenvalue problems, in order to at least get the reference solution for the reduced problem when considering the parameter μ of the snapshots.

In the greedy procedure, the parameters $(\mu_n)_{\geq 1}$ need to be selected satisfying some criteria. In practice, it is common to choose a finite subset $\mathcal{P}_{\text{train}} \subset \mathcal{P}$ of parameter values, called hereafter a *training set* and selecting the snapshots maximizing some error surrogate for the error between solutions of the reference model and the reduced model Δ_{N_n}, as is described in Algorithm 1. In an *ideal greedy* procedure, we would take 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 error on the eigenvalue: $\Delta_{N_n}(\mu) := e^k_{N_n}(\mu)$

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

with

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

However, these quantities are of course not available in general, so one has to resort to *a posteriori* error estimate for an *efficient greedy* algorithm. Therefore, the aim of the following section is to detail different strategies to define an *a posteriori* error estimator $\Delta_N(\mu)$ 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 exact eigenvalue problem.

Algorithm 1 GREEDY ALGORITHM

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 $\mathcal{V} := \mathcal{V}_n \subset \mathbb{R}^{\mathcal{N}}$

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 exact eigenvalue problems (2.1) and (2.2), and the solutions of the reduced eigenvalue problems (2.4) and (2.5).

This section is organized as follows. Sections 3.1 and 3.2 are respectively dedicated to error estimates on the eigenvectors and on the eigenvalue. In Section 4.1, we draw connections between the estimators we introduce in our context of non-symmetric eigenvalue problems, and the classical ones used for symmetric eigenvalue problems. Finally, we introduce in Section 4.2 the pratical *a posteriori* error estimators that will be used for numerical experiments in Section 5.

To simplify notation, the subscript μ is omitted in this section, as only one parameter value $\mu \in \mathcal{P}$ is considered. Therefore, the quantities u, u^*, λ, k are the solutions of the high fidelity problem, while $u_N, u_N^*, \lambda_N, k_N$ are the solutions of the reduced problem. We are therefore aiming at deriving 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

$$R_N = (B - k_N A)\, u_N, \tag{3.1}$$

$$R_N^* = \left(B^T - k_N A^T\right) u_N^*. \tag{3.2}$$

We moreover define the vector

$$\tilde{u}^* = \frac{A^T u^*}{\|A^T u^*\|}, \tag{3.3}$$

and the matrix

$$M = A^{-1} B, \tag{3.4}$$

which is well-defined since A is invertible from Assumption 2.1. Note that it then holds that

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

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 $\mathrm{Span}\{\tilde{u}^*\}^\perp$ and $\mathrm{Span}\{u\}^\perp$ respectively. More precisely, P and P^* are defined by

$$P = I - \frac{u(\tilde{u}^*)^T}{\langle u, \tilde{u}^*\rangle}, \tag{3.5}$$

$$P^* = I - \frac{\tilde{u}^* u^T}{\langle u, \tilde{u}^*\rangle}, \tag{3.6}$$

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

Lemma 3.1. *The spectral projector onto the eigenspace of M associated with the simple eigenvalue k is $I - P$, where P is defined by* (3.5).

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, which is defined by:

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

where $\mathcal{C}_k$ is a closed contour in the complex plane such that k is the only eigenvalue of M contained inside the loop. 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 3.2. *There holds*

(i) $P^2 = P$;

(ii) Ker $P = \text{Span}\{u\}$, Ran $P = [\text{Span}\{\tilde{u}^*\}]^\perp$ *and these two spaces are stable 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 stable 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 $v \in \text{Ker } P$, $PMv = MPv = 0$. Besides, for all $v \in \text{Ran } P$, it holds that $Pv = v$, and $Mv \in \text{Ran } P$ 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 on P^*, using similar arguments as in the proof of Lemma 3.2.

Lemma 3.3. *There holds*

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

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

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

We have now gathered enough results to prove Lemma 2.3.

Proof of Lemma 2.3. Let us argue by contradiction. If $\langle u_\mu^*, A_\mu u_\mu \rangle = \langle \tilde{u}_\mu^*, u_\mu \rangle = 0$, then $\text{Span}\{u_\mu\} \subset (\text{Span}\{\tilde{u}_\mu^*\})^\perp$. Yet, we have $\text{Span}\{u_\mu\} + (\text{Span}\{\tilde{u}_\mu^*\})^\perp = \mathbb{R}^{\mathcal{N}}$ using Lemma 3.2-(ii). This yields to a contradiction and concludes the proof. □

Let us introduce some notation. By Lemma 3.2, the operator $PMP - k_N I$ leaves RanP $= [\text{Span}\{\tilde{u}^*\}]^\perp$ invariant. Besides, provided that $k_N \notin \sigma\left(PMP|_{[\text{Span}\{\tilde{u}^*\}]^\perp}\right)$, the operator $(PMP - k_N I)\,|_{[\text{Span}\{\tilde{u}^*\}]^\perp}$ is invertible, seen as an operator from $[\text{Span}\{\tilde{u}^*\}]^\perp$ onto $[\text{Span}\{\tilde{u}^*\}]^\perp$. We can thus define the Moore–Penrose inverse of this operator, denoted by $(PMP - k_N I)^+$ and defined (by linearity) as follows:

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

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

Proposition 3.4 (Eigenvector estimation)**.** *Let* $u_N, u_N^* \in \mathbb{R}^{\mathcal{N}} \backslash \{0\}$ *and let* $k_N \in \mathbb{R}$ *such that* $k_N \notin \sigma((PMP)|_{[\text{Span}\{\tilde{u}^*\}]^\perp})$ *and* $k_N \notin \sigma((P^* M^T P^*)|_{[\text{Span}\{u\}]^\perp}$. *Then, the following estimates hold:*

$$\inf_{v \in \text{Span}\{u\}} \|u_N - v\| \leq C_N^u \|R_N\|, \tag{3.7}$$

$$\inf_{v^* \in \text{Span}\{u^*\}} \|u_N^* - v^*\| \leq C_N^{u^*} \|R_N^*\|, \tag{3.8}$$

with

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

Here and in the following, 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 3.5. *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 3.6. *The results obtained in Proposition 3.4 match those of [14, Proposition 4] for $k = 0$, noting the slightly different definition of the residual to take into account the generalized eigenvalue problem.*

Proof. First, there holds

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

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 $Pv = 0$. Besides, 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 3.2 (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 $v \in [\mathrm{Span}\{\tilde{u}^*\}]^\perp$,

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

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

Then using Lemma 3.2 (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 (3.1), we obtain

$$Pu_N = P\left(PMP - k_N I\right)^+ PA^{-1} R_N. \tag{3.9}$$

Thus,

$$\|Pu_N\| \leqslant \left\| P\left(PMP - k_N I\right)^+ PA^{-1} \right\| \|R_N\|. \tag{3.10}$$

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 using (3.3) and (3.6)

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

Now, using Lemma 3.3 (iii) and similar arguments as above, there holds

$$\begin{aligned}P^*\left(A^Tu_N^*\right) &= P^*\left(P^*M^TP^*-k_NI\right)^+\left(P^*M^TP^*-k_NI\right)P^*A^Tu_N^*\\ &= P^*\left(P^*M^TP^*-k_NI\right)^+P^*\left(M^T-k_NI\right)A^Tu_N^*\\ &= P^*\left(P^*M^TP^*-k_NI\right)^+P^*\left(B-k_NA\right)^Tu_N^*.\end{aligned}$$

Hence

$$P^*\left(A^Tu_N^*\right) = P^*\left(P^*M^TP^*-k_NI\right)^+P^*R_N^*. \tag{3.11}$$

Then,

$$\left\|A^{-T}P^*A^Tu_N^*\right\| \leqslant \left\|A^{-T}P^*\left(P^*M^TP^*-k_NI\right)^+P^*\right\|\|R_N^*\|,$$

which proves (3.8). □

3.2 Error estimate on the eigenvalue

We now provide an estimate for the eigenvalue.

Proposition 3.7 (Eigenvalue estimation)**.** *Let $u_N, u_N^* \in \mathbb{R}^{\mathcal{N}}$. Under Assumption 2.1 and the assumptions 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*

$$|k_N-k| \leq C_N^k\eta_N^k, \tag{3.12}$$

with

$$\eta_N^k := \frac{\|R_N\|\|R_N^*\|}{|\langle u_N^*, Au_N\rangle|}, \tag{3.13}$$

and

$$C_N^k := \left\|[P^*\left(P^*M^TP^*-k_NI\right)^+P^*]^T(M-kI)P\left(PMP-k_NI\right)^+PA^{-1}\right\|. \tag{3.14}$$

Remark 3.8. *Note that in this result, the vectors u_N and u_N^* may not be solutions of a reduced eigenvalue problem of the form (2.4) or (2.5). The only requirement of Proposition 3.7 is that k_N has to be related to u_N and u_N^* by the formula stated in the proposition.*

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

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

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

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

According to Lemma 2.3, we can set

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

so that we find

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

Using (3.9) and (3.11) finishes the proof. □

4 Practical estimates of efficiencies and prefactors

As in the previous section, to simplify notation, the subscript μ is again omitted in this section.

In view of Proposition 3.4 and Proposition 3.7, 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, respectively,

$$\Delta_N^k := \overline{C}_N^k \eta_N^k, \qquad \Delta_N^u := \overline{C}_N^u \|R_N\|, \qquad \Delta_N^{u^*} := \overline{C}_N^{u^*}\|R_N^*\|, \tag{4.1}$$

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 3.4 and Proposition 3.7.

For the applications we are interested in, as will be illustrated below in Section 5.5.1, we observe that the operators are perturbations of symmetric operators. This is why we investigate in Section 4.1 the links between the prefactor C_N^k introduced above and well-known results about this constant in the symmetric case. However, this does not yield practical efficient formulas for the prefactors. This is why we propose in Section 4.2 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 use in the numerical results to build practical *a posteriori* error estimators in the greedy algorithm to select the reduced space. This heuristic approach gives very interesting numerical results for neutronics applications as will be illustrated in Section 5.

4.1 Some connections with the symmetric case

The goal of this section is to draw links between the prefactor C_N^k defined in Section 3.2, and the prefactor which is traditionnally used for the computation of *a posteriori* error estimators for symmetric eigenvalue problems. We begin this section by recalling well-known results about *a posteriori* error estimators for symmetric eigenvalue problems in Section 4.1.1. In particular, we recall that the value of this prefactor in the symmetric context is directly linked with the value of the spectral gap of the exact problem. We then provide two different approaches to relate the value of the constant C_N^k defined by (3.14) to the value of the prefactor in the symmetric case: (i) we prove in Section 4.1.2 that in the non-symmetric case, the constant C_N^k can be estimated using the distance between the approximate eigenvalue and the *numerical range* of the non-symmetric operator; (ii) in Section 4.1.3 we study the perturbative regime where the non-symmetric operator can be seen as a small perturbation of a symmetric operator, and check that the prefactor for the non-symmetric operator is also a small perturbation of the well-known expression of the prefactor in the symmetric case.

4.1.1 Symmetric case

The aim of this section is to recall some well-known results about *a posteriori* error estimators for symmetric eigenvalue problems, i.e. in the case where A is a positive definite symmetric matrix and $B = I$, so that $M = A^{-1}$. In this case, all the eigenvalues of M are real and positive, k being its 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, there holds that $u = u^*$ 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 (3.14))

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

and we have the following proposition.

Proposition 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

$$k > k_N > k_2 > 0. \tag{4.2}$$

Then, there holds

$$C_N^k = \frac{k_2(k-k_2)}{(k_N-k_2)^2} = \frac{\|PA^{-1}\|\|M-kI\|}{\mathrm{dist}(k_N, \sigma((PMP)|_{\mathrm{Span}\{u\}^\perp}))^2}. \tag{4.3}$$

Proof. Denoting by $k = k_1 > k_2 \geq \cdots \geq k_{\mathcal{N}}$ the eigenvalues of $M = A^{-1}$ and by $u_1, u_2, \ldots, u_{\mathcal{N}}$ corresponding eigenvectors, there holds

$$A^{-1} = M = \sum_{i=1}^{\mathcal{N}} k_i u_i u_i^T,$$

and

$$P = \sum_{i=2}^{\mathcal{N}} u_i u_i^T.$$

Then, using functional calculus, there holds

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

Since the operator norm is associated with the Euclidean vector norm, the operator norm corresponds to the largest eigenvalue of the (symmetric) operator, so 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 (4.2), and $\mathrm{dist}(k_N, \sigma((PMP)|_{\mathrm{Span}\{u\}^\perp}))^2 = (k_N-k_2)^2$, 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 context.

4.1.2 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 of the non-symmetric operator. Let us first recall the notion of numerical range.

Definition 4.2. *Let $Q \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$. The numerical range of the matrix Q is defined by*

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

Lemma 4.3. *Let $Q \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ let and $z \notin \sigma(Q)$. Then,*

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

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

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

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

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

□

Proposition 4.4. *Under the same assumptions as in Proposition 3.7, there holds*

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

Note that the bound given by Proposition 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 (3.14), it holds that

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

Using Lemma 4.3, we easily obtain the result. □

The upper bound on C_N^k stated in Proposition 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 assumption.

4.1.3 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

$$\begin{cases} A = A^\varepsilon = S + \varepsilon T \text{ with } S^T = S,\, T^T = -T,\, \varepsilon > 0, \\ B = I. \end{cases} \tag{4.4}$$

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 denote by $k_S := \frac{1}{\lambda_S}$ and by $k_{S,i} := \frac{1}{\lambda_{S,i}}$ 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 direct eigenvector, $u^{*,\varepsilon}$ an associated adjoint eigenvector, $\tilde{u}^{*,\varepsilon}$ defined as in (3.3) 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 $\mathcal{V} = \mathrm{Span}\{u_S\}$. 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

$$C_N^{k,\mathrm{sym}} = \frac{k_{S,2}(k_S - k_{S,2})}{(k_N - k_{S,2})^2} \tag{4.5}$$

as ε goes to 0.

We first perform a first-order expansion of the eigenvectors and eigenvalues in ε (cf. Chapter 2 of [18]).

Lemma 4.5. *Let us assume* (4.4) *and*

$$\|u^\varepsilon\|^2 = \|u_S\|^2 = 1 \quad \text{and} \quad \langle u^\varepsilon, u_S\rangle > 0. \tag{4.6}$$

Then, as ε goes to 0,

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

Proof. Using the results of [18, 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,

$$Su_{A,1} + Tu_{A,0} = \lambda_{A,0}u_{A,1} + \lambda_{A,1}u_{A,0}. \tag{4.7}$$

Using (4.6), 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

$$\langle u_{A,0}, u_{A,1}\rangle = 0. \tag{4.8}$$

Using the latter and projecting (4.7) 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 (4.7) transforms into

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

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

$$u_{A,1} = -\,(S - \lambda_S I)^{-1}_{|\mathrm{Span}\{u_S\}^\perp}\, Tu_S. \tag{4.9}$$

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 4.6. *Let us assume* (4.4) *and* (4.6)*. Then, as* ε *goes to* 0*,*

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

where

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

$u_{A,1}$ *being defined in* (4.9)*.*

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 (3.14), namely

$$\mathcal{M}^\varepsilon = [P^\varepsilon \left(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I\right)^+ P^\varepsilon](M^\varepsilon - k^\varepsilon I) P^\varepsilon \left(P^\varepsilon M^\varepsilon P^\varepsilon - k_N I\right)^+ P^\varepsilon (A^\varepsilon)^{-1}. \tag{4.10}$$

Lemma 4.7. *Let us assume* (4.4) *and* (4.6)*. Then, as* ε *goes to* 0*,*

$$\mathcal{M}^\varepsilon = \mathcal{M}_0 + \varepsilon \mathcal{M}_1 + O(\varepsilon^2), \tag{4.11}$$

with

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

and

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

Proof. First,

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

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

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

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

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

Hence, one can write

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

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

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

we easily obtain (4.11). □

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

Proposition 4.8. *Let us assume* (4.4) *and* (4.6)*. Let us also assume that*

$$k_S \geq k_N > k_{S,2} > 0, \tag{4.12}$$

and that $k_{S,2}$ is not degenerate. Then, for ε sufficiently small, there holds

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

where $C_N^{k,\mathrm{sym}}$ is defined by (4.5).

Proof. Starting from (4.11), 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

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

From this, we deduce that

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

Note also that the same holds for Γ_S with eigenvalues

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

Then, noting that $k_{S,2}$ 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

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

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

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

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

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 observed on Figure 1. This is a strong indication that the estimate of Proposition 4.8 is sharp.

In our practical applications of interest, we will indeed observe that the operator is a perturbation of a symmetric operator, but the estimate of the prefactor by the one obtained using the symmetric part is not sufficiently good over a large range of the values of the parameters μ, in particular because the spectral gap (see Assumption (4.12)) is not uniformly bounded from below (see Section 5.5.1 for a discussion). This is why we will resort to a practical heuristic method to approximate the prefactor, as is now explained in the next section 4.2.

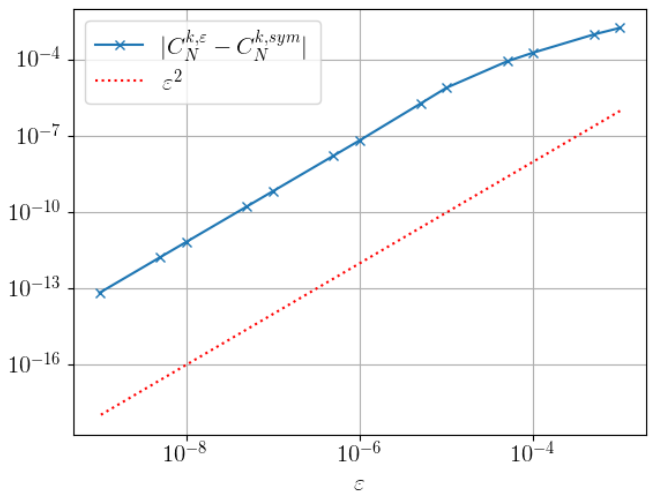

Figure 1: $|C_N^{k,\varepsilon} - C_N^{k,sym}|$ as a function of ε

4.2 Practical *a posteriori* error estimator

The aim of this section is to present the heuristic algorithm we use in order to estimate the prefactors C_N^k, C_N^u and $C_N^{u^*}$ defined in Proposition 3.7 and Proposition 3.4 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 used to build *a posteriori* error estimates for the greedy algorithm presented in Algorithm 1.

This heuristic procedure is based on the use of an estimation 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 other words, the estimation set $\mathcal{P}_{\mathrm{pref}}$ is chosen so that $\mathcal{P}_{\mathrm{train}} \cap \mathcal{P}_{\mathrm{pref}} = \emptyset$. High-fidelity solutions of the eigenvalue problems (2.1) are computed in the offline

phase for all $\mu \in \mathcal{P}_{\text{pref}}$.

For a given reduced space V_N, let us introduce the efficiency ratios: for all $\mu \in \mathcal{P}$,

$$\mathcal{E}_N^k(\mu) := \frac{|k_{\mu,N} - k_\mu|}{\eta_N^k(\mu)},\ \mathcal{E}_N^u(\mu) := \frac{\|u_{\mu,N} - u_\mu\|}{\|R_N(\mu)\|} \text{ and } \mathcal{E}_N^{u^*}(\mu) := \frac{\|u^*_{\mu,N} - u^*_\mu\|}{\|R^*_N(\mu)\|}. \tag{4.13}$$

The latter quantities are computed in the offline phase for all $\mu \in \mathcal{P}_{\text{pref}}$. By definition, it holds that for all $\mu \in \mathcal{P}$,

$$\mathcal{E}_N^k(\mu) \le C_N^k(\mu),\ \mathcal{E}_N^u(\mu) \le C_N^u(\mu) \text{ and } \mathcal{E}_N^{u^*}(\mu) \le C_N^{u^*}(\mu). \tag{4.14}$$

Our heuristic approach then consists in 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

$$\overline{C}_N^k := \max_{\mu \in \mathcal{P}_{\text{pref}}} \mathcal{E}_N^k(\mu), \qquad \overline{C}_N^u := \max_{\mu \in \mathcal{P}_{\text{pref}}} \mathcal{E}_N^u(\mu), \text{ and } \qquad \overline{C}_N^{u^*} := \max_{\mu \in \mathcal{P}_{\text{pref}}} \mathcal{E}_N^{u^*}(\mu), \tag{4.15}$$

the practical *a posteriori* error estimates used in the greedy algorithm are then defined by

$$\Delta_N^k(\mu) := \overline{C}_N^k \eta_N^k(\mu), \qquad \Delta_N^u(\mu) := \overline{C}_N^u \|R_N(\mu)\|, \text{ and } \qquad \Delta_N^{u^*}(\mu) := \overline{C}_N^{u^*} \|R^*_N(\mu)\|. \tag{4.16}$$

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

5 Numerical results

The aim of this section is to illustrate the behaviour of the proposed reduced basis method on examples arising from neutronics applications. The considered physical model is presented in Section 5.1. In Section 5.2, we describe the high-fidelity discretization of the problem using a finite element method. The parametric dependency of the coefficients of the mathematical equations describing the model enables the matrices to be assembled using a so-called affine decomposition, discussed in Section 5.3. The eigenvalue solver is described in Section 5.4. Finally, numerical tests presented in Section 5.5 give an application of the reduced basis method to nuclear core computations.

5.1 The continuous model: two-group neutron diffusion equations

The stationary neutron flux density in a reactor core is determined by solving the transport equation which depends on six variables: position (3-dimensional), velocity direction (2-dimensional), the velocity norm or energy (1-dimensional).

It physically states the balance between the emission of neutrons by fission and the absorption, scattering, and leakage of neutrons at the boundary of the spatial domain. The most common discretization of the energy variable is the multigroup approximation where the energy domain is divided into subintervals called energy groups. The reactor core is modeled by a bounded, connected and open subset Ω of $\mathbb{R}^d$ (where typically $d = 3$, but one ot two-dimensional are also considered) having a Lipschitz boundary which is piecewise regular. In practice, the neutron flux density is usually modeled by the multigroup neutron diffusion equations [9, Chapter 7] at the reactor core scale.

Let us now make precise the specific two-group neutron diffusion model we consider in this work. For a given value $\mu \in \mathcal{P}$ (the parameter set $\mathcal{P}$ will be presented below), we consider the two-group neutron diffusion equations where the neutron flux $u_\mu^{\text{ex}} := \left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right) \in H_0^1(\Omega)^2$ is decomposed into the neutron flux of high energy $\phi_{1,\mu}^{\text{ex}} \in H_0^1(\Omega)$ and thermal energy $\phi_{2,\mu}^{\text{ex}} \in H_0^1(\Omega)$. This model reads as

Find $\left(u_\mu^{\text{ex}} := \left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right), \lambda_\mu^{\text{ex}}\right) \in H_0^1(\Omega)^2 \times \mathbb{R}$ such that λ_μ^{ex} is an eigenvalue with minimal modulus, and

$$\begin{cases} -\operatorname{div}\left(D_{1,\mu}\nabla\phi_{1,\mu}^{\text{ex}}\right) + \Sigma_{11,\mu}\phi_{1,\mu}^{\text{ex}} + \Sigma_{12,\mu}\phi_{2,\mu}^{\text{ex}} = \lambda_\mu^{\text{ex}} F_{1,\mu}\left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right) & \text{in } \Omega, \\ -\operatorname{div}\left(D_{2,\mu}\nabla\phi_{2,\mu}^{\text{ex}}\right) + \Sigma_{21,\mu}\phi_{1,\mu}^{\text{ex}} + \Sigma_{22,\mu}\phi_{2,\mu}^{\text{ex}} = \lambda_\mu^{\text{ex}} F_{2,\mu}\left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right) & \text{in } \Omega, \\ \phi_{i,\mu}^{\text{ex}} = 0, \quad \text{on } \partial\Omega, \ i = 1, 2, \end{cases} \tag{5.1}$$

supplemented with a normalization condition on u_μ^{ex}. Here, for all $i, j \in \{1, 2\}$, all $\mu \in \mathcal{P}$ and all $\phi_1, \phi_2 \in H_0^1(\Omega)$,

- $F_{i,\mu}(\phi_1, \phi_2) := \chi_{i,\mu}\left((\nu\Sigma_f)_{1,\mu}\phi_1 + (\nu\Sigma_f)_{2,\mu}\phi_2\right)$;
- $\chi_{i,\mu} : \Omega \to \mathbb{R}$ is the neutron total spectrum of group i;
- $\nu_{i,\mu} : \Omega \to \mathbb{R}$ is the average number of neutrons emitted per fission of group i;
- $\Sigma_{fi,\mu} : \Omega \to \mathbb{R}$ is the fission cross section of group i;
- $D_{i,\mu} : \Omega \to \mathbb{R}_+$ is the diffusion coefficient of group i;
- $\Sigma_{ij,\mu} : \Omega \to \mathbb{R}$ with $\Sigma_{ij,\mu} = \begin{cases} \Sigma_{ti,\mu} - \Sigma_{s,0,ii,\mu} & \text{if } i = j, \\ -\Sigma_{s,0,ij,\mu} & \text{otherwise;} \end{cases}$
- $\Sigma_{ti,\mu} : \Omega \to \mathbb{R}$ is the total cross section of group i;
- $\Sigma_{s,0,ij,\mu} : \Omega \to \mathbb{R}$ is the Legendre moment of order 0 of the scattering cross section from group i to group j.

Note that in the equations above, we used the short-hand notation $(\nu\Sigma_f)_{i,\mu}$ to refer to the product $\nu_{i,\mu}\Sigma_{fi,\mu}$ for $i = 1, 2$. The so-called effective multiplication factor $k_\mu^{\text{ex}} := \dfrac{1}{\lambda_\mu^{\text{ex}}}$ measures the balance between the production and loss of

neutrons. If $k_\mu^{\text{ex}} = 1$, the nuclear chain reaction is self-sustaining; if $k_\mu^{\text{ex}} > 1$, the chain reaction is diverging; if $k_\mu^{\text{ex}} < 1$, the chain reaction vanishes.

Let us now describe the considered parametric dependency. We introduce a partition $(\Omega_m)_{m=1}^M$ of the domain Ω with $M \in \mathbb{N}^*$ so that for all $1 \leq m \leq M$, Ω_m is a domain with Lipschitz, piecewise regular boundaries. For $i = 1, 2$, the coefficients $D_{i,\mu}$, $\Sigma_{ij,\mu}$, $\chi_{i,\mu}$, $(\nu\Sigma_f)_{i,\mu}$ are assumed to be piecewise regular on each domain Ω_m for $1 \leq m \leq M$. The parameter set $\mathcal{P}$ then refers to the set of values that each of these coefficients can possibly take on each subdomain $(\Omega_m)_{1\leq m\leq M}$. In other words, the choice of a parameter value $\mu \in \mathcal{P}$ corresponds to a choice of the values of each of these coefficients on all the subdomains.

In the following, we assume that the set of admissible parameter values $\mathcal{P}$ is such that all the coefficients of the model belong to $L^\infty(\Omega)$ and that there exists $\alpha > 0$ and $0 < \varepsilon < 1$ such that for all $\mu \in \mathcal{P}$ and all $i, j \in \{1, 2\}$, $i \neq j$, almost everywhere in Ω,

$$\begin{aligned}
&\alpha \leq D_{i,\mu} \quad \text{a.e. in } \Omega,\\
&\alpha \leq \Sigma_{ii,\mu} \quad \text{a.e. in } \Omega,\\
&|\Sigma_{ij,\mu}| \leq \varepsilon \Sigma_{ii,\mu} \quad \text{a.e. in } \Omega,\\
&0 \leq (\nu\Sigma_f)_{i,\mu} \quad \text{a.e. in } \Omega,\\
&\text{there exists } \tilde{i}, \tilde{j} \in \{1, 2\} \text{ such that } \chi_{\tilde{i},\mu}(\nu\Sigma_f)_{\tilde{j},\mu} \neq 0 \in L^\infty(\Omega),
\end{aligned}$$

so that Problem (5.1) is well-posed for all $\mu \in \mathcal{P}$. The variational formulation associated to Problem (5.1) writes:

Find $\left(u_\mu^{\text{ex}} := \left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right), \lambda_\mu^{\text{ex}}\right) \in (H_0^1(\Omega)^2 \times \mathbb{R})$ such that λ_μ^{ex} is an eigenvalue with minimal modulus and,

$$a_\mu(\left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right), (\psi_1, \psi_2)) = \lambda_\mu^{\text{ex}} b_\mu(\left(\phi_{1,\mu}^{\text{ex}}, \phi_{2,\mu}^{\text{ex}}\right), (\psi_1, \psi_2)) \quad \text{for all } (\psi_1, \psi_2) \in H_0^1(\Omega)^2, \tag{5.2}$$

where for all $(\phi_1, \phi_2), (\psi_1, \psi_2) \in H_0^1(\Omega)^2$,

$$\begin{aligned}
a_\mu((\phi_1, \phi_2), (\psi_1, \psi_2)) &:= \int_\Omega (D_{1,\mu}\nabla\phi_1) \cdot \nabla\psi_1 + \Sigma_{11,\mu}\phi_1\psi_1 + \Sigma_{12,\mu}\phi_2\psi_1\\
&\quad + \int_\Omega (D_{2,\mu}\nabla\phi_2) \cdot \nabla\psi_2 + \Sigma_{21,\mu}\phi_1\psi_2 + \Sigma_{22,\mu}\phi_2\psi_2,\\
b_\mu((\phi_1, \phi_2), (\psi_1, \psi_2)) &:= \int_\Omega \chi_{1,\mu}\left((\nu\Sigma_f)_{1,\mu}\phi_1 + (\nu\Sigma_f)_{2,\mu}\phi_2\right)\psi_1\\
&\quad + \int_\Omega \chi_{2,\mu}\left((\nu\Sigma_f)_{1,\mu}\phi_1 + (\nu\Sigma_f)_{2,\mu}\phi_2\right)\psi_2,
\end{aligned}$$

supplemented with a normalization condition on u_μ^{ex}. The associated adjoint problem then reads,

Find $\left(u_\mu^{*,\text{ex}} := \left(\phi_{1,\mu}^{*,\text{ex}}, \phi_{2,\mu}^{*,\text{ex}}\right), \lambda_\mu^{\text{ex}}\right) \in (H_0^1(\Omega)^2 \times \mathbb{R})$ such that λ_μ^{ex} is an eigenvalue with minimal modulus and,

$$a_\mu((\psi_1, \psi_2), \left(\phi_{1,\mu}^{*,\text{ex}}, \phi_{2,\mu}^{*,\text{ex}}\right))) = \lambda_\mu^{\text{ex}} b_\mu((\psi_1, \psi_2), \left(\phi_{1,\mu}^{*,\text{ex}}, \phi_{2,\mu}^{*,\text{ex}}\right)) \quad \text{for all } (\psi_1, \psi_2) \in H_0^1(\Omega)^2, \tag{5.3}$$

supplemented with a a normalization condition on $u_\mu^{*,\text{ex}}$.

5.2 The high-fidelity discretization

We describe in this section the high-fidelity discretization of the continuous problem introduced in the previous section, that we consider as the reference problem in our reduced basis context. Let $\mathcal{T}_{\mathcal{N}}$ be a shape-regular mesh of Ω and $\widetilde{V}_{\mathcal{N}}$ be an associated conformal finite element approximation space of dimension $\widetilde{\mathcal{N}}$. We also denote 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 to Problem (5.2) writes

Find $\left(u_\mu^{\mathcal{N}} := \left(\phi_{1,\mu}^{\mathcal{N}}, \phi_{2,\mu}^{\mathcal{N}}\right), \lambda_\mu^{\mathcal{N}}\right) \in (V_{\mathcal{N}} \times \mathbb{R})$ such that $\lambda_\mu^{\mathcal{N}}$ is an eigenvalue with minimal modulus and,

$$a_\mu(\left(\phi_{1,\mu}^{\mathcal{N}}, \phi_{2,\mu}^{\mathcal{N}}\right), \left(\psi_1^{\mathcal{N}}, \psi_2^{\mathcal{N}}\right)) = \lambda_\mu^{\mathcal{N}} b_\mu(\left(\phi_{1,\mu}^{\mathcal{N}}, \phi_{2,\mu}^{\mathcal{N}}\right), \left(\psi_1^{\mathcal{N}}, \psi_2^{\mathcal{N}}\right)), \text{ for all } (\psi_1^{\mathcal{N}}, \psi_2^{\mathcal{N}}) \in V_{\mathcal{N}} = (\widetilde{V}_{\mathcal{N}})^2, \tag{5.4}$$

where $(\phi_{1,\mu}^{\mathcal{N}}, \phi_{2,\mu}^{\mathcal{N}})$ satisfies a normalization condition. We refer to [17] for the *a priori* error analysis of Problem (5.4). Similarly, the discrete variational formulation associated to Problem (5.3) reads

Find $\left(u_\mu^{*,\mathcal{N}} := (\phi_{1,\mu}^{*,\mathcal{N}}, \phi_{2,\mu}^{*,\mathcal{N}}), \lambda_\mu^{\mathcal{N}}\right) \in (V_{\mathcal{N}} \times \mathbb{R})$ such that $\lambda_\mu^{\mathcal{N}}$ is an eigenvalue with minimal modulus and,

$$a_\mu(\left(\psi_1^{\mathcal{N}}, \psi_2^{\mathcal{N}}\right), (\phi_{1,\mu}^{*,\mathcal{N}}(\mu), \phi_{2,\mu}^{*,\mathcal{N}})) = \lambda_\mu^{\mathcal{N}} b_\mu(\left(\psi_1^{\mathcal{N}}, \psi_2^{\mathcal{N}}\right), (\phi_{1,\mu}^{*,\mathcal{N}}, \phi_{2,\mu}^{*,\mathcal{N}})), \text{ for all } (\psi_1^{\mathcal{N}}, \psi_2^{\mathcal{N}}) \in V_{\mathcal{N}} = (\widetilde{V}_{\mathcal{N}})^2, \tag{5.5}$$

where $(\phi_{1,\mu}^{*,\mathcal{N}}, \phi_{2,\mu}^{*,\mathcal{N}})$ satisfies a normalization condition.

Let us denote by $(\theta^1, \cdots, \theta^{\mathcal{N}})$ a basis of $V_{\mathcal{N}}$. Problem (5.4) reads as follows in matrix form. For all $\mu \in \mathcal{P}$, $i = 1, 2$, let $u_\mu := (u_{\mu,k})_{1 \le k \le \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$ be the coordinates of $u_\mu^{\mathcal{N}}$ in the basis $(\theta^1, \cdots, \theta^{\mathcal{N}})$ so that

$$u_\mu^{\mathcal{N}} := \sum_{k=1}^{\mathcal{N}} u_{\mu,k} \theta^k. \tag{5.6}$$

Let us define the matrices $A_\mu := \left(a_\mu(\theta^k, \theta^l)\right)_{1 \le k,l \le \mathcal{N}}$ and $B_\mu := \left(b_\mu(\theta^k, \theta^l)\right)_{1 \le k,l \le \mathcal{N}}$. The pair $(u_\mu, \lambda_\mu) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R}$ is then solution to

$$A_\mu u_\mu = \lambda_\mu B_\mu u_\mu, \tag{5.7}$$

where u_μ satisfies a normalization condition. This is the high-fidelity eigenvalue problem of the form (2.1) we consider in the following numerical tests.

Likewise, for Problem (5.5), the pair $\left(u_\mu^*, \lambda_\mu\right) \in \mathbb{R}^{\mathcal{N}} \times \mathbb{R}$ is solution to

$$(A_\mu)^T u_\mu^* = \lambda_\mu (B_\mu)^T u_\mu^* \tag{5.8}$$

together with a normalization condition on u_μ^*. Here, $u_\mu^* = (u_{\mu,k}^*)_{1 \le k \le \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$ is the vector of coordinates of the function $u_\mu^{*,\mathcal{N}}$ in the basis $(\theta^1, \ldots, \theta^{\mathcal{N}})$, i.e.

$$u_\mu^{*,\mathcal{N}} = \sum_{k=1}^{\mathcal{N}} u_{\mu,k}^* \theta^k.$$

Problem (5.8) is the adjoint high-fidelity eigenvalue problem of the form (2.2) that we consider in the following numerical tests.

5.3 Affine decomposition of the coefficients

In the following numerical tests, the domain Ω is chosen as $[0, L]^2$ for some $L > 0$. We introduce a partition $(\Omega_k)_{k=1}^K$ of the domain Ω and the parameter functions entering in the definition of Problem 5.1 are assumed to be piecewise constant on each Ω_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 $D_1, \Sigma_{11}, \Sigma_{12}, D_2, \Sigma_{21}, \Sigma_{22}, \chi_1, \chi_2, \Sigma_{f1}, \Sigma_{f2}$ on the domain Ω_k, for each $1 \leq k \leq K$. We remark that for all $\mu \in \mathcal{P}$, the matrices A_μ and B_μ write

$$A_\mu = \sum_{k=1}^{K} \sum_{p=1}^{6} f_p(\mu_k) A_{k,p} + M_{bc} \text{ and } B_\mu = \sum_{k=1}^{K} \sum_{q=1}^{4} g_q(\mu_k) B_{k,q},$$

where $f(\mu_k)$ and $g(\mu_k)$ are the vectors defined by

$$\begin{aligned} f(\mu_k) &= (D_1(\mu_k), \Sigma_{11}(\mu_k), \Sigma_{12}(\mu_k), D_2(\mu_k), \Sigma_{21}(\mu_k), \Sigma_{22}(\mu_k)) \\ g(\mu_k) &= ((\chi_1 \nu \Sigma_{f1})(\mu_k), (\chi_1 \nu \Sigma_{f2})(\mu_k), (\chi_2 \nu \Sigma_{f1})(\mu_k), (\chi_2 \nu \Sigma_{f2})(\mu_k)), \end{aligned}$$

$A_{k,p}$ and $B_{k,q}$ ($1 \leq k \leq K$, $1 \leq p \leq 6$, $1 \leq q \leq 4$) are parameter-independent $\mathcal{N} \times \mathcal{N}$ matrices, and $M_{bc} \in \mathbb{R}^{\mathcal{N} \times \mathcal{N}}$ is a parameter-independent matrix which stems from the discretization of the boundary condition. As a consequence, all these matrices can be pre-computed in order to efficiently assemble the matrices A_μ and B_μ online, and estimate the residuals $R_N(\mu)$ and $R_N^*(\mu)$ as we now explain.

Thanks to the affine decomposition of the matrices A_μ and B_μ above, the residual norm is easily computable online, as it only requires algebraic operations over vectors of the size of the (small) reduced basis, which is N. Indeed, let $(\xi_1, \ldots, \xi_N)$ be an orthonormal basis of the chosen reduced space for the scalar product $\langle \cdot, \cdot \rangle$, and let $V_N \in \mathbb{R}^{\mathcal{N} \times N}$ be the matrix containing the coordinates of the basis $(\xi_1, \ldots, \xi_N)$ in the canonical basis of $\mathbb{R}^{\mathcal{N}}$. For $1 \leq k, l \leq K, 1 \leq p, p' \leq 6$, $1 \leq q, q' \leq 4$, we define, in the offline stage, the reduced matrices of dimension $N \times N$ as follows:

$$\begin{aligned} D^N_{k,l,p,p'} &= V_N^t A_{k,p}^t \mathbb{X}^{-1} A_{l,p'} V_N \\ E^N_{k,l,p,q} &= V_N^t A_{k,p}^t \mathbb{X}^{-1} B_{l,q} V_N \\ F^N_{k,l,q,q'} &= V_N^t B_{k,q}^t \mathbb{X}^{-1} B_{l,q'} V_N \\ D^N_{bc,k,p} &= V_N^t M_{bc}^t \mathbb{X}^{-1} A_{k,p} V_N \\ E^N_{bc,k,q} &= V_N^t M_{bc}^t \mathbb{X}^{-1} B_{k,q} V_N \\ F^N_{bc} &= V_N^t M_{bc}^t \mathbb{X}^{-1} M_{bc} V_N, \end{aligned}$$

where $\mathbb{X}$ stands for the Gram matrix of size $\mathcal{N} \times \mathcal{N}$ for the considered scalar product $\langle \cdot, \cdot \rangle$, commonly called the mass matrix, and A^t denotes the transpose

of the matrix A. Then, in the online stage, for a given parameter μ, we can assemble the residual norm as

$$\|R_N(\mu)\| := \|\left(B_\mu - k_{\mu,N} A_\mu\right) u_{\mu,N}\| = \sqrt{c_{\mu,N}^t G_{\mu,N} c_{\mu,N}},$$

with

$$\begin{aligned} G_{\mu,N} =& |k_{\mu,N}|^2 \left(\sum_{k,l=1}^{K} \sum_{p,p'=1}^{6} f_p(\mu_k) f_{p'}(\mu_l) D^N_{k,l,p,p'} + \sum_{k=1}^{K} \sum_{p=1}^{6} f_p(\mu_k) \left(D^N_{bc,k,p} + (D^N_{bc,k,p})^t \right) + F^N_{bc} \right) \\ & - k_{\mu,N} \left(\sum_{k,l=1}^{K} \sum_{p=1}^{6} \sum_{q=1}^{4} f_p(\mu_k) g_q(\mu_l) \left(E^N_{k,l,p,q} + (E^N_{k,l,p,q})^t \right) \right) + \sum_{k=1}^{K} \sum_{q=1}^{4} g_q(\mu_k) \left(E^N_{bc,k,q} + (E^N_{bc,k,q})^t \right) \\ & + \sum_{k,l=1}^{K} \sum_{q,q'=1}^{4} g_q(\mu_k) g_{q'}(\mu_l) F^N_{k,l,q,q'}. \end{aligned}$$

A similar construction is readily possible for $\|R^*_N(\mu)\|$.

5.4 Eigenvalue solver

The eigenvalue solver, for both high-fidelity and reduced-order models, relies on the inverse power method given in Algorithm 2. In practice, this algorithm is run with relative error tolerances set to $\tau_u = 10^{-6}$ and $\tau_\lambda = 10^{-7}$.

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: $(u, \lambda) = \left(u_i, \frac{1}{k_i}\right)$

The direct high-fidelity eigenvalue problem (2.1) is solved by applying Algorithm 2 with $\mathtt{A} = A_\mu$ and $\mathtt{B} = B_\mu$. Likewise, the adjoint eigenvalue problem (2.2) is solved by applying Algorithm 2 with $\mathtt{A} = A_\mu^T$ and $\mathtt{B} = B_\mu^T$. The resolutions of the reduced eigenvalue problems are performed similarly.

5.5 Numerical tests

The aim of this section is to illustrate the numerical behaviour of the reduced basis method and the proposed *a posteriori* error estimators on two different numerical test cases. Let us introduce the notation

$$e_N^k(\mu) = |k_\mu - k_{\mu,N}|\,, \quad e_N^{k,rel}(\mu) = \frac{|k_\mu - k_{\mu,N}|}{|k_\mu|},$$

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

$$e_N^{u^*}(\mu) = \|u_\mu^* - u_{\mu,N}^*\|_{\ell^2},$$

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 (5.6). 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.

5.5.1 Test case 1: 2D two-group toy example

The reduced basis method is first run on 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 2 shows the mesh used for the test case as well as the decomposition of Ω into four subdomains. Here, we set $B_\mu = I$ for all $\mu \in \mathcal{P}$.

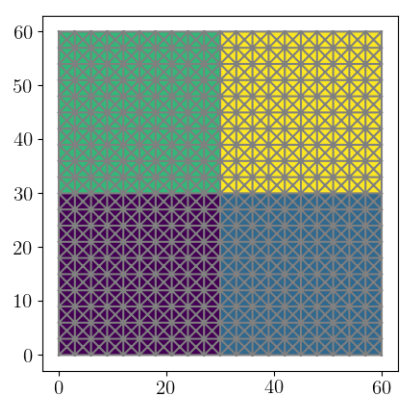

Figure 2: Domain of calculation for the two-group toy example with its associated mesh

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

- $\Sigma_{s,0,ij}$: uniform law on $[0, 0.15]$, $1 \leq i, j \leq 2$;
- Σ_{t1} and Σ_{t2}: uniform law on $[2(\Sigma_{s,0,12} + \Sigma_{s,0,21}), 0.7]$;
- $D_i = \dfrac{1}{3\Sigma_{ti}}$, $i = 1, 2$;
- $\chi_i \nu \Sigma_{fj} = \delta_{ij}$, $1 \leq i, j \leq 2$.

The coefficients are chosen so that the coercivity of Problems (5.2) and (5.3) 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 $\#\mathcal{P}_{\text{train}} = 300$, $\#\mathcal{P}_{\text{test}} = 50$ and $\mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset$. In the offline stage, the greedy algorithm is performed using the *a posteriori* estimator

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

defined in (3.13) for all $\mu \in \mathcal{P}$ (in other words, we choose here $\bar{C}_N^k(\mu) = 1$ for all μ, following the notation (4.1)).

The left part of Figure 3 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 3 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.

It is also interesting to look at the behavior of the implemented *a posteriori* error estimators. The relation between the error $|k_N(\mu) - k(\mu)|$ 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 (3.14). The value of $C_N^k(\mu)$ on the test set $\mathcal{P}_{\text{test}}$ is presented in Figure 4. In that particular case, we fall into the framework developed in Section 4.1.3. Indeed, when we compute the perturbation magnitude ε_μ as

$$\varepsilon_\mu = \frac{\left\| \frac{A_\mu - A_\mu^T}{2} \right\|}{\left\| \frac{A_\mu + A_\mu^T}{2} \right\|}, \tag{5.9}$$

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 (4.5) 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

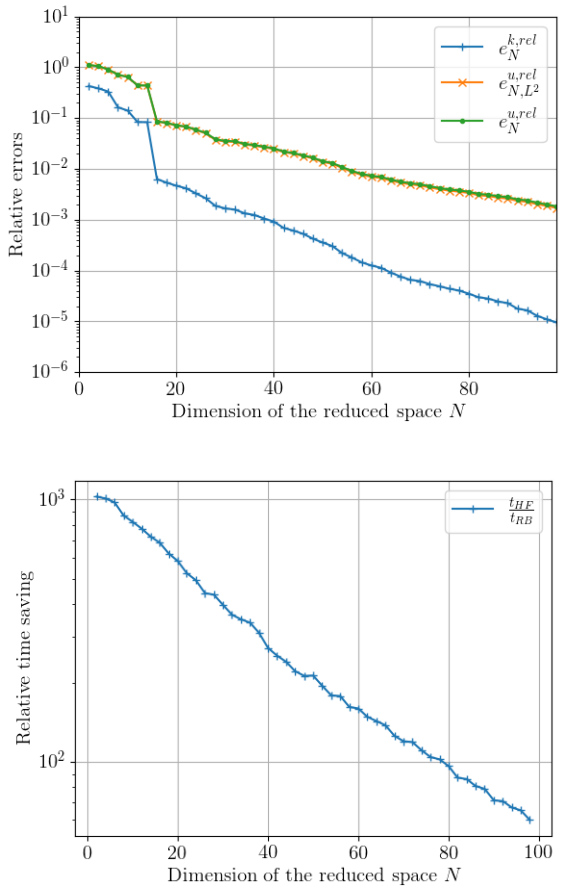

Figure 3: (Left) Mean relative errors over $\mathcal{P}_{\text{test}}$; (Right) Relative time saving factor $\dfrac{t_{\text{HF}}}{t_{\text{RB}}}$ as a function of the dimension of the reduced space N.

Figure 4. Actually, in the cases where the prefactors differ a lot, we observe that condition (4.12) in Proposition (4.8) is not satisfied, which explains why the perturbative expansions may not be sharp.

Figure 5 compares the behavior of the simple *a posteriori* error estimators $\|R_N(\mu)\|$, $\|R_N^*(\mu)\|$ and $\eta_N^k(\mu)$ defined in (3.13), 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 from $N \geq 20$ which confirms 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

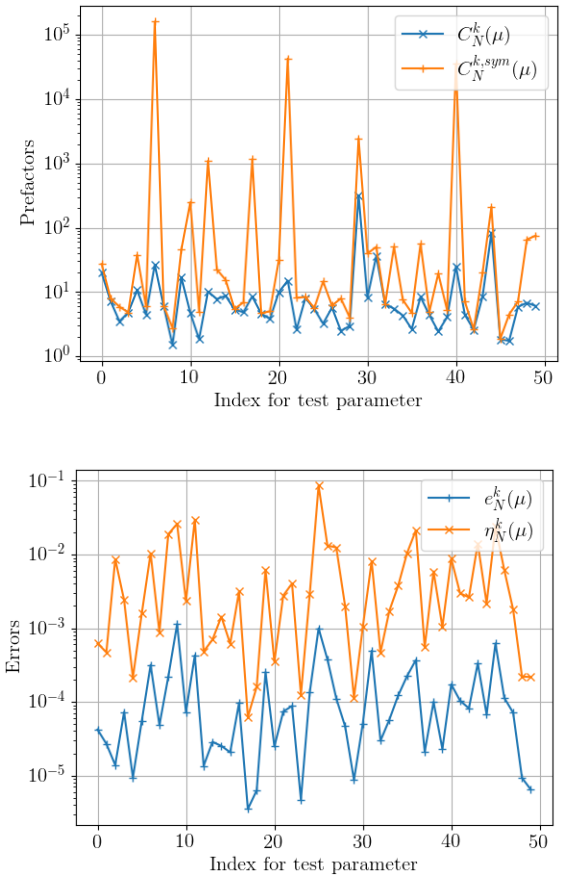

Figure 4: (Left) Variations of the prefactor $C_N^k(\mu)$ and $C_N^{k,\mathrm{sym}}(\mu)$ over $\mathcal{P}_{\mathrm{test}}$ for $N = 100$.
(Right) Parametric variations of the real eigenvalue error $e_N^k(\mu)$ (in blue) and the associated *a posteriori* error estimator $\eta_N^k(\mu)$ (in orange) over $\mathcal{P}_{\mathrm{test}}$, for $N = 100$.

4 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 illustrate the importance of introducing prefactors $\bar{C}_N^k(\mu)$, $\bar{C}_N^u(\mu)$ and $\bar{C}_N^{u^*}(\mu)$ to estimate the true errors, see (4.1). This is important in particular in order to stop the greedy procedure when the real error is below a given threshold. This will be discussed below.

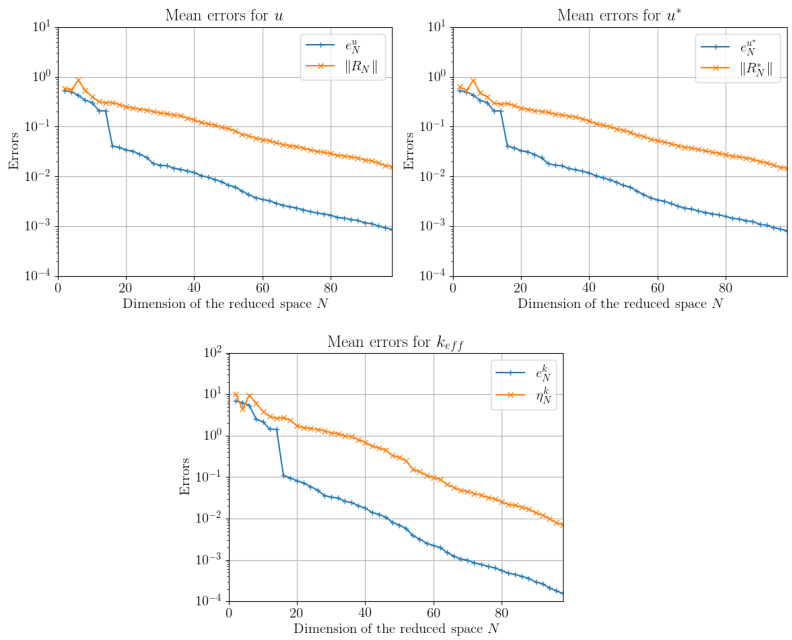

Figure 5: Mean values for errors and associated *a posteriori* error estimators over $\mathcal{P}_{\text{test}}$. (Left) e_N^u and $\|R_N\|$; (Middle) $e_N^{u^*}$ and $\|R_N^*\|$; (Right) e_N^k and η_N^k.

5.5.2 Test case 2: 2D two-group "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 6 shows, it is made up 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 one being 21.504 cm long. It is discretized into $\mathcal{N} = 2602$ degrees of freedom. Here, there holds $B_\mu \neq I$, and the Dirichlet boundary condition in Problem (5.1) is replaced by a Robin condition called void boundary condition which writes

$$D_i(r,\mu)\nabla\phi_i(r,\mu).\vec{n} + \frac{1}{2}\phi_i(r,\mu) = 0 \quad \text{on } \partial\Omega, \quad 1 \leq i \leq 2,$$

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

In this test case, the parameter μ stands for five parameters which determine all the physical parameters entering (5.1). More precisely, by recalling the partition $(\Omega_k)_{k=1}^K$ of the domain Ω, 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 Ω_k:

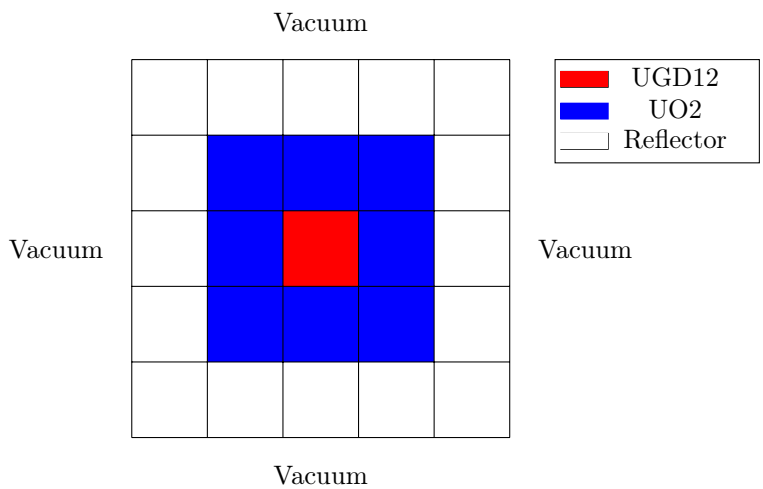

Figure 6: Median cross-sectional view of the *MiniCore* ($z = 234.36$ cm)

- the nature of the material in Ω_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

$$\#\mathcal{P}_{\text{train}} = 1000, \qquad \#\mathcal{P}_{\text{test}} = 50, \text{ and } \qquad \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset.$$

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 Algorithm 1). Then, the greedy procedure is continued using the *a posteriori* estimator $\|R_N\| + \|R_N^*\|$, as the quantity of interest here is the two-group flux $(\phi_1^{\mathcal{N}}, \phi_2^{\mathcal{N}})$ as well as its adjoint $(\phi_1^{*,\mathcal{N}}, \phi_2^{*,\mathcal{N}})$.

The left part of Figure 7 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 8 shows, for a certain $\mu_0 \in \mathcal{P}$ and for $N = 100$, the maximum error on the associated first-group flux does not exceed 3.2×10^{-4}; as for the second group, the right part of Figure 8 shows that the flux error is locally gathered in an area of low flux, quite far from the hot spot.

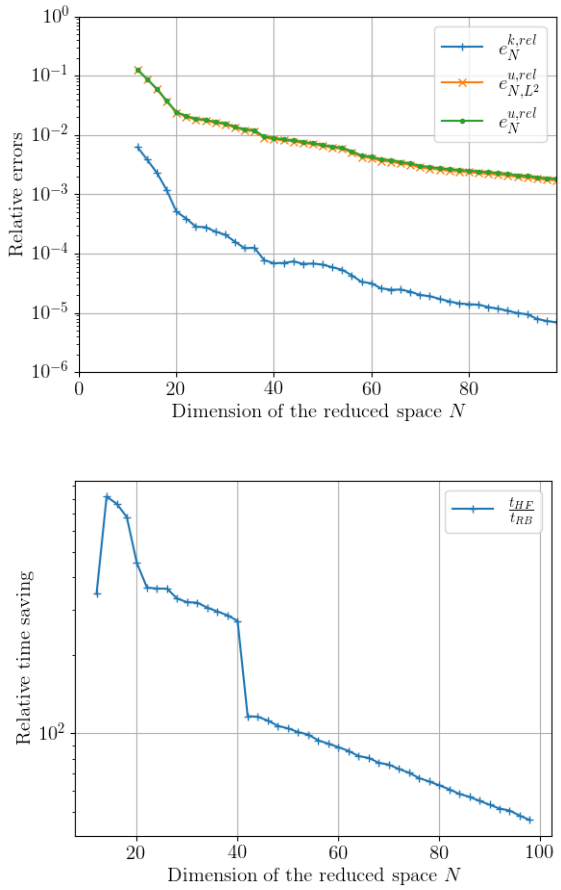

Figure 7: (Left) Mean relative errors over $\mathcal{P}_{\text{test}}$; (Right) Relative time saving factor $\dfrac{t_{\text{HF}}}{t_{\text{RB}}}$ as a function of the dimension of the reduced space N.

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 7 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.

We now study the certification of the method performed by the *a posteriori* estimator. Figure 9 shows that, although the residuals display similar values as those for the real eigenvector errors, for the eigenvalue, the order of magnitude

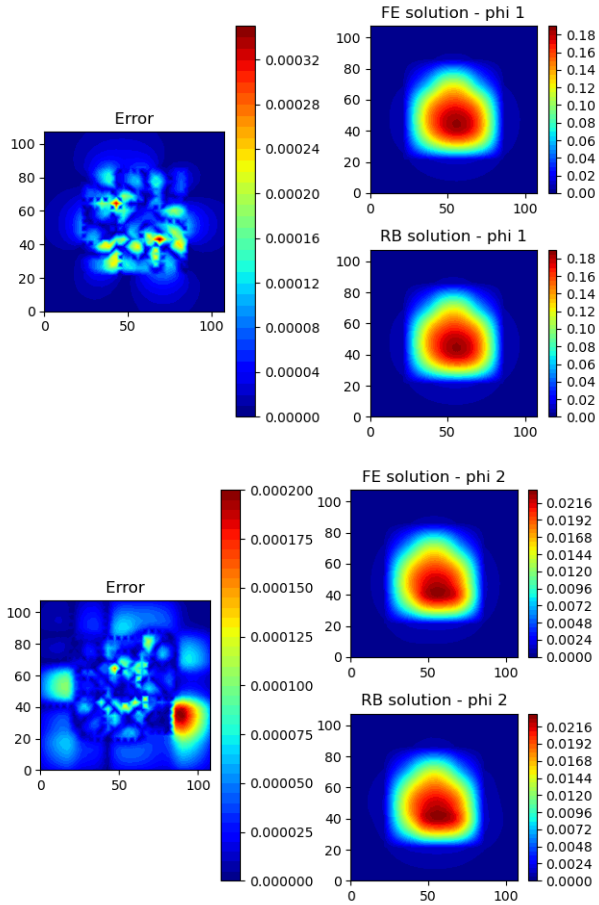

Figure 8: (Left) Plots of the first energy group of high-fidelity (upper right) and reduced (lower right) solutions u_{μ_0} and $u_{\mu_0,N}$, and their error (left) $|u_{\mu_0} - u_{\mu_0,N}|$, for $N = 100$ and for $\mu_0 \in \mathcal{P}_{\text{test}}$; (Right) Plots of the second energy group of high-fidelity (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 $\mu_0 \in \mathcal{P}_{\text{test}}$

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 by Figure 10, 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 4.2 is computed. The right plot of Figure 10 shows that the efficiency $\mathcal{E}_N^k$ defined in (4.13) levels off for $N = 100$ at the order of magnitude of 10^{-1}, and does not depend too much

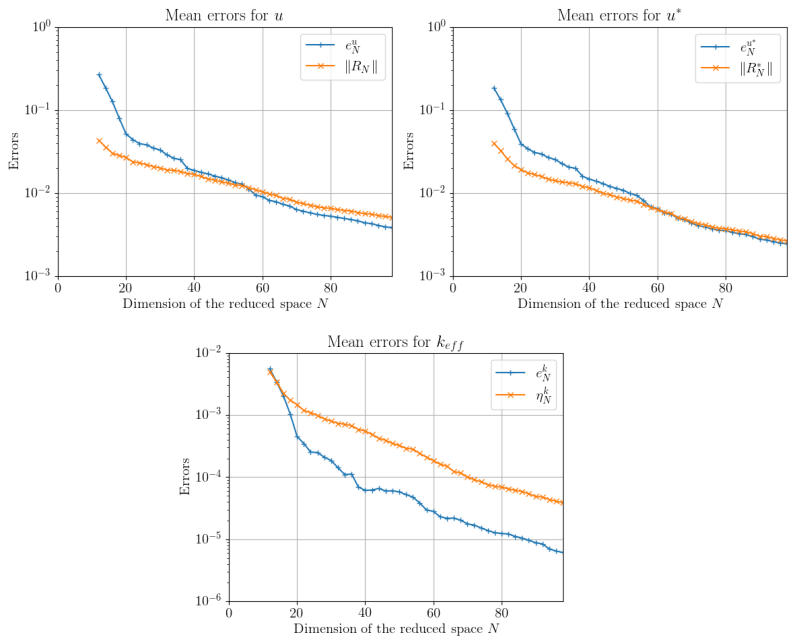

Figure 9: Mean values for errors and associated *a posteriori* error estimators over $\mathcal{P}_{\text{test}}$.
(Left) e_N^u and $\|R_N\|$; (Middle) $e_N^{u^*}$ and $\|R_N^*\|$; (Right) e_N^k and η_N^k.

on the parameter μ. Therefore, we propose to apply the procedure outlined in Section 4.2 to build a posteriori error estimators of the form (4.16), with constants $\overline{C}_N^k$, $\overline{C}_N^u$ and $\overline{C}_N^{u^*}$ approximated by (4.15). This requires to choose a set $\mathcal{P}_{\text{pref}}$, that we randomly chose in $\mathcal{P}$ such that

$$\mathcal{P}_{\text{pref}} \subset \mathcal{P}, \qquad \#\mathcal{P}_{\text{pref}} = 10, \text{ and } \qquad \mathcal{P}_{\text{pref}} \cap \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \emptyset.$$

As a result of this procedure, Figure 11 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 tends to be an optimal stopping indicator.

Finally, we gather in Table 1 the measured computational times for several quantities of interest and main stages in 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 there holds

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

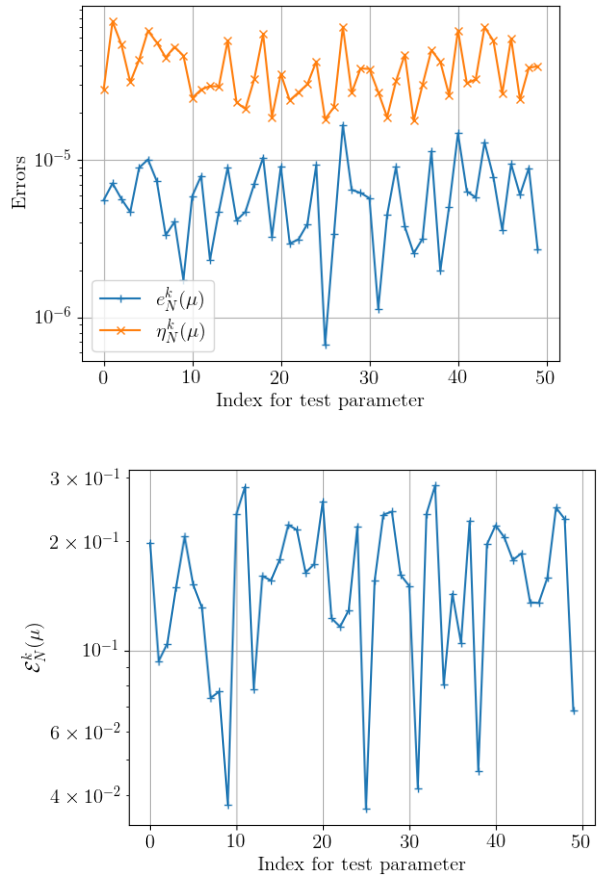

Figure 10: (Left) Parametric variations of the real eigenvalue error $e_N^k(\mu)$ (in blue) and its associated *a posteriori* error estimator $\eta_N^k(\mu)$ (in orange) over $\mathcal{P}_{\text{test}}$, for $N = 100$; (Right) Parametric variations of the practical efficiency $\mathcal{E}_N^k(\mu)$ over $\mathcal{P}_{\text{test}}$, for $N = 100$.

that is

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

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

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme

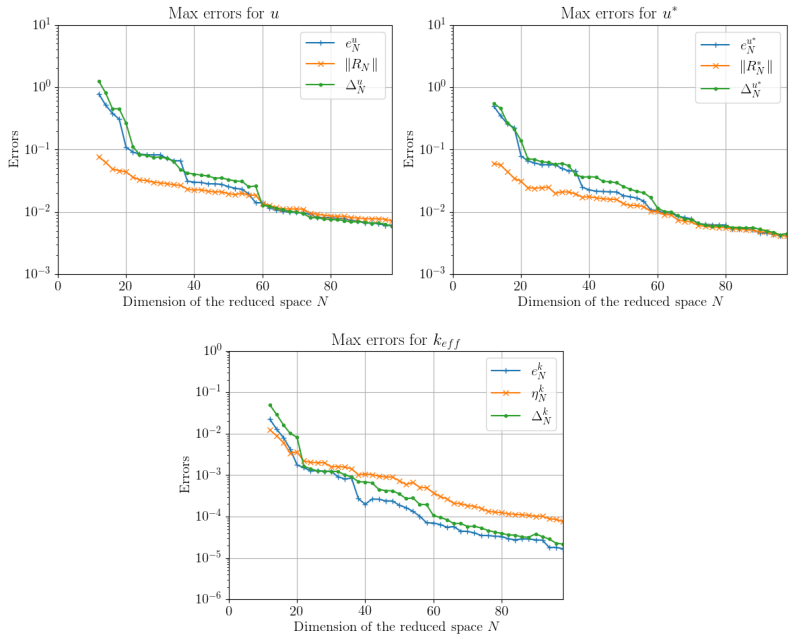

Figure 11: Maximum values for errors and associated *a posteriori* error estimators over $\mathcal{P}_{\text{test}}$. Upper left: e_N^u, $\|R_N\|$ and Δ_N^u; upper right: $e_N^{u^*}$, $\|R_N^*\|$ and $\Delta_N^{u^*}$; lower: e_N^k, η_N^k and Δ_N^k.

	Mean computational time
Offline stage	$\approx$ 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 1: Mean computational times for the Efficient Greedy reduced basis method applied to the 2D two-group *minicore* in Python, for $N = 100$

(grant agreement EMC2 No 810367). GD was supported by the French 'Investissements d'Avenir' program, project Agence Nationale de la Recherche (ISITE-BFC) (contract ANR-15-IDEX-0003). GD was also supported by the Ecole des Ponts-ParisTech.

References

[1] G. Allaire, X. Blanc, B. Desprès, and F. Golse, *Transport et diffusion*, Ecole Polytechnique, 2019.

[2] I. Babuška and J. Osborn, *Eigenvalue problems*, (1991).

[3] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numer., 19 (2010), pp. 1–120.

[4] D. Boffi, A. Halim, and G. Priyadarshi, *Reduced basis approximation of parametric eigenvalue problems in presence of clusters and intersections*, (2023).

[5] 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, (2010).

[6] A. G. Buchan, C. C. Pain, F. Fang, and I. M. 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.

[7] M. D. DeChaine and M. A. Feltus, *Nuclear fuel management optimization using genetic algorithms*, Nuclear Technology, 111 (1995), pp. 109–114.

[8] 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.

[9] J. J. Duderstadt and L. J. Hamilton, *Nuclear reactor analysis*, John Wiley & Sons, Inc., 1976.

[10] P. Edel, *Reduced basis method for parameter-dependent linear equations. Application to time-harmonic problems in electromagnetism and in aeroacoustics*, PhD thesis, Sorbonne Université, Nov. 2022.

[11] 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.

[12] J. Gedicke and C. Carstensen, *A posteriori error estimators for nonsymmetric eigenvalue problems*, (2009).

[13] 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.

[14] S. Giani, L. Grubišić, and A. Międlar, *Robust error estimates for approximations of non-self-adjoint eigenvalue problems*, Numerische Mathematik, (2016).

[15] J. S. Hesthaven, G. Rozza, and B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, SpringerBriefs in Mathematics, Springer, 2016.

[16] 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.

[17] E. Jamelot and F. Madiot, *Numerical analysis of the neutron multi-group SP_N equations*, Comptes Rendus. Mathématique, 359 (2021), pp. 533–545.

[18] T. Kato, *Perturbation theory for linear operators*, vol. 132, Springer Science & Business Media, 2013.

[19] S. Lorenzi, *An adjoint proper orthogonal decomposition method for a neutronics reduced order model*, Annals of Nuclear Energy, 114 (2018), pp. 245–258.

[20] 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.

[21] C. Prud'Homme, D. V. Rovas, K. Veroy, L. Machiels, Y. Maday, A. T. Patera, and G. Turinici, *Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods*, J. Fluids Eng., 124 (2002), pp. 70–80.

[22] A. Quarteroni, A. Manzoni, and F. Negri, *Reduced Basis Methods for Partial Differential Equations: An Introduction*, Springer, Aug. 2015.

[23] A. Sartori, A. Cammi, L. Luzzi, M. E. Ricotti, and G. Rozza, *Reduced order methods: applications to nuclear reactor core spatial dynamics-15566*, in ICAPP 2015 Proceedings, 2015.

[24] P. J. Turinsky, *Nuclear fuel management optimization: A work in progress*, Nuclear technology, 151 (2005), pp. 3–8.