Reduced basis for the Diffusion equation

Nicolas PODVIN

04/12/19

Contents

Introduction

This report was realized during a 5 month internship at the "Commissariat à l'Énergie Atomique et aux Énergies Alternatives" (CEA). The main purpose of this internship is to study reduced basis approaches for the parametrized (multigroup) Neutron diffusion equations.

In certain applications such as the optimization of the configuration of the assemblies in a nuclear core reactor, it is needed to perform numerical simulations over many different configurations. This may become prohibitive in terms of computational costs.

The reduced basis approach [7] consists in reducing the Finite Element model by projecting the approximation space of dimension N (typically large) onto a considerably smaller and well-chosen space, therefore reducing the computational cost while controlling the accuracy.

Applications of the Reduced Basis method are numerous in the field of neutronics : see [4] for the stationary case and [8] in the transient case. Those results illustrate the efficiency of a POD space, where the projection was made in a Galerkin framework. Another approach is presented in [6], where the adjoint problem is included in the projection, creating a specific Petrov Galerkin framework.

We focus in this work on an eigenproblem formulation of the diffusion equation. We start in Part I with the one group diffusion equation, which can be reformulated as a coercive generalized eigenvalue problem. Chapter 1 presents the well-posedness of the problem and the numerical method employed for the discretization. Chapter 2 introduces the different reduced basis approaches employed. Chapter 3 is devoted to the numerical results.

The two group diffusion equation is considered in Part II. Though the well-posedness is still assured, the non-symmetric framework brings some challenges. Similarly to Chapter 1, Chapter 4 presents the well-posedness of the problem and the numerical method employed for the discretization. Chapter 5 details the various reduced basis approaches regarding the Proper Orthogonal Decomposition and the definition of the reduced problem. The numerical results are presented in Chapter 6.

Part I

One Group diffusion equation

Chapter 1

Setting of the model

1.1 Continuous problem

Let Ω be a *bounded, connected* and *open* subset of $\mathbb{R}^d$ with $d = \{1, 2, 3\}$. Let $(\Omega_k)_{k\in\{1,...,Q_k\}}$ be a partition of the domain Ω into open subdomains with piecewise smooth boundaries, for a given integer $k \in \{1, ..., Q_k\}$.

We look at the *parametric-PDE* that we proposed to solve via different approaches all along this paper, in its most simplistic form : the *one group discrete Neutron Diffusion Equation.* Let μ denotes the parameter we will consider, which lives in a space $P \subset \mathbb{R}^{Q_k}$ for the purpose of this study. Our equation will then be,

$$\begin{aligned}&\text{Find } (\phi_\mu, \lambda_\mu) \text{ such that} &(1.1)\\ &-\mathrm{div}\,(\ D(\mu)\ \nabla\phi_\mu\ ) + (\Sigma_t(\mu) - \Sigma_{s0}(\mu))\ \phi_\mu \ = \ \lambda_\mu\ \nu\Sigma_f(\mu)\ \phi_\mu \ \text{ in } \Omega, \quad \phi_\mu = 0 \text{ on } \partial\Omega,\end{aligned}$$

where (see [6]),

- $D(\mu)$: Diffusion coefficient,
- $\Sigma_t(\mu)$: Total macroscopic cross section of the considered neutron population,
- $\Sigma_{s0}(\mu)$: Scattering macroscopic cross section of the neutron population, considering the diffusion to be isotropic (anisotropy of order 0),
- $\Sigma_f(\mu)$: Fission macroscopic cross section,
- ν : Number of neutrons emitted by a fission generated by a neutron.

We define the space $\mathcal{PW}^{1,\infty}(\Omega)$ (see [6]) by

$$\mathcal{PW}^{1,\infty}(\Omega_i) \ = \ \{D \in L^\infty(\Omega), D_{|\Omega_i} \in W^{1,\infty}(\Omega_i), 1 \le i \le Q_k\},$$

and make the following assumptions on the coefficients of (1.1).

Assumptions 1. *Assume that,*

- Ω *is open, connected and bounded,*

- *Each* Ω_i *for* $i \in \{1, ..., Q_k\}$ *is open and its boundary is piecewise smooth,*
- $\Sigma_t(\mu) \in \mathcal{PW}^{1,\infty}(\Omega)$,
- $\Sigma_{s0}(\mu) \in \mathcal{PW}^{1,\infty}(\Omega)$,
- $\nu(\mu)\Sigma_f(\mu) \in \mathcal{PW}^{1,\infty}(\Omega)$,
- $\exists (\Sigma_t)_*, (\Sigma_t)^* > 0, \forall i \in \{1, ..., Q_k\}, \quad \forall \mu \in P, \quad (\Sigma_t)_* \leq (\Sigma_t)_{|\Omega_i}(\mu) \leq (\Sigma_t)^*,$
- $\exists (\Sigma_{s0})_*, (\Sigma_{s0})^* > 0, \forall i \in \{1, ..., Q_k\}, \quad \forall \mu \in P, \quad (\Sigma_{s0})_* \leq (\Sigma_{s0})_{|\Omega_i}(\mu) \leq (\Sigma_{s0})^*,$
- $\exists (\nu\Sigma_f)_*, (\nu\Sigma_f)^* > 0, \forall i \in \{1, ..., Q_k\}, \quad \forall \mu \in P, \quad (\nu\Sigma_f)_* \leq (\nu\Sigma_f)_{|\Omega_i}(\mu) \leq (\nu\Sigma_f)^*.$

The variational formulation of problem (1.1) writes,

$$\begin{aligned} &\text{Find } (\phi_\mu, \lambda_\mu) \in V \times \mathbb{R} \text{ such that for all } v \in V, \\ &a_\mu(\phi_\mu, v) \;=\; \lambda_\mu \; b_\mu(\phi_\mu, v). \end{aligned} \tag{1.2}$$

Physically, the unknown $\phi(\underline{x})$ represents the neutron density at point $\underline{x} \in \Omega$. We are interested in finding the eigenvalue λ_{eff} of lowest magnitude and ϕ is therefore the associated eigenvector. We then define the *multiplication factor* $k_{eff} := \frac{1}{\lambda_{eff}}$. It can be interpreted as the ratio of the number of neutrons emitted by fission over the number of neutrons lost by absorption, scattering or leakage.

For simplicity, we shall henceforth denote $\lambda_{eff,\mu}$ as λ_μ and similarly, $\phi_{eff,\mu}$ as ϕ_μ to lighten the notations.

The equation (1.1) is parametric in the sense that all coefficients $\Sigma_t, \Sigma_{s0}, \nu$ and Σ_f all depend on a parameter μ, and let us denote by P the space in which μ lives.

By defining henceforth $H = L^2(\Omega)$ and $V = H^1_0(\Omega)$, let us define the continuous forms $a_\mu : V \times V \to \mathbb{R}$ and $b_\mu : V \times V \to \mathbb{R}$ as,

$$\begin{aligned} \forall u, v \in V \quad a_\mu(u, v) &= \int_\Omega \frac{1}{3\Sigma_t(\mu)} \nabla u \cdot \nabla v + (\Sigma_t(\mu) - \Sigma_{s0}(\mu))uv, \\ \forall u, v \in V \quad b_\mu(u, v) &= \int_\Omega (\nu(\mu)\Sigma_f(\mu))uv. \end{aligned}$$

Given Assumptions 1, we can verify that a_μ is *symmetric, continuous* and uniformly *coercive* w.r.t the V-norm, where the V-norm is taken as the $H^1(\Omega)$ norm and not as the $H^1_0(\Omega)$ semi-norm,

$$\begin{aligned} \exists \alpha > 0, \ \forall \mu \in P, \ \forall u \in V, \ \ a_\mu(u, u) &\geq \alpha \left\| u \right\|_V^2, \\ \exists C > 0, \ \forall \mu \in P, \ \forall u, v \in V, \ \ a_\mu(u, v) &\leq C \left\| u \right\|_V \left\| v \right\|_V. \end{aligned}$$

Thanks to the symmetric coercive framework, the solution $(\phi_\mu, \lambda_\mu) \in V \times \mathbb{R}$ of the weak formulation.

Theorem 1. *Let Assumptions (1) holds. Then, problem (1.2) is well posed (up to a multiplication factor for ϕ_μ) and is equivalent to (1.1)*

$$\forall v \in V \quad a_\mu(\phi_\mu, v) \; = \; \lambda_\mu \; b_\mu(\phi_\mu, v).$$

In addition, ϕ_μ is a positive eigenvector, and λ_μ is a simple eigenvalue.

Proof. Let $\mu \in P$. Let us define the following source problem for any $s \in H$,

$$\begin{aligned} &\text{Find } \phi_\mu \in V \text{ such that for all } v \in V, \\ a(\phi_\mu, v) \; &= \; \langle \nu\Sigma_f \; s, v\rangle_H \, . \end{aligned} \tag{1.3}$$

Applying the Lax-Milgram theorem, we know that there exists a unique $\phi_\mu \in V$ satisfying (1.3).

Let us define

$$\begin{aligned} A : H \longrightarrow V \quad &, \quad s \mapsto \phi_\mu \text{ solution of (1.3)}, \\ \mathcal{A} : H \longrightarrow H \quad &, \quad s \mapsto As. \end{aligned}$$

Step 1 : Compactness of $\mathcal{A}$

Given that V is embedded in H with a compact embedding (see [3]),

$$\exists C_i > 0, \|u\|_V \leq C_i \|u\|_H \, .$$

By denoting $i_{V\to H}$ the relevant injection, we have

$$\mathcal{A} = i_{V\to H} \circ A.$$

Therefore, to prove the compactness of $\mathcal{A}$, it is sufficient to prove the continuity of A. We then have, for $s \in H$ and by denoting ϕ_μ,

$$\begin{aligned} \|\phi_\mu\|_V^2 \; &\leq \; \alpha^{-1} \; a_\mu(\phi_\mu, \phi_\mu) \\ &= \; \alpha^{-1} \langle \nu\Sigma_f \; s \; , \; \phi_\mu \rangle_H \\ &\leq \; \alpha^{-1}(\nu\Sigma_f)^* \|s\|_H \|\phi_\mu\|_H \; \text{ by Assumptions (1)}, \\ &\leq \; \alpha^{-1}(\nu\Sigma_f)^* \|\phi_\mu\|_V \|s\|_H \, . \end{aligned} \tag{1.4}$$

Therefore, we have for all $s \in H$,

$$\|As\|_V \leq \alpha^{-1}(\nu\Sigma_f)^* \|s\|_H \, ,$$

which proves the continuity of A, and thus the compactness of $\mathcal{A}$.

Step 2 : Existence, Uniqueness and Positivity

We conclude using Krein-Rutman theorem (see [1, Theorem 6.2.14])

Theorem 2. *Let X be a Banach space , $C \subset X$ a closed cone with interior $Int(C) \neq \emptyset$ and assume that $C \cap (-C) = \{0\}$. Let $T : X \longrightarrow X$ be a compact operator that is positive, i.e. $T(C) \subset C$. Therefore, there exists $u \in Int(C)$ and $\lambda > 0$ such that $Tu = \lambda u$. In addition, λ is simple, the largest -in terms of magnitude- of the eigenvalues of T and the only eigenvalue associated to an element of $Int(C)$.*

Taking the space $H = L^2(\Omega)$ and the cone of non negative functions, the only hypothesis left to verify is the strict positivity of our operator, which can be asserted by the *weak maximum principle*, see [3, Chapter 9].

□

The smallest real eigenvalue λ_μ satisfies the Rayleigh properties.

Theorem 3. *Let $(\phi_\mu, \lambda_\mu) \in V \times \mathbb{R}$ be the solution of (1.2). Then,*

$$\begin{aligned} \lambda_\mu &= \underset{w \in V \backslash \{0\}}{min} \frac{a_\mu(w,w)}{b_\mu(w,w)}, \\ \phi_\mu &= \underset{w \in V \backslash \{0\}}{argmin} \frac{a_\mu(w,w)}{b_\mu(w,w)}. \end{aligned}$$

Proof. First, let us remark that b_μ define a scalar product and that H provided with this scalar product is still hilbertian. This comes form the hypothesis on $(\nu\Sigma_f)$ in (1) : for all $v \in H$ and for all $\mu \in P$,

$$(\nu\Sigma_f)_* \|v\|_H^2 \leq b_\mu(v,v) \leq (\nu\Sigma_f)^* \|v\|_H^2 .$$

Let $(\phi_k)_{k \in \mathbb{N}^*}$ be a hilbertian basis adapted to $(H, b_\mu(.,.))$. We have then, by defining $\lambda_k := \beta_k^{-1}$ that $\left(\frac{\phi_k}{\sqrt{\lambda_k}}\right)_{k \in \mathbb{N}^*}$ forms a hilbertian basis of $(V, a_\mu(.,.))$.
Indeed, it is clear that each ϕ_k belongs to V as they are eigenvectors of the problem. In addition, we have for all $(k,l) \in (\mathbb{N}^*)^2$

$$a_\mu\left(\frac{\phi_k}{\sqrt{\lambda_k}}, \frac{\phi_l}{\sqrt{\lambda_l}}\right) = \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_l}} b_\mu(\phi_k, \phi_l).$$

As $b(\phi_k, \phi_l) = \delta_{kl}$, this shows that $\left(\frac{\phi_k}{\sqrt{\lambda_k}}\right)_{k \in \mathbb{N}^*}$ form a hilbertian basis of V.

These observations on the relationships between a_μ-hilbertian basis and b_μ-hilbertian basis will help in the expression of the Rayleigh quotient. Let us now characterize both H and V with these.

We have from Parseval's equalities (and inequalities)

$$\begin{aligned} H &= \{v = \sum_{k \in \mathbb{N}^*} b_\mu(\phi_k, v)\phi_k \ / \ \sum_{k \in \mathbb{N}^*} b_\mu(\phi_k, v)^2 < \infty\}, \\ V &= \{v = \sum_{k \in \mathbb{N}^*} b_\mu(\phi_k, v)\phi_k \ / \ \sum_{k \in \mathbb{N}^*} \lambda_k b_\mu(\phi_k, v)^2 < \infty\}. \end{aligned}$$

Let $v \in V\backslash\{0\}$, and set $\alpha_k = \sqrt{b_\mu(\phi_k, v)}$ for all $k \in \mathbb{N}^*$. We have

$$R(v) = \frac{a_\mu(v,v)}{b_\mu(v,v)} = \frac{\sum_{k\in\mathbb{N}^*} \lambda_k \alpha_k^2}{\sum_{k\in\mathbb{N}^*} \alpha_k^2}.$$

If we define $c_k := \dfrac{\alpha_k^2}{\sum_{k\in\mathbb{N}^*} \alpha_k^2}$, we have

$$R(v) = \sum_{k\in\mathbb{N}^*} \lambda_k c_k \ , \ \sum_{k\in\mathbb{N}^*} c_k = 1.$$

It is then clear that λ_1 is the minimum of $R(.)$, obtained for the choice ϕ_1, as the sequence of λ_k is increasing.

□

This property is only available in the symmetric context : if more than two groups are considered for the Neutron Diffusion equation, this Rayleigh property does not hold.

In conclusion, the partial differential equation considered is well-posed. We shall now present its discretization to approximate the solution.

1.2 Discretization of the problem

In this study, we shall solve this equation using a *Finite Element approximation* (see [4]). Let $\mathcal{T}_h$ be a triangulation of the domain $\Omega =: \cup_{i=1}^{N_h-1} K_i$, we define for $k \in \mathbb{N}$ the space $\mathcal{P}_{k,h}$ by

$$\mathcal{P}_{k,h} = \{v_h \in \mathcal{C}^0(\overline{\Omega}) \cap V \ / \ \forall i \in \{1, ..., N_h - 1\} \ \ v_{h_{|K_i}} \in \mathbb{P}_k\},$$

N_h therefore denotes the dimension of the space. Let $(\chi_k)_{k\in\{1,...,N_h\}}$ be a basis of $\mathcal{P}_{k,h}$. We can then express any element $v_h \in V_h$ as

$$v_h = \sum_{k=1}^{N_h} v_{h,k} \ \chi_k.$$

The discrete problem associated to (1.2) writes,

$$\begin{aligned} &\text{Find } (\phi_{\mu,h}, \lambda_{\mu,h}) \in V_h \times \mathbb{R} \text{ such that for all } v_h \in V_h, \\ &a_\mu(\phi_{\mu,h}, v_h) \ = \ \lambda_{\mu,h} \ b_\mu(\phi_{\mu,h}, v_h). \end{aligned} \tag{1.5}$$

Let us denote the vector of the representation of an element of V_h in the basis (χ_k), basis that henceforth shall be referred to as *Finite Element basis*, as $\underline{V_h} = (v_{h,k})_k \in \mathbb{R}^{N_h}$. Let us define as well matrices sometime referred as *Stiffness* and *Mass* matrices

$$\begin{aligned} A_{\mu,h}[i,j] &= a_\mu(\chi_j, \chi_i), \\ B_{\mu,h}[i,j] &= b_\mu(\chi_j, \chi_i). \end{aligned}$$

We can then reformulate equation (1.5) as,

$$\begin{aligned}&\text{Find the couple } (\underline{\phi_h}, \lambda_h) \in \mathbb{R}^{N_h} \times \mathbb{R}^+ \text{ such that}\\ &A_{\mu,h}\ \underline{\phi_h} = \lambda B_{\mu,h}\ \underline{\phi_h}.\end{aligned} \tag{1.6}$$

As $A_{\mu,h}$ and $B_{\mu,h}$ are both *symmetric definite positive matrices*, this problem is well posed as a standard symmetric generalized eigenvalue problem. Moreover, thanks to $V_h \subset V$ and Theorem 3, we have

$$\lambda_h \ \leq\ \lambda_{\mu,h}.$$

Thanks to the Rayleigh formulation, we can refine our estimate of λ_h.

Theorem 4. *Let Assumption (1) holds. Let $\Pi_h : V \longrightarrow V_h$ denote the a_μ-orthogonal projector on V_h. Let ϕ_μ denote the solution of (1.2). If $\Pi_h\phi_\mu$ is not null, then*

$$\lambda_{\mu,h} \ \leq\ \lambda_\mu \ \frac{b_\mu(\phi_\mu,\phi_\mu)}{b_\mu(\Pi_h\phi_\mu,\Pi_h\phi_\mu)} \leq \lambda_\mu \ \frac{(\nu\Sigma_f)^* \left\|\phi_\mu\right\|^2}{\nu(\Sigma_f)_* \left\|\Pi_h\phi_\mu\right\|_H^2}.$$

Proof. We can adapt the proof of Theorem 3 on a finite-dimensional case to see that

$$\begin{aligned}\lambda_{\mu,h} &= \min_{w_h\in V_h\setminus\{0\}} \frac{a_\mu(w_h,w_h)}{b_\mu(w_h,w_h)},\\ \phi_{\mu,h} &= \operatorname*{argmin}_{w_h\in V_h\setminus\{0\}} \frac{a_\mu(w_h,w_h)}{b_\mu(w_h,w_h)}.\end{aligned}$$

Therefore, if $\Pi_h\phi_\mu$ is not null, then

$$\lambda_{\mu,h} \leq \frac{a_\mu(\Pi_h\phi_\mu,\Pi_h\phi_\mu)}{b_\mu(\Pi_h\phi_\mu,\Pi_h\phi_\mu)}.$$

By definition of the a_μ-orthogonal projector Π_h, we have for all $v \in V$

$$a_\mu(v,v) = a_\mu(\Pi_h v,\Pi_h v) + a_\mu(v-\Pi_h v, v-\Pi_h v).$$

As a_μ is coercive w.r.t the V-norm, the second term is positive and therefore

$$a_\mu(\Pi_h v,\Pi_h v) \leq a_\mu(v,v).$$

Therefore

$$\lambda_{\mu,h} \leq \frac{a_\mu(\phi_\mu,\phi_\mu)}{b_\mu(\Pi_h\phi_\mu,\Pi_h\phi_\mu)} \ =\ \lambda_\mu \ \frac{b_\mu(\phi_\mu,\phi_\mu)}{b_\mu(\Pi_h\phi_\mu,\Pi_h\phi_\mu)}.$$

The boundedness of $(\nu\Sigma_f)$ allows us to conclude.

□

We conclude this part by presenting convergence results. Applying the results of Babuška-Osborn theory [2] recalled in [3, Section 9], we have the following a priori estimate.

Theorem 5. *Assume that $(V_h)_h$ corresponds to a Finite Element approximation set of spaces. Let (ϕ_μ, λ_μ) and $(\phi_{\mu,h}, \lambda_{\mu,h})$ be respectively solution to Problems (1.2) and (1.5). Assuming that λ_μ is simple and that $\|\phi_\mu\|_H = 1$, we have*

$$\begin{aligned} |\lambda_\mu - \lambda_{\mu,h}| &\leq C \left[\inf_{v_h \in V_h} \|\phi_\mu - v_h\|_V \right]^2, \\ \|\phi_\mu - \phi_{\mu,h}\|_V &\leq C \inf_{v_h \in V_h} \|\phi_\mu - v_h\|_V, \end{aligned}$$

where C is a positive constant independent of h. In particular, if the polynomial approximation is of order 1, we have

$$\begin{aligned} |\lambda_\mu - \lambda_{\mu,h}| &\leq Ch^2, \\ \|\phi_\mu - \phi_{\mu,h}\|_V &\leq Ch. \end{aligned}$$

Chapter 2

Reduced basis methods

The situation is as follows : we have a Galerkin parametrized FE Eigenproblem that gives a high-fidelity approximation of (1.1). In this chapter, we aim at finding a method which reduces considerably the computation time while preserving a certain accuracy on the solution.

2.1 Reduced problem

Let us recall the high-fidelity problem (1.6)

$$A_{h,\mu}\phi_h = \lambda_h B_{h,\mu}\phi_h, \tag{2.1}$$

with

$$\begin{aligned} (A_{h,\mu})_{ij} &= a_\mu(\chi_i, \chi_j), \\ (B_{h,\mu})_{ij} &= b_\mu(\chi_i, \chi_j). \end{aligned} \tag{2.2}$$

For our reduced problem, we look for a solution in a given N-dimensional reduced space spanned by our reduced basis $(\xi_1, \ldots, \xi_N)$ written as

$$\phi_N = \sum_{k=1}^{N_h} \phi_N^k \xi_k.$$

Let $\mathcal{V}_N := \mathrm{Span}(\xi_1, ..., \xi_N)$ and V_N denotes the matrix representation of (ξ_k) in the Galerkin basis. Therefore, the considered reduced problem is a Galerkin projection of (2.1),

$$\begin{aligned} &\text{Find } (\phi_N, \lambda_N) \text{ such that for all } v_N \in \mathcal{V}_N \\ &a_\mu(\phi_N, v_N) = \lambda_N b_\mu(\phi_N, v_N). \end{aligned} \tag{2.3}$$

By defining

$$\begin{aligned} A_N &= (a_\mu(\xi_j,\xi_i))_{i,j} = V_N^T A_{h,\mu} V_N, \\ B_N &= (b_\mu(\xi_j,\xi_i))_{i,j} = V_N^T B_{h,\mu} V_N, \end{aligned}$$

we can reformulate (2.3) in an algebraic formulation

$$\text{Find } (\phi_N, \lambda_N) \text{ such that} \qquad A_N \phi_N = \lambda_N B_N \phi_N. \tag{2.4}$$

For clarity, we chose to drop the μ in the terms A_N, B_N, ϕ_N and λ_N, but the reader should bear in mind that all those terms are μ-dependent.

Finally, by denoting ϕ_N the approached solution to our problem, we can evaluate the error for any norm Z and its Gram matrix in the FE basis Z_h

$$\|\phi_h - \phi_N\|_Z^2 = U_h^T \ Z_h \ (V_N \phi_N).$$

Due to the affine parametrization of our problem, we can separate every μ-independent terms and compute them in an *offline phase*, then assemble those with the μ-dependent parameters to compute the approached solution in an *online phase*.

Let us observe that Eigenvectors are unique up to a multiplication factor, so in order to approach the FE Eigenvector, we choose

$$\begin{aligned} b(\phi_h, \phi_h) &= 1, \\ b(\phi_N, \phi_N) &= 1, \\ b(\phi_h, \phi_N) &> 0. \end{aligned}$$

Imposing ϕ_N and ϕ_h to have positive components, this allows to uniquely define the eigenvector solutions of (2.3) and (1.5).

However, the issue of the space $\mathcal{V}_N$ still lingers. We will now see different approaches to compute a suitable reduced basis.

2.2 Proper Orthogonal Decomposition

We first propose to use a *Proper Orthogonal Decomposition* (POD) technique [8].
We start by computing a certain number n_s of high fidelity solutions $\phi_h(\mu_1), \dots, \phi_h(\mu_{n_s})$, called *snapshots*, from the set of parameters $P_{train} = \{\mu_1, \dots, \mu_{n_s}\}$. We label this set of solutions as the *training set*.

By denoting $\Phi_h^1, \dots, \Phi_h^{n_s}$ their vector representation in the finite element basis $(\chi_k)_{k\in\{1,\dots,N_h\}}$, we assemble the so-called *snapshot matrix* S

$$S = [\Phi_h^1 | \dots | \Phi_h^{n_s}].$$

The POD allows us to find the N-dimensional matrix that is closest to S in term of a certain euclidean norm $\|.\|_X$. By denoting X_h the Gram matrix of this norm for the FE basis, and by defining $\mathcal{V}_N^X$ the set of all X-orthogonal matrices

$$\mathcal{V}_N^X = \{W \in \mathbb{R}^{N_h \times N},\ W^T X_h W = I_N\}.$$

We look for the matrix V_N of $\mathcal{V}_N^X$ that minimizes the square of the X-norm of the error between each snapshot ϕ_h^i and its X-orthogonal projection onto the subset spanned by V_N, that we shall denote as $\mathbb{P}_{V_N}^X$ (see [8]),

$$V_N = \underset{W \in \mathcal{V}_N^X}{\text{argmin}} \sum_{i=1}^{n_s} \left\| \Phi_h^i - \mathbb{P}_W^X \Phi_h^i \right\|_X^2 .$$

Writing $W = [w_1|...|w_N]$, we can write

$$\forall U \in \mathbb{R}^{N_h} \ \ \mathbb{P}_W^X = \sum_{i=1}^{N} < U, w_i >_X w_i = W W^T X_h U.$$

Therefore

$$V_N = \underset{W \in \mathcal{V}_N^X}{\text{argmin}} \sum_{i=1}^{n_s} \left\| \Phi_h^i - W W^T X_h \Phi_h^i \right\|_X^2 .$$

As it happens, the procedure that allow to compute the best N-dimensional approximation of a N_h matrix is called *Singular Value Decomposition* or SVD for short.
By setting $\tilde{S} = X_h^{1/2} S$, and by computing the SVD of $\tilde{S}$,

$$\begin{aligned} X_h^{1/2} S &= U\ \Sigma\ Z^T, \\ U &= [\tilde{\xi_1}|...|\tilde{\xi_{N_h}}] \in \mathbb{R}^{N_h \times N_h}, \\ \Sigma &= diag(\sigma_1(S), ...\sigma_{min(n_s,N_h)}(S)), \\ Z &= [\tilde{\psi_1}|...|\tilde{\psi_{n_h}}] \in \mathbb{R}^{n_s \times n_s}. \end{aligned}$$

With the singular values $(\sigma_i(S))$ are in decreasing order, and U, Z are two orthogonal matrices. We have the following result

Theorem 6. *The best approximation for S is $V_N = [X_h^{-1/2}\tilde{\xi_1}|...|X_h^{-1/2}\tilde{\xi_N}]$,*

$$[X_h^{-1/2}\tilde{\xi_1}|...|X_h^{-1/2}\tilde{\xi_N}] = \underset{W \in \mathcal{V}_N^X}{argmin} \sum_{i=1}^{n_s} \left\| \Phi_h^i - W W^T X_h \Phi_h^i \right\|_X^2 .$$

Proof. First, let us observe that by defining

$$\tilde{S} = X_h^{1/2} S \ \text{ and } \ \tilde{W} = X_h^{1/2} W,$$

we have

$$\begin{aligned}
\min_{W \in \mathcal{V}_N^X} \sum_{i=1}^{n_s} \left\| \phi_h^i - WW^T X_h \phi_h^i \right\|_X^2 &= \min_{W \in \mathcal{V}_N^X} \sum_{i=1}^{n_s} \left\| X_h^{1/2} \phi_h^i - X_h^{1/2} WW^T X_h \phi_h^i \right\|_2^2 \\
&= \min_{W \in \mathcal{V}_N^X} \left\| X_h^{1/2} S - X_h^{1/2} WW^T X_h S \right\|_2^2 \qquad (2.5) \\
&= \left\| \tilde{S} - \tilde{W}\tilde{W}^T \tilde{S} \right\|_F^2 . \qquad (2.6)
\end{aligned}$$

Where $\|.\|_F$ denotes the Frobenius norm (i.e. the standard euclidean matrix norm). The Eckart–Young–Mirsky theorem [8] explicit the matrix solution of this minimization problem

Eckart-Young-Mirsky Theorem. *Given the SVD decomposition of a matrix $A \in \mathbb{R}^{m \times n}$,*

$$A = [\xi_1|...|\xi_m] \; diag_{m,n}(\sigma_1, ..., \sigma_{min(m,n)}) \; [\psi_1|...|\psi_n]^T,$$

we have for a given integer $k \leq rank(A)$,

$$\left\| A - \sum_{k=1}^{k} \sigma_i \xi_i \psi_i^T \right\|_F = \underset{rank(B) \leq k}{argmin} \; \|A - B\|_F = \sqrt{\sum_{i=k+1}^{rank(A)} \sigma_i}.$$

Therefore, the best N-dimensional approximation in the Frobenius sense of the matrix $\tilde{S}$ is

$$\tilde{S_N} = \sum_{k=1}^{N} \sigma_k \tilde{\xi_i} \tilde{\psi_i}^T .$$

By the SVD decomposition, we have for an integer i,

$$\tilde{S}\tilde{\psi_i} \; = \; \sigma_i \tilde{\xi_i} \;\; \text{and} \;\; \tilde{S}^T \tilde{\xi_i} \; = \; \sigma_i \tilde{\psi_i} \quad \text{, therefore} \quad \tilde{\psi_i} \; = \; \frac{1}{\sigma_i} \tilde{S}^T \tilde{\xi_i}.$$

By defining $\tilde{V_N} = [\tilde{\xi_1}|...|\tilde{\xi_N}]$, we have

$$\tilde{S_N} = \sum_{k=1}^{N} \sigma_i \frac{1}{\sigma_i} \tilde{\xi_i} (\tilde{S} \tilde{\xi_i})^T = \tilde{V_N} \tilde{V_N}^T \tilde{S}.$$

We conclude by remarking that $V_N = X_h^{-1/2} \tilde{V_N}$.

□

On a side note, the computation of $X_h^{1/2}$ can be tiresome : to cut its computational cost, we found that the singular values (σ_k) are solution to the following Eigenproblem

$$S^T X_h S \times \tilde{\psi_i} \; = \; \sigma_i^2(S) \tilde{\psi_i} \quad \forall i \in \{1, \ldots, n_s\}.$$

Then, we compute the first N vectors to compute V_N,

$$\xi_i = X_h^{-1/2}\tilde{\xi}_i = \sigma_i^{-1}(S)S\tilde{\psi}_i.$$

Let us remark that this procedure will lower the accuracy on the computation of σ_i , due to the squaring operation.

The X-orthogonal basis $(\xi_1, \ldots, \xi_N)$ is a suitable *reduced basis* for our problem.

2.3 Faster computation : the Greedy algorithm

Following the POD discussion, we might ask the purpose of another procedure : after all, for a given integer N, the POD provided the N-dimension approximated space that approximated best -w.r.t a certain norm X- the training set of finite element solutions.

Well, the aim of the greedy procedure is to iteratively construct the reduced basis, instead of building it all at once as the POD procedure does, which would ideally reduce the cost of computation.

2.3.1 The Ideal Greedy

To study the efficiency of the iterative process, we look at an algorithm we labelled "Ideal Greedy".

We start with a randomly sampled parameter μ_1, and we build the space $\mathcal{V}_N$ by normalizing w.r.t a certain norm X the eigenvector solution of (1.2).Then at each step k , we select the parameter μ^* satisfying the following optimal property,

$$\mu^* = \underset{\mu \in P_{train}}{argmax} \left\| \phi_h(\mu) - \phi_k(\mu) \right\|_X ,$$

where ϕ_k is the RB solution to the k - dimensional reduced problem, that is, the solution of (2.3) where the space $\mathcal{V}_N$ is spanned by the previously computed X-orthonormal basis $(\xi_1, ..., \xi_k)$.

We therefore have the following algorithm,

Algorithm 1 Ideal Greedy algorithm to compute N-dimensional RB space

```
Get random sample $\mu_1 \in P_{train}$
$\phi_1 \leftarrow \phi_h(\mu_1)$
$\xi_k \leftarrow \phi_1 / \|\phi_1\|_X$
$V_N \leftarrow [\ \xi_k\ ]$
$\epsilon = 1$
for $k \in \{1, ..., N-1\})$ do
    $\mu_{k+1} \leftarrow \underset{\mu \in P_{train}}{argmax} \|\phi_h(\mu) - \phi_k(\mu)\|_X$
    $\xi_k \leftarrow \text{ON}(\ \phi_h(\mu_{k+1}), V_k, X\ )$
    $V_k \leftarrow [\ V_k \mid \xi_k\ ]$
end for
```

where the function "ON(u, W, X)" computes an X-orthonormalization of the vector u, in order to be orthogonal to all column of a matrix W.

At this stage, one might see that as the space is constructed iteratively, at each step, a minimum over all the training set is required, nullifying the purpose of this algorithm. Indeed, as the POD presented in Section 2.1 computes the N-dimensional space closest -w.r.t a certain norm- of the snapshots, any other procedure will yield greater errors w.r.t. the snapshots. The purpose of greedy algorithm is to cut computation time while keeping the error as low as possible. However, a greedy algorithm more expensive than the POD serves no purpose, since it does not reduces neither the error nor the computation time. To circumvent this issue, we will present a variant of this algorithm -labelled "Efficient Greedy"-, that will use an *a posteriori estimator* to evaluate the error $\|\phi_h(\mu) - \phi_k(\mu)\|_X$ at each step k.

2.3.2 A posteriori estimator

Let us define the *residual* as an element of the dual of V, $Res(w, \beta) \in V'$, for a couple $(w, \beta) \in V \times \mathbb{R}$ by

$$\langle Res(w, \beta), v \rangle_{V',V} = \beta b_\mu(w, v) - a_\mu(w, v) \ \ \forall v \in V.$$

If the couple (u, λ) is a solution to (1.1), we have

$$\langle Res(u, \lambda), v \rangle_{V',V} = 0 \ \ \forall v \in V.$$

As $(V, \|.\|_{a,\mu})$ is a Hilbert space, the Riesz-Fréchet representation theorem implies the existence and uniqueness of a variable $\phi^*_{a,\mu} \in V$ such that

$$\forall v \in V \quad \langle Res(w, \beta), v \rangle_{V',V} = a_\mu(\phi^*_{a,\mu}(w, \beta), v).$$

To lighten the notations, let us rename $res_{a,\mu}$ the variable $\phi^*_{a,\mu}$.

Let us write (ϕ_h, λ_h) the FE solution to (1.5) and (ϕ_N, λ_N) a RB solution to (2.3) (i.e. a Galerkin projection of the FE problem onto a N-dimensional space). To clarify the situation, let us recap the link between $Res(.,.)$ and its Riesz representation $res(.,.)$ over different special cases.

$Res(w,\beta)$	$res_{a,\mu}(w,\beta)$
$Res(u,\lambda)$	0
$Res(\phi_h,\lambda_h)$	$res^h_{a,\mu}$
$Res(\phi_N,\lambda_N)$	$res^N_{a,\mu}$

However, for our study, which concern the approximation of the FE solution by a reduced problem, it might be more appropriate to define the residual as an element of the dual of V_h: space adapted to the Galerkin FE approximation rather than onto V'. We then define, for any couple $(w_h,\beta)\in V_h\times\mathbb{R}$,

$$\left\langle Res^{FE}(w_h,\beta), v\right\rangle_{V_h',V_h} = \beta b_\mu(w_h,v_h) - a_\mu(w_h,v_h) \quad \forall v_h\in V_h. \tag{2.7}$$

We have similarly, according to the Riesz representation theorem,

$Res^{FE}(w_h,\beta)$	$res^{FE}_{a,\mu}(w,\beta)$
$Res^{FE}(u,\lambda)$	undefined
$Res^{FE}(\phi_h,\lambda_h)$	0
$Res^{FE}(\phi_N,\lambda_N)$	$res^{FE,N}_{a,\mu}$

We look for the energetic error between the FE solution and the RB approximation

$$\begin{aligned} \|\phi_h-\phi_N\|^2_{a,\mu} &= a_\mu(\phi_h-\phi_N,\phi_h-\phi_N), \\ &= \lambda_h b_\mu(\phi_h,\phi_N) + a_\mu(\phi_N,\phi_h-\phi_N). \end{aligned}$$

We select the vector ϕ_h and ϕ_N to be of b_μ-norm 1 and pointing in the same direction (still in the b_μ sense)

$$\begin{aligned} b(\phi_h,\phi_h) &= 1, \\ b(\phi_N,\phi_N) &= 1, \\ b(\phi_h,\phi_N) &\geq 0. \end{aligned} \tag{2.8}$$

We can define an *a posteriori estimator* for the energetic error
Theorem 7.

$$\|\phi_h-\phi_N\|^2_{a,\mu} \leq \left\|Res^{FE}(\phi_N,\lambda_N)\right\|^2_{-a,\mu} + (\lambda_N+\lambda_h)\,\|\phi_h-\phi_N\|^2_{b,\mu}\,.$$

Proof. To lighten the notations, let us denote $e_n := \phi_h-\phi_N$. We have

$$\begin{aligned} \|e_n\|^2_{a_\mu} &= a_\mu(e_n,e_n) \\ &= a_\mu(e_n,\phi_h) - a_\mu(e_n,\phi_N) \\ &= \lambda_h b_\mu(e_n,\phi_h) - a_\mu(e_n,\phi_N). \end{aligned}$$

Using the fact that b_μ is symmetric and (2.8), we have

$$b_\mu(e_n,e_n) = 2(1-b_\mu(u_h,u_N)) = 2b_\mu(u_h,e_N) = -2b_\mu(e_N,u_N). \tag{2.9}$$

Therefore,

$$\begin{aligned}
\|e_n\|_{a_\mu}^2 &= \frac{\lambda_h}{2} b_\mu(e_n, e_n) - a_\mu(e_n, \phi_N) + \lambda_N b_\mu(e_n, \phi_N) - \lambda_N b_\mu(e_n, \phi_N) \\
&= \frac{\lambda_h + \lambda_N}{2} \|e_n\|_{b_\mu}^2 - \left\langle Res^{FE}(\phi_N, \lambda_N), e_n \right\rangle_{V',V} \\
&\leq \frac{\lambda_h + \lambda_N}{2} \|e_n\|_{b_\mu}^2 + \left\| Res^{FE}(\phi_N, \lambda_N)_{-a_\mu} \right\| \|e_n\|_{a_\mu} \\
&\leq \frac{1}{2} \|Res^{FE}(\phi_N, \lambda_N)\|_{-a_\mu}^2 + \frac{1}{2} \|\phi_h - \phi_N\|_{a_\mu}^2 + \frac{\lambda + \lambda_h}{2} \|\phi_N - \phi_h\|_{b_\mu}^2,
\end{aligned}$$

where we used (2.9) and (2.7) in the second line. It allows us to conclude.

□

Let us recall that

$$\begin{aligned}
\left\| Res^{FE}(\phi_N, \lambda_N) \right\|_{-a,\mu}^2 &= \sup_{v \in V_h} \frac{\left\langle Res^{FE}(\phi_N, \lambda_N), v \right\rangle_{V_h', V_h}}{a_\mu(v, v)} \\
&= a_\mu(res_{a,\mu}^{FE,N}, res_{a,\mu}^{FE,N}).
\end{aligned}$$

2.3.3 Computation of the dual norm of the residual

By Riesz's representation theorem,

$$a_\mu(res_{a,\mu}^{FE,N}, v_h) = \lambda_N b_\mu(\phi_N, v_h) - a_\mu(\phi_N, v_h).$$

Algebraically, by denoting $\underline{R}_{a,\mu}^{FE,N}$ the representation of $res_{a,\mu}^{FE,N}$ in the FE basis

$$\underline{R}_{a,\mu}^{FE,N} = A_{h,\mu}^{-1}(\lambda_N B_{h,\mu} - A_{h,\mu}) V_N \underline{\phi_N}. \tag{2.10}$$

Finally,

$$\left\| Res^{FE}(\phi_N, \lambda_N) \right\|_{-a,\mu}^2 = (\underline{R}_{a,\mu}^{FE,N})^T \; A_{h,\mu} \; \underline{R}_{a,\mu}^{FE,N}.$$

Unfortunately, the inversion of $A_{h,\mu}$ breaks the μ-affine decomposition, because the inverse operation is neither linear nor affine. Thus, the "offline / online" decomposition is no longer possible, so the computation of each residual requires to solve (2.10), which will be expensive to compute for large FE spaces. A solution might be to estimate the residual by another separated RB approximation.

Note that the residual is a suitable *a posteriori estimator* only if the a_μ-norm dominates the b_μ-norm. As a first approach, we will evaluate whether it is an appropriate *a posteriori estimator*.

2.3.4 The Efficient Greedy algorithm

We then present the computable greedy algorithm,

Algorithm 2 Ideal Greedy algorithm to compute N-dimensional RB space

```
Get random sample $\mu_1 \in P_{train}$
$\phi_1 \leftarrow \phi_h(\mu_1)$
$\xi_k \leftarrow \phi_1 / \left\| \phi_1 \right\|_X$
$V_N \leftarrow [\ \xi_k\ ]$
$\epsilon = 1$
for $k \in \{1, ..., N-1\})$ do
    $\mu_{k+1} \leftarrow \underset{\mu \in P_{train}}{argmax} \left\| Res^{FE,N}(\phi_k, \lambda_k) \right\|_{-a,\mu}$
    $\xi_k \leftarrow$ ON( $\phi_h(\mu_{k+1}), V_k, X$ )
    $V_k \leftarrow [\ V_k, \xi_k\ ]$
end for
```

This allows us to only compute one FE solution per iteration, which considerably reduces computational times, given that the estimator is sufficiently cheap to compute.

Chapter 3

Numerical Results

3.1 Verification of the Finite Element solver

We used the Python library *Fenics* to compute the Finite Element solution to the partial differential equation 1.1. We defined our geometry $\Omega := [0, L]^2$ with $L = 60$, and discretized it as a 2×2 grid for the parameter space. Our approximation functions were Lagrangian $\mathcal{P}_{1,h}$ functions. The mesh grid was selected as a cross triangular mesh(see Figure 3.1).

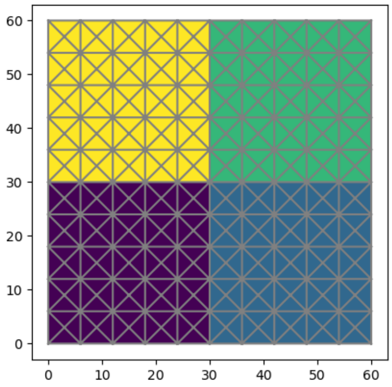

Figure 3.1: Domain Ω and subdomains Ω_k for the parameter space (colored blocks), with a example of a FE mesh discretization of 221 degrees of freedom (10 elements along one border)

We first aimed at verifying the results of Theorem 5. However, as we do not dispose of the exact solution of (1.1), we used a fine mesh to compute a solution with 33025 degrees of freedom (128 degrees of freedom per edge), and we compare this solution to coarser mesh approximations. The results are shown in Figure 3.2.

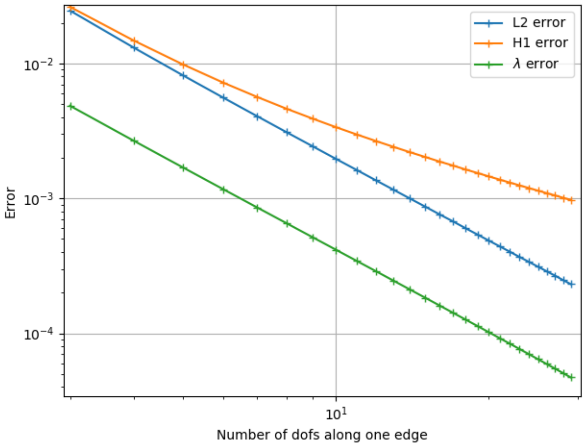

Figure 3.2: Log-scale plot of the $L^2(\Omega)$, $H^1(\Omega)$ and λ_μ error for a given admissible parameter μ

We can see that asymptotically, the slope of the V-log error is -1, and the slope of the λ_μ-log error is -2, which confirms the results of Theorem 5.

3.2 POD analysis

3.2.1 Methodology

We first need to determinate the *POD* projection norm that we shall use. To assess its importance, we ran sample tests for different projection norm and their impact on the overall errors. We tried

- the H-norm : $\| \, . \, \|_{L^2(\Omega)}$,
- the V-norm : $\| \, . \, \|_{H^1(\Omega)}$,
- the V-semi-norm : $\| \, \nabla . \, \|_{L^2(\Omega)}$,
- a mean energetic norm over all our training sample : $U \in \mathbb{R}^{N_h} \mapsto \dfrac{1}{\#P_{train}} \displaystyle\sum_{k=1}^{\#P_{train}} \sqrt{U^T A_{\mu_k,h} U}$, labelled as *EMM* for *Energy Mean Matrix* projection.

The parameters μ are randomly sampled according to a uniform sampling. Each coefficient Σ_t, Σ_{s0} and $\nu\Sigma_f$ are taking value in $[0, 1]$. To ensure the coercivity of a_μ, the constraint $\Sigma_t \geq \Sigma_{s0}$ was added. A different sample was made for each subregion of the domain, therefore μ contains $Q_k = 4$ different values for each of the 3 coefficients Σ_t, Σ_{s0} and $\nu\Sigma_f$.

The set P_{train} is constituted of 100 training samples, each of a 221 degrees of freedom for computational efficiency. The norms were computed with the Gram matrices in the FE basis of the relevant scalar product.

3.2.2 Results

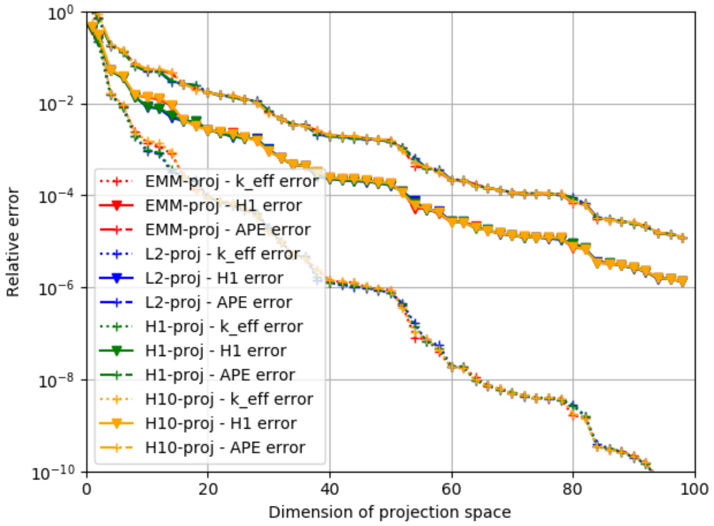

Figure 3.3: Colors - Red : The EMM projection norms, Blue : The $L^2(\Omega)$ projection, Green : The $H^1(\Omega)$ projection, Yellow : The $H^1(\Omega)$ semi-norm projection. Linestyle - Dotted : k_{eff} error, Full : V-error, Dashed : "APE", i.e. dual norm of the residual

We can see in Figure 3.3 that any of the 4 choices yields good approximations. We shall henceforth project with the V-norm.

To better assess our reduced basis method, we now randomly choose a set P_{test} of parameters $(\mu_k)_{k\in\{1,\ldots,\#P_{test}\}}$. We then plot the mean, max and min error over P_{test}. We choose 20 as the number of test samples, as seen in Figure 3.4.

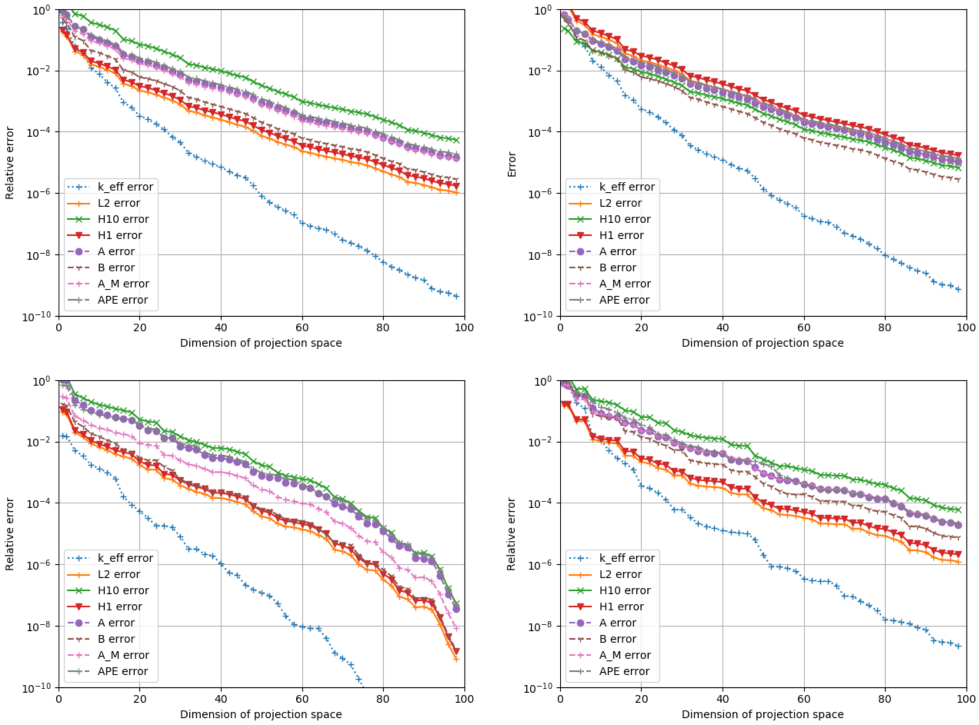

Figure 3.4: Error over 20 random sampling of parameters. In reading order : mean relative error, mean error, min relative error and max relative error. NB : "A_M" stand for the *Energy Mean Matrix* norm previously defined

This method is quite satisfying, we can cluster the errors into 4 groups in increasing relative accuracy,

- k_{eff} error : it only takes 40 basis vector to have a accuracy of 10^{-5};
- $L^2(\Omega)$, $H^1(\Omega)$ and b_μ : the error on the "mass" term is represented here. Indeed, having the $H^1(\Omega)$ error being close to the $L^2(\Omega)$ error means that the $H^1_0(\Omega)$ norm is relatively small compared to them;
- a_μ, $\underset{\mu \in P_{train}}{\text{mean}}\, a_\mu$ and APE $:= \|Res\|_{-a,\mu}$: the energy-related errors.
- $H^1_0(\Omega)$: although relatively small in magnitude, the relative error considerations underlines that the relative accuracy obtained on the $H^1_0(\Omega)$ are the slowest to obtain.

We also verified that the b_μ-error was negligible compared to the a_μ-error, which justifies taking the dual norm of the residual as the *a posteriori estimator* (see Theorem 7).

As all of our simulation were done with a coarse mesh, to illustrate that similar results were obtained for finer meshes, we present results on a 5101-degrees of freedom mesh (see Figure 3.5).

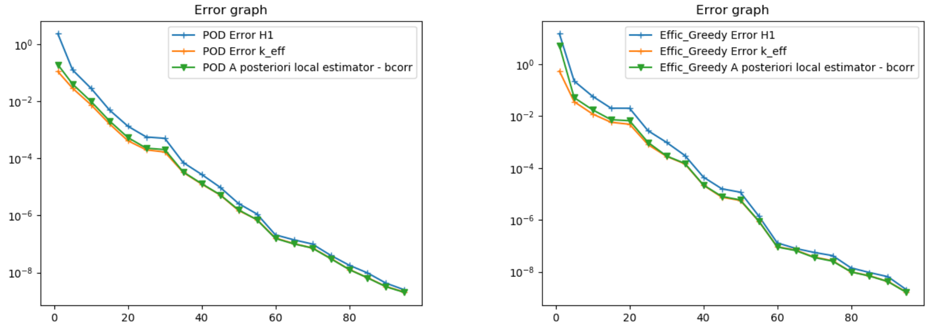

Figure 3.5: Error for the POD (left) and the efficient greedy algorithm (right) for a finer mesh of 5101 degrees of freedom

3.3 Comparison between the different RB approaches

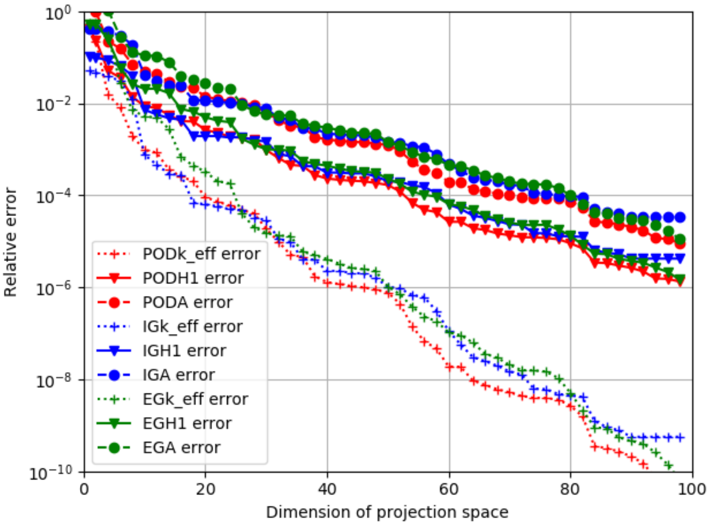

Figure 3.6: Comparison of the 3 RB approaches. Color - Red : POD, Blue : Ideal Greedy, Green : Efficient greedy with the residual's dual norm as *a posteriori estimator*. Linestyle - Dotted : k_{eff} error, Solid : V error, Dashed : Energy a_μ error.

Figure 3.6 compare the 3 approaches - POD, Ideal and Efficient greedy- on a random parameter μ. We observe few non significant variations when we sampled other parameters. We can see that the efficient greedy is quite accurate w.r.t the POD accuracy, but the POD still seems to yields the best results.

3.4 Computational cost

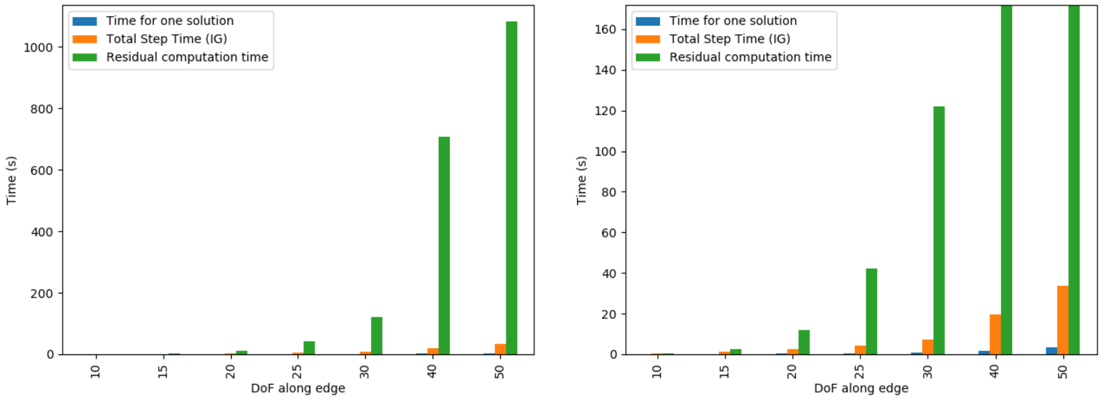

Figure 3.7: Computational time for a training set with 50 snapshots

Figure 3.7 presents the computational cost of the RB methods. The aim was to compare the RB bases construction for the greedy algorithms at each step with the FE solver. We can see that the residual computation is incredibly costly once we increase the size of the mesh. Therefore another reduction method is required to compute the residual in order for this method to be competitive with the others. The *Ideal Greedy* is still quite costly w.r.t the FE computational time, which make it unreliable in practice.

Part II

Two Group Diffusion Equation

Chapter 4

Setting of the model

4.1 Continuous problem

The two Group Diffusion parametric equation is a generalization to its simpler version presented in the previous part. Still denoting $\mu \in P$ the parameter, it stands as follows

$$\begin{aligned}
&\text{Find } (\phi_1, \phi_2, \lambda) \in V \times V \times \mathbb{R} \text{ such that} \\
&-\text{div}(D_1(\mu)\nabla\phi_1) + \Sigma_{11}(\mu)\phi_1 - \Sigma_{12}(\mu)\phi_2 = \lambda\ (Mf_{11}(\mu)\phi_1 + Mf_{12}(\mu)\phi_2) \\
&-\text{div}(D_2(\mu)\nabla\phi_2) - \Sigma_{21}(\mu)\phi_1 + \Sigma_{22}(\mu)\phi_2 = \lambda\ (Mf_{21}(\mu)\phi_1 + Mf_{22}(\mu)\phi_2).
\end{aligned} \tag{4.1}$$

The parameters are (see [6] for a detailed description of each),

- $D_i(\mu)$: Diffusion coefficient for group i (inversely related to the total macroscopic cross section of group i);
- $\Sigma_{ii}(\mu)$: Removal macroscopic cross section of group i;
- $\Sigma_{ij}(\mu)$: Scattering macroscopic cross section from group j to group i, for $i \neq j$;
- $Mf_{ij}(\mu)$: Fission macroscopic cross section from group j to group i.

Assumptions 2. *We present the following assumptions,*

- Ω *is open, connected and bounded;*
- *Each* Ω_i *for* $i \in \{1, ..., Q_k\}$ *is open and its boundary is piecewise smooth;*
- $D_i(\mu) \in \mathcal{PW}^{1,\infty}(\Omega)$ *for* $i \in \{1, 2\}$*;*
- $\Sigma_{ij}(\mu) \in \mathcal{PW}^{1,\infty}(\Omega)$ *for* $(i, j) \in \{1, 2\}^2$*;*
- $Mf_{ij}(\mu) \in \mathcal{PW}^{1,\infty}(\Omega)$ *for* $(i, j) \in \{1, 2\}^2$ *;*
- $\exists (D_i)_*, (D_i)^* > 0, \forall k \in \{1, ..., Q_k\} \quad \forall \mu \in P \quad (D_i)_* \leq (D_i)_{|\Omega_k}(\mu) \leq (D_i)^*$ *for* $i \in \{1, 2\}$*;*
- $\exists (\Sigma_{ij})_*, (\Sigma_{ij})^* > 0, \forall k \in \{1, ..., Q_k\} \quad \forall \mu \in P \quad (\Sigma_{ij})_* \leq (\Sigma_{ij})_{|\Omega_k}(\mu) \leq (\Sigma_{ij})^*$ *for* $(i, j) \in \{1, 2\}^2$*;*
- $\exists (Mf_{ij})_*, (Mf_{ij})^* > 0, \forall k \in \{1, ..., Q_k\} \quad \forall \mu \in P \quad (Mf_{ij})_* \leq (Mf_{ij})_{|\Omega_k}(\mu) \leq (Mf_{ij})^*$ *for* $(i, j) \in \{1, 2\}^2$.

We also impose thee following condition, which is sufficient but not necessary for many theorems that will be presented, and a sharper version of the following hypothesis may be found in [6],

- $\Sigma_{ii} \leq \frac{1}{2}(\Sigma_{12} - \Sigma_{12})$ *for* $i \in \{1, 2\}$.

For the variational formulation of problem (4.1), we define the vector $u = (\phi_1, \phi_2) \in V \times V$, and we define the continuous forms $a_\mu : V \times V \longrightarrow \mathbb{R}$ and $b_\mu : V \times V \longrightarrow \mathbb{R}$ by, for a given $u = (\phi_1, \phi_2) \in V \times V$ and $v = (v_1, v_2) \in V \times V$,

$$\begin{aligned}
a_\mu(u, v) &= \int_\Omega D_1(\mu)\nabla\phi_1 \cdot \nabla v_1 + \int_\Omega \Sigma_{11}(\mu)\phi_1 v_1 - \int_\Omega \Sigma_{12}(\mu)\phi_1 v_2 \\
&+ \int_\Omega D_2(\mu)\nabla\phi_2 \cdot \nabla v_2 - \int_\Omega \Sigma_{21}(\mu)\phi_2 v_1 + \int_\Omega \Sigma_{22}(\mu)\phi_2 v_2, \\
b_\mu(u, v) &= \int_\Omega M f_{11}(\mu)\phi_1 v_1 + \int_\Omega M f_{12}(\mu)\phi_1 v_2 \\
&+ \int_\Omega M f_{21}(\mu)\phi_2 v_1 + \int_\Omega M f_{22}(\mu)\phi_2 v_2.
\end{aligned}$$

Therefore the variational problem writes

$$\begin{aligned}
&\text{Find } (\phi_\mu, \lambda_\mu) \in V \times V \times \mathbb{R} \text{ such that for all } v \in V \times V, \\
&a_\mu(\phi_\mu, v) = \lambda_\mu \, b_\mu(\phi_\mu, v).
\end{aligned} \tag{4.2}$$

The well-posedness of the equation (4.2) is similar to the symmetric one due to the coercive assumption of a_μ, that we suppose holds thanks to Assumptions 2.

The problem is still well-posed in the sense that,

Theorem 8. *Let Assumptions 2 holds. Then problem (4.2) is well posed (up to a multiplication factor for ϕ_μ) and is equivalent to (4.1),*

$$\forall v \in V \quad a_\mu(\phi_\mu, v) = \lambda_\mu \, b_\mu(\phi_\mu, v).$$

In addition, λ_μ is the unique eigenvalue associated to ϕ_μ, a positive vector, and is a simple eigenvalue.

The proof is identical to the one presented in the one group symmetric case.

4.2 Discretization of the problem

We use the same Finite Element context as in the symmetric case. To that aim, we define the vector $\underline{U} = (\underline{\phi_1}, \underline{\phi_2})^T \in \mathbb{R}^{2N_h}$ and we solve the coupled problem

$$A_{\mu,h}\underline{U} = \lambda B_{\mu,h}\underline{U}.$$

with the "artificial" FE basis $(\psi_k, k \in [|1, 2N_h|])$ defined as

$$\begin{array}{rcl} \forall k \in [|1, N_h|] \;\; \psi_k & = & \chi_k, \\ \forall k \in [|1, N_h|] \;\; \psi_{N_h+k} & = & \chi_k, \end{array}$$

and the matrices

$$\begin{array}{rcl} A_{\mu,h} & = & (a_\mu(\psi_j, \psi_i))_{i,j\in\{1,\ldots,2N_h\}}, \\ B_{\mu,h} & = & (a_\mu(\psi_j, \psi_i))_{i,j\in\{1,\ldots,2N_h\}}. \end{array}$$

We therefore have a "block matrix" problem, each bloc containing $N_h \times N_h$ elements.

Chapter 5

Reduced basis methods

5.1 Proper Orthogonal Decomposition

In this section, we apply the POD methodology to build a reduced basis for the non symmetric two group Diffusion equation [5].

Given the way we chose to solve this problem, by introducing the vector $\underline{U} = (\phi_1 \ \phi_2)^T \in \mathbb{R}^{2N_h}$, we can see that multiple approaches are possible to compute a suitable POD-reduced basis.

First, we can look at the simple problem on $\underline{U}$, and therefore compute the *POD* as we did in the symmetric case. Given a number n_s of snapshots $(\underline{U}(\mu_k))_{k\in[1,\dots,n_s]}$, we look for a basis

$$V_N = [\xi_1|\dots|\xi_N] \ , \ \xi_k \in \mathbb{R}^{2N_h}.$$

Then, selecting for a projected N-dimensional problem the solution

$$\phi_N^* = \sum_{k=1}^{N} \alpha_k \times \xi_k.$$

This approach will be defined as **Multi Group** or MG for short: it is the same as the symmetric POD, but with an initial space of dimension $2N_h$ instead of N_h.

Secondly, we can shift point of views and look at the problem as a two group coupled equation, and calculate the *POD* separately on each group separately: given some snapshots $(\phi_1(\mu_k))_{k\in[1,\dots,n_s]}$ and $(\phi_2(\mu_k))_{k\in[1,\dots,n_s]}$, we look for two basis

$$\begin{aligned} V_N^1 &= [\psi_1^1|\dots|\psi_{N_1}^1] \ , \ \psi_k^1 \in \mathbb{R}^{N_h}, \\ V_N^2 &= [\psi_1^2|\dots|\psi_{N_2}^2] \ , \ \psi_k^2 \in \mathbb{R}^{N_h}. \end{aligned}$$

Then, we look for the solution in the space

$$\mathcal{V}_N := \text{Span}(\begin{pmatrix}\psi_i^1\\0\end{pmatrix})_{i\in\{1,\dots,N_1\}} + \text{Span}(\begin{pmatrix}0\\\psi_i^2\end{pmatrix})_{i\in\{1,\dots,N_2\}} :$$

$$\phi_N = \begin{pmatrix}\sum_{k=1}^{N_1}\alpha_k\psi_k^1\\\sum_{k=1}^{N_2}\beta_k\psi_k^2\end{pmatrix} \tag{5.1}$$

Which translates, in terms of the reduced problem to solve on $\underline{U}$, by taking the projection matrix

$$V_N = \begin{pmatrix}V_N^1 & 0\\0 & V_N^2\end{pmatrix}. \tag{5.2}$$

This approach is denoted as **Group Separated** or GS for short [5].

5.2 Reduced Problem

As the relevant bilinear forms $a_\mu(.,.)$ and $b_\mu(.,.)$ are defined on the same space, the immediate projection approach is to consider a simple *Galerkin* projection

$$A_{N,\mu} = V_N^T A_{h,\mu} V_N \ , \ B_{N,\mu} = V_N^T B_{h,\mu} V_N.$$

However an interesting approach is the so-called "Adjoint POD" or APOD [7], which is a Petrov Galerkin projection where the trial space is a POD on snapshots of the relevant problem and the test space is a POD on snapshot of the adjoint problem.

Therefore, if we denote by V_N the matrix of the POD and V_N^* the matrix of the Adjoint POD, the projected matrices are

$$A_{N,\mu} = V_N^T A_{h,\mu} V_N^* \ , \ B_{N,\mu} = V_N^T B_{h,\mu} V_N^*.$$

Interestingly, a special method was introduced to compute the adjoint basis : instead of two independent POD, the adjoint basis is computed with the same linear combination than the first one [7]. More precisely, recalling that S_N is the snapshot matrix and denoting S_N^* the adjoint snapshot matrix (assumed here to have the same size),

$$S_N = U\Sigma Z^T,$$

Which implies

$$\begin{aligned} U &= S_N Z\Sigma^{-1}, \\ U^* &= S_N^* Z\Sigma^{-1}. \end{aligned}$$

The previous equation indeed states that as the basis vector of V_N are expressed as linear combination of the snapshots S_N, the same coefficient are used to expressed the basis vector of V_N^* in terms of S_N^*,

$$\begin{aligned}\forall k \in [|1, ..., N|] \;\; \xi_k &= \sum_{k=1}^{n_s} C_k \times \phi_k, \\ \xi_k &= \sum_{k=1}^{n_s} C_k \times u_k^*.\end{aligned}$$

With $C_k = [Z\Sigma^{-1}]_k$ the k-th line of the matrix $Z\Sigma^{-1}$ [7].

However, we can also choose to incorporate both spaces with the adjoint solutions, by considering the Galerkin projection

$$A_{N,\mu} = (V_N^{sum})^T A_{h,\mu} V_N^{sum} \ , \ B_{N,\mu} = (V_N^{sum})^T B_{h,\mu} V_N^{sum}.$$

with V_N^{sum} the matrix representation of the space $\mathcal{V}_N + \mathcal{V}_N^*$. This approach shall be denoted as the *Galerkin Sum Projection*.

Chapter 6

Numerical results

6.1 Matrix conditioning

In order to assess whether a certain scheme is stable, we look at the condition number (see [4]) defined for the l^2 norm as

$$\kappa(A) = \|A\| \times \left\|A^{-1}\right\| = \frac{\sigma_{max}(A)}{\sigma_{min}(A)},$$

where $\sigma(M)$ denotes the singular values of M.

We say that a matrix M is *ill-conditionned* if $\kappa(M) >> 1$. Denoting $\mathbb{M}$ and $\mathbb{G}$ the mass and stiffness matrix for the heat equation,

$$\begin{aligned} \mathbb{M}_{i,j} &= \int_\Omega \chi_i \chi_j, \\ \mathbb{G}_{i,j} &= \int_\Omega \nabla\chi_i \cdot \nabla\chi_j. \end{aligned}$$

We have classically (see [4]), for our uniform mesh of unit size h, the existence of two positive constants c and c' such that,

$$\begin{aligned} \kappa(\mathbb{M}) &\leq c, \\ \kappa(\mathbb{G}) &\leq c' \times h^{-2}. \end{aligned}$$

The conditioning of a non singular matrix is a measure of how "non singular" it is: more precisely, the inverse of the conditioning denotes a distance between the matrix and the set of singular matrix,

$$\forall M \in GL_n(\mathbb{R}) \quad \kappa(M)^{-1} = \min_{E \in M_n(\mathbb{R})/GL_n(\mathbb{R})} \frac{\|E\|}{\|M\|}.$$

Therefore, the more a matrix is ill-conditioned, the closer it is to a singular matrix.

However, due to the different choices of norms and the multiple projection matrices, it is important to identify which matrix can be used to assess the stability of a method.

Assumptions 3. *We have*

- *Y denotes the gram matrix associated to the V-norm in the space spanned by V_N*
- *The RB method is obtained by a Galerkin Projection, thus $A_N := V_N^T A_h V_N$*
- *The matrix A_N is coercive w.r.t Y in the sense that for all $U \in \mathbb{R}^N$, $U^T A_N U \geq \alpha U^T Y U$*

Theorem 9. *Let Assumptions 3 holds. The matrix $A_{N,Y} := Y^{-1/2} A_N Y^{-1/2}$ has a bounded condition number, and there exists $C > 0$ such that*

$$\kappa(A_{N,Y}) \leq \frac{C}{\alpha_N}.$$

Proof. Let $U \in \mathbb{R}^N$. We have according to Assumption 3

$$U^T A_N U \geq \alpha_N U^T Y U.$$

By setting $\Phi_Y := Y^{1/2} U$ and $A_{N,Y} := Y^{-1/2} A_N Y^{-1/2}$, the previous equation is equivalent to

$$\Phi_Y^T A_{N,Y} \Phi_Y \geq \alpha_N \Phi_Y^T \Phi_Y. \tag{6.1}$$

Applying the *Singular Value Decomposition* of $A_{N,Y}$, we obtain

$$\begin{array}{rcl} A_{N,Y} : & = & G_Y \Sigma_Y Z_Y^T, \\ G_Y & = & [\xi_Y^1 | ... | \xi_Y^N] \in \mathbb{R}^{N \times N}, \\ \Sigma_Y & = & \text{diag}\,(\sigma_1, ..., \sigma_N), \\ Z_Y & = & [\psi_Y^1 | ... | \psi_Y^N] \in \mathbb{R}^{N \times N}. \end{array}$$

We then apply (6.1) by taking $\Phi_Y = \psi_Y^k$ for a given integer $k \in \{1, ..., N\}$,

$$\sigma_k(A_{N,Y}) \geq \frac{\alpha_N}{(\psi_Y^k)^T \xi_Y^k}.$$

This implies,

$$\sigma_{min}(A_{N,Y}) \geq \frac{\alpha_N}{((\psi_Y^N)^T \xi_Y^N}. \tag{6.2}$$

In addition, the V-continuity of $a_\mu(.,.)$ yields

$$\exists C > 0, \forall (U, V) \in \mathbb{R}^N \quad U^T A_N V \leq C (U^T Y U)^{1/2} (V^T Y V)^{1/2}.$$

Squaring this relation, and taking U as a left singular vector and V as a right singular vector will yield

$$\sigma_{max}(A_{N,Y}) \leq C. \tag{6.3}$$

Therefore, combining (6.2) and (6.3) gives the result.

□

This result can be similarly generalized for a *Petrov Galerkin* projection (see [8, Section 2.4.6]. Therefore, we can look at the conditioning of the matrix $A_{N,Y}$ to inform on the stability of the scheme.

6.2 Galerkin POD Projection

Here we study the influence of the choice of group approaches (namely, MG or GS) on the errors, over our domain $\Omega := [0, L]^2$. These comparisons were established on a single test parameter, and the training set contained 300 snapshots, containing 221 degrees of freedom each, which is quite gross, but enabled us to make quick computations. Neither increasing the degrees of freedom nor changing the test parameter mattered much to the results, and therefore we present these ones.

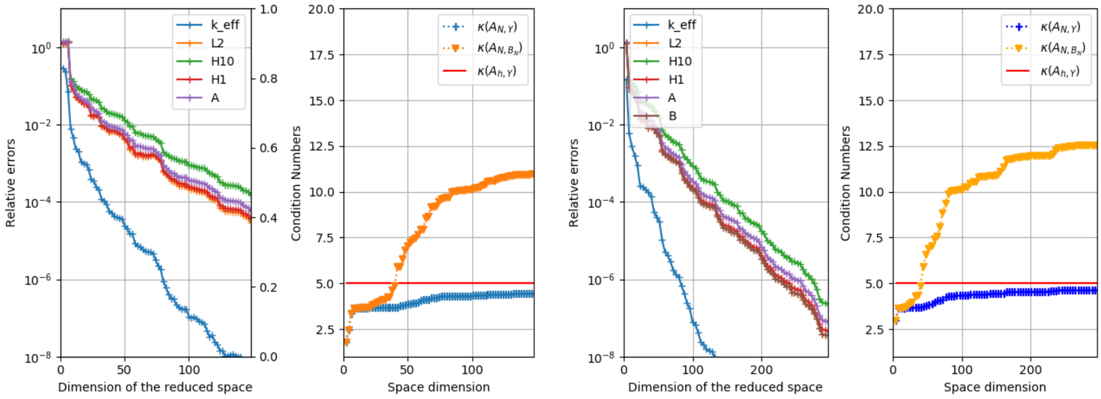

Figure 6.1: (Left to right) : Errors for different norms between the high fidelity FE solution and the reduced basis solution, for the MG approach, with a Galerkin projection on the POD-basis V_N ; Conditioning of the A_N matrix resulting from that projection. Idem for the GS approach

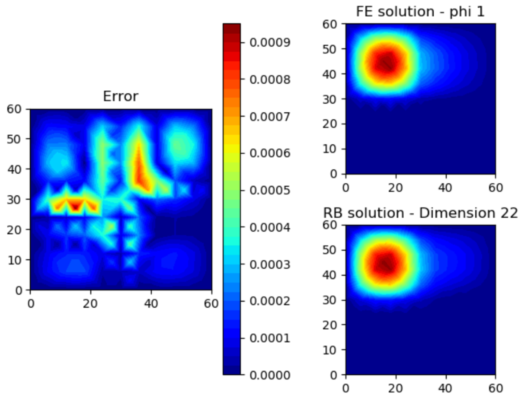

Figure 6.2: Example of the error on the eigenvector for the MG Galerkin approach: left : Plot of the difference of the FE and RB eigenvector on a gross mesh containing 221 degrees of freedom - right : The germane eigenvector and its RB approximation of dimension 22

Figure 6.2 compare the two group-approaches for the Galerkin POD projection. A first comment would be on the dimension scale chosen : indeed, the first graph is twice as stretched as the second one, ranging from 0 to 200 instead of 0 to 100. We made that choice in order to highlights the influence of the overall information instead of the dimension of the space. As seen in section 5.1, for a given number of snapshots, the GS spaces matrices have twice as much vectors than the MG ones, resulting in a doubling of the spaces' dimensions. But the information comes from one vector. Therefore, to compare the efficiency of the two methods, it seemed more natural to take the information provided by the snapshot as a point of comparison instead of the dimension, which was artificially doubled by the GS method. This can lead to believe than even with half as many snapshots, the GS method is more precise, especially on the prediction of k_{eff}.

Taking another point of view, and looking at the dimension of the spaces, we can see that the results are very close between MG and GS. Although this could imply that both method yield similar results, given that for a given integer N, MG requires N computation whereas GS only requires half of N, we can deduce that the GS approach is generally more precise.

We also plotted the condition number of the matrix $A_{N,B_N} := B_N^{-1/2} A_N B_N^{-1/2}$, as this matrix has the same eigencouples as our problem. Therefore, the stability of the problem can be measured by looking at the inf-sup constant of this matrix. We here presented its condition number as a substitute for the inf-sup constant, although the relation between those two quantities is unclear.

Figure 6.2 represent the error made on each node of the 221 degrees of freedom mesh for the relevant test parameter used in those plots. We can see that for a RB dimension of 22, the maximum error on those nodes are less than 10^{-3}, and generally located on the boundaries of the parameter grid, i.e. on $\Omega_i \cap \Omega_j$, where the flux is expected to have the greater gradient.

6.3 Petrov Galerkin POD / Adjoint POD

We present the results for the APOD projection method. This approach was motivated by the nature of our both the problem and the adjoint problem,

$$\begin{aligned} &\text{Find } \phi_\mu \in V \text{ such that for all } v \in V, \quad a_\mu(\phi_\mu, v) = \lambda_\mu b_\mu(\phi_\mu, v), \\ &\text{Find } \phi_\mu \in V \text{ such that for all } v \in V, \quad a_\mu(v, \phi_\mu) = \lambda_\mu^* b_\mu(v, \phi_\mu), \end{aligned}$$

written in algebraic form as

$$\begin{aligned} A_{\mu,h}\underline{U} &= \lambda B_{\mu,h}\underline{U}, \\ \underline{U}^* A_{\mu,h} &= \lambda \underline{U}^* B_{\mu,h}. \end{aligned}$$

Because of this formulation, we were naturally drawn to choose the matrix V_N as the trial space matrix and V_N^* as the test space matrix.

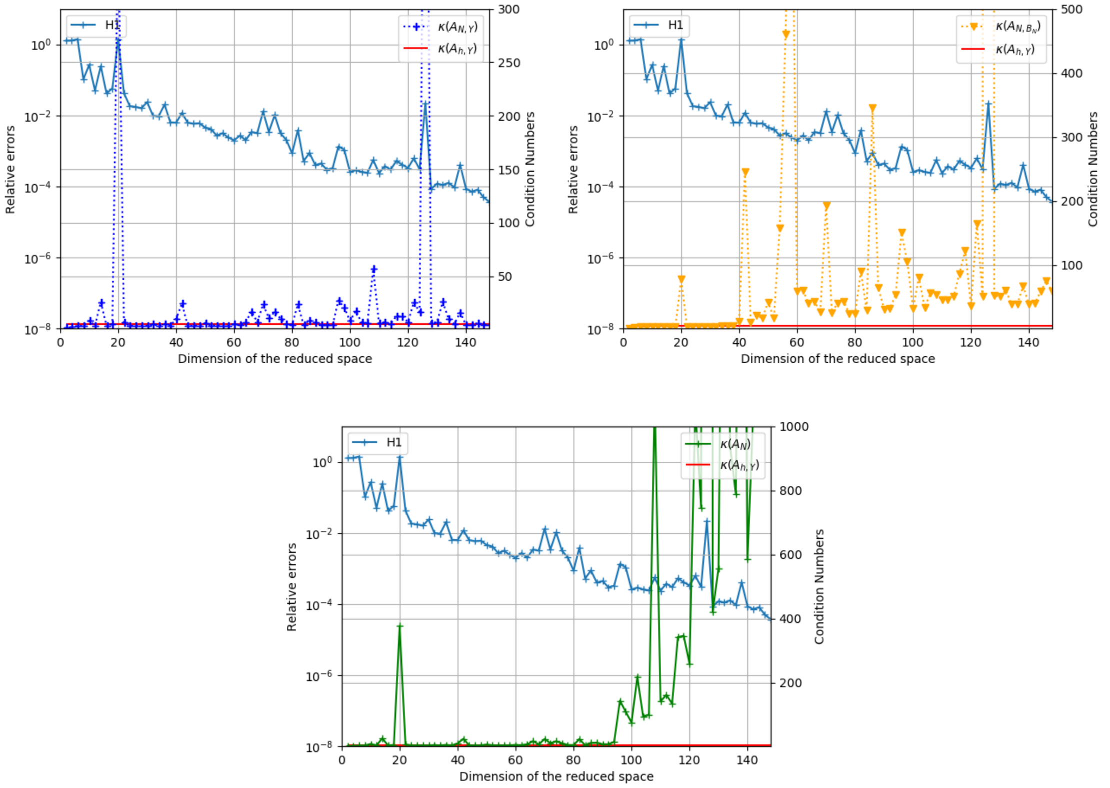

Figure 6.3: (Left to right) : Errors for the V-norm between the high fidelity FE solution and the reduced basis solution, for the MG approach, with a Petrov Galerkin projection $(V_N,\ V_n^*)$; Conditioning of the matrices resulting from that projection : Blue : $A_{N,Y}$, Orange : A_{N,B_N}, Green : A_N.

To assess which condition number is the best indicator of stability, we tried to illustrate the peaks of conditioning for three matrices : $A_{N,Y}$, A_{N,B_N} and A_N, as seen in Figure 6.3. It seems $A_{N,Y}$ is best to identify any jump in the error, as was expected by Theorem 9.

We chose not to present the GS method, as it was similar, even though it appeared to be more stabled

On a side note, the method presented in [7] appeared to be worse than the method presented here in terms of stability.

6.4 Galerkin POD "Sum" Projection

In an attempt to correct the results in the previous section, we tried to include the adjoint vector, but in a Galerkin projection framework. By denoting $\mathcal{V}_N$ and $\mathcal{V}_N^*$ the spaces spanned by the POD basis in both the snapshots and adjoint snapshots, we tried to do a Galerkin projection on the sum of these spaces, denoted by $\mathcal{V}_N + \mathcal{V}_N^*$. In a algebraic sense, this amounts to add the columns of V_N^* to V_N.

To analyze the influence of the adjoint vector, we added at each step one vector of V_N and then the corresponding adjoint vector of V_N^* that was orthonormalized by a *revealing rang*

QR decomposition. This allowed us to plot the diagonal of "R" in the QR decomposition, thus revealing the numerical rank at each step of our algorithm.

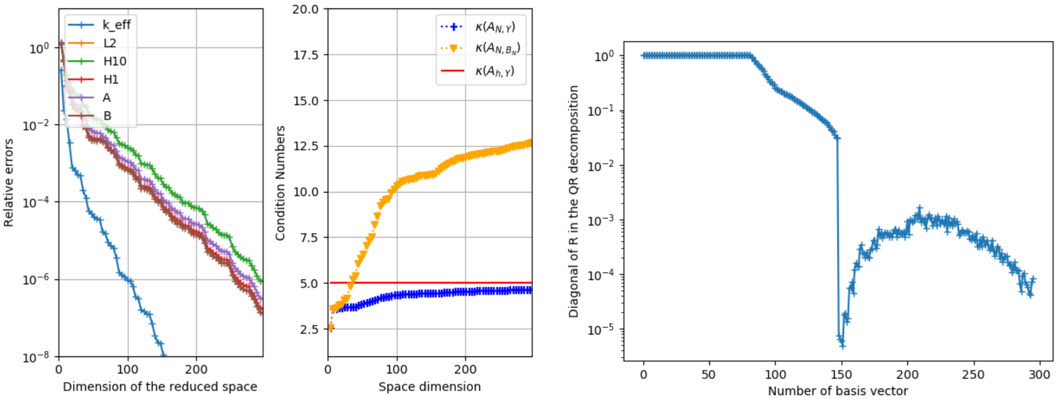

Figure 6.4: Left : Error on the Petrov Galerkin sum MG-approach - Right : Diagonal of R revealing numerical rank

A shown in Figure 6.4, this method seems as stable as the simple Galerkin projection with similar. It also illustrate the influence of the added adjoint vector as seen on the right of Figure 6.4, where the diagonal of R in the QR decomposition is shown. As the diagonal terms go to zero, we say that their contribution to the rank of the matrix decreases : hence the "numerical" rank is the number of basis vector for which the diagonal is above a certain threshold. We clearly see that the contribution to the rank is split in two, denoting the separation of V_N vectors and V_N^* vectors.

Conclusion

On the symmetric problem, as the *a priori* analysis was quite thorough, the main purpose of this study was to confirm that reduction methods were indeed a suitable method to approximate the Finite Element. This was confirmed for all three reduced basis approaches. We were able to cut computational cost further in our study by using a *parameter-independent* norm for our projection basis without altering much the accuracy of the solutions, which was not expected in the *a priori* analysis. In addition, the greedy algorithm outperformed our expectations in the simple "toy problem" that we considered.

On the non-symmetric problem, the numerical simulation serves as intuition builders for orienting the *a priori* analysis. We only focuses in POD technique due to the multiple possibilities presented : on the definition of the reduced POD space (whether to separate the groups or not) and on the choice of projection, for instance Galerkin or Petrov Galerkin projection and including an adjoint space or not. We tested three possible methods, and two turned out to be stable and precise while one remained unstable. We also presented a stability estimator in the condition number of a certain matrix, used to assert the stability of a system, and empirically verify its utility.

Further extensions could be possible, regarding the sampling of parameters [8], a finer partition of the domain Ω and a more complex geometry. In particular, the implementation of computable greedy algorithm for more complex problem, and a more efficient way to compute *a posteriori estimator* are a direct extension to this paper. In terms of stability, a deeper study on the relevance of the condition number, especially on Petrov Galerkin framework and weak coercive forms, could solidify the stability discussion made in this paper.

Bibliography

[1] G. Allaire, X. Blanc, and et.al. Transport et diffusion. 2015.

[2] I. Babuska and J. Osborn. Handbook of numerical analysis, vol ii, part 1, 1991.

[3] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[4] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159. Springer Science & Business Media, 2013.

[5] P. German and J. C. Ragusa. Reduced-order modeling of parameterized multi-group diffusion k-eigenvalue problems. *Annals of Nuclear Energy*, 134:144–157, 2019.

[6] L. Giret. *Numerical Analysis of a Non-Conforming Domain Decomposition for the Multigroup SPN Equations.* PhD thesis, 2018.

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

[8] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced basis methods for partial differential equations: an introduction*, volume 92. Springer, 2015.