Mathematical Analysis of Robustness of Two-Level Domain Decomposition Methods with respect to Inexact Coarse Solves

F. Nataf

Mathematical Analysis of Robustness of Two-Level Domain Decomposition Methods with respect to Approximate Coarse Solves

F. Nataf[1]

[1]Laboratoire J.L. Lions, UPMC, CNRS, Equipe LJL-INRIA Alpines,
nataf@jll.math.upmc.fr, Paris, France

November 8, 2017

Contents

Abstract

Convergence of domain decomposition methods rely heavily on the efficiency of the coarse space used in the second level. The GenEO coarse space has been shown to lead to a fully robust two-level Schwarz preconditioner which scales well over multiple cores [27, 19] as has been proved rigorously in [27]. The robustness is due to its good approximation properties for problems with highly heterogeneous material parameters. It is available in the finite element packages FreeFem++ [5], Feel++ [12] and recently in Dune [1] and is implemented as a standalone library in HPDDM [6]. But the coarse component of the preconditioner can ultimately become a bottleneck if the number of subdomains is very large and exact solves are used. It is therefore interesting to consider the effect of approximate coarse solves. In this paper, robustness of GenEO methods is analyzed with respect to approximate coarse solves. Interestingly, the GenEO-2 method introduced in [3] has to be modified in order to be able to prove its robustness in this context.

TBD

1 Introduction

Convergence of domain decomposition methods rely heavily on the efficiency of the coarse space used in the second level, see [9, 15, 11] and references therein. The GenEO coarse space has been shown to lead to a fully robust two-level Schwarz preconditioner which scales well over multiple cores [27, 19] as has been proved rigorously in [27]. The robustness is due to its good approximation properties for problems with highly heterogeneous material parameters. It is available in the finite element packages FreeFem++ [5], Feel++ [12] and recently in Dune [1] and is implemented as a standalone library in HPDDM [6]. But the coarse component of the preconditioner can ultimately become a bottleneck if the number of subdomains is very large and exact solves are used. It is therefore interesting to consider the effect of approximate coarse solves. In this paper, robustness of GenEO methods is analyzed with respect to approximate coarse solves. Interestingly, the GenEO-2 method introduced in [3] has to be modified in order to be able to prove its robustness in this context. In the context of domain decomposition methods, the robustness of the BDDC w.r.t. approximate coarse solves has been studied in [16, 17] and in [8]. We focus here on GenEO methods.

The general framework of our work is the following. Let M^{-1} be a preconditioner enhanced by a second level correction based on a rectangular matrix Z whose columns are a basis of a coarse space V_0. The coarse space correction is

$$Z(Z^T\,A\,Z)^{-1}Z^T\,, \tag{1}$$

and the coarse operator is defined by

$$E := Z^T\,A\,Z\,. \tag{2}$$

Let M^{-1} denote a one-level preconditioner, the two-level method is defined by:

$$M_2^{-1} := Z\,E^{-1}\,Z^T + (I_d - Z\,E^{-1}\,Z^T A)M^{-1}(I_d - AZ\,E^{-1}\,Z^T)\,,$$

see the balancing domain decomposition method by J. Mandel [7] and also the BFGS algorithm as described in e.g. [10].

We consider Geneo methods, where the coarse space V_0 spanned by the columns of Z is built from solving generalized eigenvalue problems in the subdomains. Since it is a purely parallel task with no communication involved, this part of the computation is not penalizing parallelism. Actually, in strong scaling experiments where the number of degrees of freedom of subdomains is smaller and smaller the scaling of this task is perfect. On the other hand, as the size of matrix $Z^T\,A\,Z$ typically increases linearly with the number of subdomains, the solving of the corresponding linear systems for instance with a LU factorization becomes a bottleneck in two-level domain decomposition methods. It is therefore interesting to estimate the robustness of the modified two-level method when in (2) the operator E is approximated by some operator $\tilde{E}$:

$$\boxed{\tilde{E} \simeq E\,,}$$

since it paves the way to approximate coarse solves and three or more level methods. More precisely, formula (2) is modified and the preconditioner we study is defined by:

$$\tilde{M}_2^{-1} := Z\,\tilde{E}^{-1}\,Z^T + (I_d - Z\,\tilde{E}^{-1}\,Z^T A)M^{-1}(I_d - AZ\,\tilde{E}^{-1}\,Z^T)\,.$$

2 Basic definitions

The problem to be solved is defined via a variational formulation on a domain $\Omega \subset \mathbb{R}^d$ for $d \in \mathbb{N}$:

$$\text{Find } u \in V \text{ such that : } a_\Omega(u,v) = l(v)\,, \quad \forall v \in V\,,$$

where V is a Hilbert space of functions from Ω with real values. The problem we consider is given through a symmetric positive definite bilinear form a_Ω that is defined in terms of an integral over any open set $\omega \subset \Omega$. Typical examples are the Darcy equation ($\mathbf{K}$ is a diffusion tensor)

$$a_\omega(u,v) := \int_\omega \mathbf{K}\,\nabla u \cdot \nabla v\,dx\,,$$

or the elasticity system ($\boldsymbol{C}$ is the fourth-order stiffness tensor and $\boldsymbol{\varepsilon}(\boldsymbol{u})$ is the strain tensor of a displacement field $\boldsymbol{u}$):

$$a_\omega(\boldsymbol{u},\,\boldsymbol{v}) := \int_\omega \boldsymbol{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\,dx\,.$$

The problem is discretized by a finite element method. Let $\mathcal{N}$ denote the set of degrees of freedom and $(\phi_k)_{k\in\mathcal{N}}$ be a finite element basis on a mesh $\mathcal{T}_h$. Let $A \in \mathbb{R}^{\#\mathcal{N}\times\#\mathcal{N}}$ be the associated finite element matrix, $A_{kl} := a_\Omega(\phi_l,\phi_k)$, $k,l \in \mathcal{N}$. For some given right hand side $\mathbf{F} \in \mathbb{R}^{\#\mathcal{N}}$, we have to solve a linear system in $\mathbf{U}$ of the form

$$A\mathbf{U} = \mathbf{F}\,.$$

Domain Ω is decomposed into N (overlapping or non overlapping) subdomains $(\Omega_i)_{1\leqslant i\leqslant N}$ so that all subdomains are a union of cells of the mesh $\mathcal{T}_h$. This decomposition induces a natural decomposition of the set of indices $\mathcal{N}$ into N subsets of indices $(\mathcal{N}_i)_{1\leqslant i\leqslant N}$:

$$\mathcal{N}_i := \{k \in \mathcal{N} \mid meas(\mathrm{supp}(\phi_k) \cap \Omega_i) > 0\}\,,\ 1 \leqslant i \leqslant N. \tag{3}$$

For all $1 \leqslant i \leqslant N$, let R_i be the restriction matrix from $\mathbb{R}^{\#\mathcal{N}}$ to the subset $R^{\#\mathcal{N}_i}$ and D_i be a diagonal matrix of size $\#\mathcal{N}_i \times \#\mathcal{N}_i$, so that we have a partition of unity at the algebraic level,

$$\sum_{i=1}^{N} R_i^T D_i R_i = I_d\,, \tag{4}$$

where $I_d \in \mathbb{R}^{\#\mathcal{N}\times\#\mathcal{N}}$ is the identity matrix.

We also define for all subdomains $1 \leqslant j \leqslant N$, $\widetilde{A}^j$, the $\#\mathcal{N}_j \times \#\mathcal{N}_j$ matrix defined by

$$\mathbf{V}_j^T \widetilde{A}^j \mathbf{U}_j := a_{\Omega_j} \left(\sum_{l \in \mathcal{N}_j} \mathbf{U}_{jl} \phi_l, \sum_{l \in \mathcal{N}_j} \mathbf{V}_{jl} \phi_l \right), \quad \mathbf{U}_j, \mathbf{V}_j \in \mathbb{R}^{\mathcal{N}_j} . \tag{5}$$

When the bilinear form a results from the variational solve of a Laplace problem, the previous matrix corresponds to the discretization of local Neumann boundary value problems. For this reason we will call it "Neumann" matrix even in a more general setting.

We also make use of two numbers k_0 and k_1 related to the domain decomposition. Let

$$k_0 := \max_{1 \leqslant i \leqslant N} \# \left\{ j \mid R_j A R_i^T \neq 0 \right\} \tag{6}$$

be the maximum multiplicity of the interaction between subdomains plus one. Let k_1 be the maximal multiplicity of subdomains intersection, i.e. the largest integer m such that there exists m different subdomains whose intersection has a non zero measure.

Let $\tilde{P}_0$ be defined as:

$$\boxed{\tilde{P}_0 := Z \tilde{E}^{-1} Z^T A ,} \tag{7}$$

the operator $\tilde{P}_0$ is thus an approximation to the A-orthogonal projection on V_0

$$\boxed{P_0 := Z E^{-1} Z^T A}$$

which corresponds to an exact coarse solve.

Note that although $\tilde{P}_0$ is not a projection it has the same kernel and range as P_0:

Lemma 2.1 *We have*

$$\ker P_0 = \ker \tilde{P}_0 = V_0^{A\perp} \quad \textit{and} \quad \operatorname{Im} P_0 = \operatorname{Im} \tilde{P}_0 = V_0 ,$$

where $V_0^{A\perp}$ is the vector space A-orthogonal to V_0, that is when $\mathbb{R}^{\#\mathcal{N}}$ is endowed with the scalar product induced by A: $(x\,,\,y)_A := (x\,,\,Ay)$.

Proof First note that the kernel of $\tilde{P}_0$ contains $\ker Z^T A$. On the other hand, we have:

$$\tilde{P}_0 x = Z \tilde{E}^{-1} Z^T A x = 0 \Rightarrow (Z \tilde{E}^{-1} Z^T A x, A x) = (\tilde{E}^{-1} Z^T A x, Z^T A x) = 0 .$$

Since $\tilde{E}$ is SPD, it means that $Z^T A x = 0$, that is $x \in \ker Z^T A$. We have thus $\ker \tilde{P}_0 = \ker Z^T A$. Note that

$$Z^T A x = 0 \Leftrightarrow \forall y \ \ (A x\,,\, Z y) = 0 \Leftrightarrow x \in V_0^{A\perp} .$$

As for the image of $\tilde{P}_0$, since the last operation in its definition is the multiplication by the matrix Z we have $\operatorname{Im} P_0 \subset V_0$. Conversely, let $y \in V_0$, there exists β such that $y = Z\beta$. It is easy to check that $y = \tilde{P}_0(Z\,(ZAZ)^{-1} \tilde{E}\,\beta)$. Thus, $\operatorname{Im} \tilde{P}_0 = V_0$.
The same arguments hold if $\tilde{E}$ is replaced by E. Thus, $\tilde{P}_0$ and P_0 have the same kernel and image. ■

3 Approximate Coarse Solves for GenEO

The GenEO coarse space was introduced in [13] and is defined as follows:

Definition 3.1 (Generalized Eigenvalue Problem for GenEO) *For each subdomain $1 \leqslant j \leqslant N$, we introduce the generalized eigenvalue problem*

$$\begin{array}{c} \textit{Find } (\mathbf{V}_{jk}, \mu_{jk}) \in \mathbb{R}^{\#\mathcal{N}_j} \backslash \{0\} \times \mathbb{R} \textit{ such that} \\ D_j R_j \, A \, R_j^T D_j \mathbf{V}_{jk} = \mu_{jk} \widetilde{A}^j \mathbf{V}_{jk} \, . \end{array} \tag{8}$$

Let $\mu > 0$ be a user-defined threshold, we define $V_{geneo}^{\mu} \subset \mathbb{R}^{\#\mathcal{N}}$ as the vector space spanned by the family of vectors $(R_j^T D_j \mathbf{V}_{jk})_{\mu_{jk} > \mu \, , 1 \leqslant j \leqslant N}$ corresponding to eigenvalues larger than μ.

Let $\tilde{\pi}_j$ be the projection from $\mathbb{R}^{\#\mathcal{N}_j}$ on $\mathrm{Span}\{\mathbf{V}_{jk} | \, \mu_{jk} > \mu\}$ parallel to $\mathrm{Span}\{\mathbf{V}_{jk} | \, \mu_{jk} \leqslant \mu\}$.

In this section, Z denotes a rectangular matrix whose columns are a basis of the coarse space V_{geneo}^{μ} defined in Definition (3.1). The dimension of Z is $\#\mathcal{N} \times \#\mathcal{N}_0$. The GenEO preconditioner with approximage coarse solve reads:

$$M_{GenEOACS}^{-1} := Z \, \tilde{E}^{-1} \, Z^T + (I_d - \tilde{P}_0) \, (\sum_{i=1}^{N} R_i^T \, (R_i \, A \, R_i^T)^{-1} \, R_i) \, (I_d - \tilde{P}_0^T) \, . \tag{9}$$

The study the spectrum of $M_{GenEOACS}^{-1} A$ is based on the Fictitious Space lemma, see [2] for more details. For this purpose, we introduce

$$H_D := \mathbb{R}^{\#\mathcal{N}_0} \times \Pi_{i=1}^{N} \mathbb{R}^{\#\mathcal{N}_i}$$

be endowed with the following bilinear form

$$\begin{array}{rcl} \tilde{b} : H_D \times H_D & \longrightarrow & \mathbb{R} \\ (\, (\mathbf{U}_0 \, , \, (\mathbf{U}_i)_{1 \leqslant i \leqslant N}), \, (\mathbf{V}_0 \, , \, (\mathbf{V}_i)_{1 \leqslant i \leqslant N}) \,) & \longmapsto & (\tilde{E} \mathbf{U}_0, \, \mathbf{V}_0) + (R_i \, A \, R_i^T \, \mathbf{U}_i, \, \mathbf{V}_i) \end{array} \tag{10}$$

and $\widetilde{\mathcal{R}} : H_D \longrightarrow H$ is defined by

$$\widetilde{\mathcal{R}}(\mathcal{U}) := Z \, \mathbf{U}_0 + (I_d - \tilde{P}_0) \, \sum_{i=1}^{N} R_i^T \mathbf{U}_i \, . \tag{11}$$

Recall that had we had used an exact coarse space solve, we would have introduced:

$$\mathcal{R}(\mathcal{U}) := Z \, \mathbf{U}_0 + (I_d - P_0) \, \sum_{i=1}^{N} R_i^T \mathbf{U}_i \, . \tag{12}$$

Note that we have

$$\widetilde{\mathcal{R}}(\mathcal{U}) = \mathcal{R}(\mathcal{U}) + (P_0 - \tilde{P}_0) \, \sum_{i=1}^{N} R_i^T \mathbf{U}_i \, .$$

It can be chekced that the resulting preconditioner with approximate coarse solve $\widetilde{\mathcal{R}} \, \widetilde{B}^{-1} \, \widetilde{\mathcal{R}}^T$ is actually equal to $M_{GenEOACS}^{-1}$, see (9). In order to apply the fictitious space Lemma, three assumptions have to be checked.

• $\widetilde{\mathcal{R}}$ is onto.
Let $\mathbf{U} \in H$, we have $\mathbf{U} = \tilde{P}_0\mathbf{U} + (I_d - \tilde{P}_0)\,\mathbf{U}$. By Lemma 2.1, $\tilde{P}_0\mathbf{U} \in V_0$ so that there exists $\mathbf{U}_0 \in \mathbb{R}^{\#\mathcal{N}_0}$ such that $\tilde{P}_0\mathbf{U} = Z\mathbf{U}_0$. Thus we have

$$\mathbf{U} = Z\mathbf{U}_0 + (I_d - \tilde{P}_0)\sum_{i=1}^{N} R_i^T D_i R_i \mathbf{U} = \widetilde{\mathcal{R}}(\mathbf{U}_0, (D_i R_i \mathbf{U})_{1\leqslant i\leqslant N}))$$

• Continuity of $\widetilde{\mathcal{R}}$
We have to estimate a constant c_R such that for all $\mathcal{U} = (\mathbf{U}_0\,,\,(\mathbf{U}_i)_{1\leqslant i\leqslant N}) \in H$, we have:

$$a(\widetilde{\mathcal{R}}(\mathcal{U}), \widetilde{\mathcal{R}}(\mathcal{U})) \leqslant c_R\,\tilde{b}(\mathcal{U}\,,\,\mathcal{U})\,.$$

Let δ be some positive number and Using that the image of $P_0 - \tilde{P}_0$ is a-orthogonal to the image of $I_d - P_0$, Cauchy-Schwarz inequality and the a-orthogonality of the projection $I_d - P_0$, we have:

$$\begin{aligned}
a(\widetilde{\mathcal{R}}(\mathcal{U}), \widetilde{\mathcal{R}}(\mathcal{U})) &= \|\mathcal{R}(\mathcal{U}) + (P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&= \|\mathcal{R}(\mathcal{U})\|_A^2 + 2\,a(Z\,\mathbf{U}_0 + (I_d - P_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\,,\,(P_0 - \tilde{P}_0)\sum_{i=1}^N R_i^T\mathbf{U}_i) \\
&\quad +\|(P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&= \|\mathcal{R}(\mathcal{U})\|_A^2 + 2\,a(Z\,\mathbf{U}_0\,,\,(P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i) \\
&\quad +\|(P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&\leqslant \|\mathcal{R}(\mathcal{U})\|_A^2 + \delta\|Z\,\mathbf{U}_0\|_A^2 + \tfrac{1}{\delta}\,\|(P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&\quad +\|(P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&\leqslant \|Z\,\mathbf{U}_0\|_A^2 + \|\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 + \delta\|Z\,\mathbf{U}_0\|_A^2 \\
&\quad +(1 + \tfrac{1}{\delta})\,\|(P_0 - \tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&\leqslant (1+\delta)\|Z\,\mathbf{U}_0\|_A^2 + (1 + \|P_0 - \tilde{P}_0\|_A^2(1 + \tfrac{1}{\delta}))\|\textstyle\sum_{i=1}^N R_i^T\mathbf{U}_i\|_A^2 \\
&\leqslant (1+\delta)\lambda_{max}(E\tilde{E}^{-1})(\tilde{E}\,\mathbf{U}_0, \mathbf{U}_0) + (1 + \|P_0 - \tilde{P}_0\|_A^2(1 + \tfrac{1}{\delta}))\,k_0\textstyle\sum_{i=1}^N \|R_i^T\mathbf{U}_i\|_A^2 \\
&\leqslant \max\Big(\,(1+\delta)\,\lambda_{max}(E\tilde{E}^{-1}),\,[1 + \|P_0 - \tilde{P}_0\|_A^2(1 + \tfrac{1}{\delta})]\,k_0\Big)\;\tilde{b}(\mathcal{U}\,,\,\mathcal{U})\,.
\end{aligned}$$

It is possible to minimize over δ the factor in front of $\tilde{b}(\mathcal{U}\,,\,\mathcal{U})$ using the

Lemma 3.1 *Let c, d, α and β be positive constant, we have*

$$\min_{\delta>0}\max(c + \alpha\delta, d + \beta\delta^{-1}) = \frac{d + c + \sqrt{(d-c)^2 + 4\alpha\beta}}{2}\,.$$

Proof The optimal value for δ corresponds to the equality $c + \alpha\delta = d + \beta\delta^{-1}$. ■

Let

$$\epsilon_A := \|P_0 - \tilde{P}_0\|_A = \|Z^T(\,(Z^T\,A\,Z)^{-1} - \tilde{E}^{-1})Z^T A\|_A, \tag{13}$$

the formula of Lemma 3.1 yields

$$c_R := \frac{k_0(1+\epsilon_A^2) + \lambda_{max}(E\tilde{E}^{-1}) + \sqrt{(k_0(1+\epsilon_A^2) - \lambda_{max}(E\tilde{E}^{-1}))^2 + 4\lambda_{max}(E\tilde{E}^{-1})k_0(\epsilon_A^2+1)}}{2}\,. \tag{14}$$

Actually, ϵ_A can be expressed in term of the minimal eigenvalue of $E\tilde{E}^{-1}$.

Lemma 3.2 *Other formula for ϵ_A:*

$$\epsilon_A = \sup_{\mathbf{U}_0 \in \mathbb{R}^{\#\mathcal{N}_0}} \frac{(E(E^{-1}-\tilde{E}^{-1})E\mathbf{U}_0, \mathbf{U}_0)}{(E\mathbf{U}_0, \mathbf{U}_0)} = \max(|1-\lambda_{min}(E\tilde{E}^{-1})|\,,\, |1-\lambda_{max}(E\tilde{E}^{-1})|)\,.$$

Proof Since $P_0 - \tilde{P}_0$ is A-symmetric, its norm is also given by

$$\epsilon_A = \sup_{\mathbf{U}} \frac{|((P_0-\tilde{P}_0)\mathbf{U}\,,\,\mathbf{U})_A|}{\|\mathbf{U}\|_A^2}$$

We can go further by using the fact that P_0 is a A-orthogonal and that P_0 and $\tilde{P}_0$ have the same kernels and images:

$$\begin{aligned}
\epsilon_A &= \sup_{\mathbf{U}} \frac{|((P_0-\tilde{P}_0)(P_0\,\mathbf{U} + (I_d-P_0)\,\mathbf{U})\,,\, P_0\,\mathbf{U} + (I_d-P_0)\mathbf{U})_A|}{\|P_0\mathbf{U}\|_A^2 + \|(I_d-P_0)\mathbf{U}\|_A^2} \\
&= \sup_{\mathbf{U}} \frac{|((P_0-\tilde{P}_0)P_0\,\mathbf{U}\,,\, P_0\,\mathbf{U})_A|}{\|P_0\mathbf{U}\|_A^2 + \|(I_d-P_0)\mathbf{U}\|_A^2} = \sup_{\mathbf{U}} \frac{|((P_0-\tilde{P}_0)P_0\,\mathbf{U}\,,\, P_0\,\mathbf{U})_A|}{\|P_0\mathbf{U}\|_A^2} \\
&= \sup_{\mathbf{U}\in V_0} \frac{|((P_0-\tilde{P}_0)\,\mathbf{U}\,,\,\mathbf{U})_A|}{\|\mathbf{U}\|_A^2} = \sup_{\mathbf{U}_0\in\mathbb{R}^{\#\mathcal{N}_0}} \frac{|(E(E^{-1}-\tilde{E}^{-1})E\mathbf{U}_0, \mathbf{U}_0)|}{(E\mathbf{U}_0, \mathbf{U}_0)} \\
&= \sup_{\mathbf{U}_0\in\mathbb{R}^{\#\mathcal{N}_0}} |1 - \frac{(\tilde{E}^{-1}E\mathbf{U}_0, E\mathbf{U}_0)}{(E\mathbf{U}_0, \mathbf{U}_0)}|\,.
\end{aligned}$$

■

This means that formula (14) for c_R can be expressed explicitely in terms of k_0 and of the minimal and maximal eigenvalue of $\tilde{E}^{-1}E$.

- Stable decomposition

Let $\mathbf{U} \in H$ be decomposed as follows:

$$\begin{aligned}
\mathbf{U} &= P_0\mathbf{U} + (I_d-P_0)\mathbf{U} = P_0\mathbf{U} + (I_d-P_0)\textstyle\sum_{j=1}^N R_j^T D_j R_j \mathbf{U} \\
&= P_0\mathbf{U} + (I_d-P_0)\textstyle\sum_{j=1}^N R_j^T D_j (I_d-\tilde{\pi}_j) R_j\mathbf{U} + \underbrace{(I_d-P_0)\sum_{j=1}^N R_j^T D_j \tilde{\pi}_j\, R_j\mathbf{U}}_{=0} \\
&= \underbrace{P_0\mathbf{U} + (\tilde{P}_0-P_0)\sum_{j=1}^N R_j^T D_j (I_d-\tilde{\pi}_j) R_j\mathbf{U}}_{:=F\mathbf{U}\in V_0} + (I_d-\tilde{P}_0)\textstyle\sum_{j=1}^N R_j^T D_j (I_d-\tilde{\pi}_j) R_j\mathbf{U}\,.
\end{aligned}$$

Let $\mathbf{U}_0 \in \mathbb{R}^{\#\mathcal{N}_0}$ be such that $Z\mathbf{U}_0 = F\mathbf{U}$, we choose the following decomposition:

$$\mathbf{U} = \widetilde{\mathcal{R}}(\mathbf{U}_0, (D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U})_{1\leqslant j\leqslant N})\,.$$

The stable decomposition consists in estimating a constant $c_T > 0$ such that:

$$c_T\,[(\tilde{E}\,\mathbf{U}_0, \mathbf{U}_0) + \sum_{j=1}^N (R_j A R_j^T D_j (I_d-\tilde{\pi}_j) R_j\mathbf{U}, D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U})] \leqslant a(\mathbf{U},\mathbf{U})\,. \tag{15}$$

Since the second term in the left hand side is the same as in the exact coarse solve method, we have (see [2], page 177, Lemma 7.15):

$$\sum_{j=1}^{N}(R_j A R_j^T D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U}, D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U}) \leqslant k_1\,\tau\,a(\mathbf{U},\mathbf{U})\,. \tag{16}$$

We now focus on the first term of the left hand side of (15). Let δ be some positive number, using again (16), the following auxiliary result holds:

$$\begin{aligned}
\|F\mathbf{U}\|_A^2 &\leqslant (1+\delta)\|P_0\mathbf{U}, P_0\mathbf{U}\|_A^2 \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\textstyle\sum_{j=1}^{N} R_j^T D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U}\|_A^2 \\
&\leqslant (1+\delta)(A\mathbf{U},\mathbf{U}) \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2\,\|\textstyle\sum_{j=1}^{N} R_j^T D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U}\|_A^2 \\
&\leqslant (1+\delta)(A\mathbf{U},\mathbf{U}) \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2\;k_0\textstyle\sum_{j=1}^{N}\|R_j^T D_j(I_d-\tilde{\pi}_j)R_j\mathbf{U}\|_A^2 \\
&\leqslant \Big(1+\delta+(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2 k_0 k_1\,\tau\Big)\;a(\mathbf{U},\mathbf{U})
\end{aligned}$$

The best possible value for δ is

$$\delta := \epsilon_A\sqrt{k_0\,k_1\,\tau}\,.$$

Hence, we have:

$$(Z^T\,A\,Z\mathbf{U}_0\,,\,\mathbf{U}_0) = \|F\mathbf{U}\|_A^2 \leqslant (1+\epsilon_A\sqrt{k_0\,k_1\,\tau})^2 a(\mathbf{U},\mathbf{U})\,. \tag{17}$$

Thus, we have:

$$\begin{aligned}
(\tilde{E}\,\mathbf{U}_0\,,\,\mathbf{U}_0) &= (\tilde{E}\,E^{-1/2}E^{1/2}\mathbf{U}_0\,,\,E^{-1/2}E^{1/2}\mathbf{U}_0) = (E^{-1/2}\tilde{E}\,E^{-1/2}E^{1/2}\mathbf{U}_0\,,\,E^{1/2}\mathbf{U}_0) \\
&\leqslant \lambda_{max}(E^{-1/2}\tilde{E}\,E^{-1/2})\,(E^{1/2}\mathbf{U}_0\,,\,E^{1/2}\mathbf{U}_0) \\
&= \lambda_{max}(E^{-1}\tilde{E})(Z^T\,A\,Z\mathbf{U}_0\,,\,\mathbf{U}_0) \\
&\leqslant \lambda_{max}(E^{-1}\tilde{E})(1+\epsilon_A\sqrt{k_0\,k_1\,\tau})^2 a(\mathbf{U},\mathbf{U})\,.
\end{aligned}$$

This last estimate along with (16) prove that in (15), it is possible to take

$$c_T = \frac{\lambda_{min}(E\tilde{E}^{-1})}{(1+\epsilon_A\sqrt{k_0\,k_1\,\tau})^2 + k_1\tau}\,. \tag{18}$$

Overall, with c_T given by (18) and c_R by (14), we have proved the following spectral estimate:

$$c_T \leqslant \lambda(M_{GenEOACS}^{-1}\,A) \leqslant c_R\,. \tag{19}$$

Constants c_T and c_R are stable with respect to ϵ_A and the spectrum of $E\tilde{E}^{-1}$ so that (19) proves the stability of preconditioner $M_{GenEOACS}^{-1}$ w.r.t. approximate solves.

4 Approximate Coarse Solves for GenEO2

The GenEO-2 coarse space construction was introduced in [4, 3] , see [2] also § 7.7, page 186. It is motivated by domain decomposition methods for which the local solves are not necessarily Dirichlet solves e.g. discretization of Robin boundary value problems, see [14]. We were not able to prove the robustness of the GenEO-2 coarse space with respect to approximate coarse solves when used in the original GenEO-2 preconditioner (39), see remark 4.2. For this reason, we study here a slight modification of the preconditioner, eq. (27), for which we prove robustness.

For all subdomains $1 \leqslant i \leqslant N$, let B_i be a matrix of size $\#\mathcal{N}_i \times \#\mathcal{N}_i$, which comes typically from the discretization of boundary value local problems using optimized transmission conditions or Neumann boundary conditions. Recall that by construction matrix $D_i\, R_i A R_i^T D_i$ is symmetric positive-semi definite and we make the extra following assumption:

Assumption 4.1 *For all subdomains $1 \leqslant i \leqslant N$, matrix B_i is symmetric positive semi-definite and either of the two conditions holds*

- *B_i is definite,*
- *$B_i = \widetilde{A}^i$ and $D_i\, R_i A R_i^T D_i$ is definite.*

In order to ease the redaction, we first consider the case where B_i is definite. The other case will be treated in Remark 4.3. We recall the coarse space defined in [4, 3, 2]. Let γ and τ be two user defined thresholds. We introduce two generalized eigenvalue problems which by Assumption 4.1 are regular.

Definition 4.1 (Generalized Eigenvalue Problem for the lower bound) *For each subdomain $1 \leqslant j \leqslant N$, we introduce the generalized eigenvalue problem*

$$\begin{array}{c} \text{Find } (\mathbf{V}_{jk}, \lambda_{jk}) \in \mathbb{R}^{\#\mathcal{N}_j} \backslash \{0\} \times \mathbb{R} \text{ such that} \\ \widetilde{A}^j \mathbf{V}_{jk} = \lambda_{jk} B_j \mathbf{V}_{jk} \ . \end{array} \tag{20}$$

Let $\tau > 0$ be a user-defined threshold and $\tilde{\pi}_j$ be the projection from $\mathbb{R}^{\#\mathcal{N}_j}$ on $V_{j\tau} = Span\{\mathbf{V}_{jk} | \lambda_{jk} < \tau\}$ parallel to $Span\{\mathbf{V}_{jk} | \lambda_{jk} \geqslant \tau\}$. We define $V^\tau_{j,geneo} \subset \mathbb{R}^{\#\mathcal{N}}$ as the vector space spanned by the family of vectors $(R_j^T D_j \mathbf{V}_{jk})_{\lambda_{jk}<\tau}$ corresponding to eigenvalues smaller than τ. Let V^τ_{geneo} be the vector space spanned by the collection over all subdomains of vector spaces $(V^\tau_{j,geneo})_{1\leqslant j\leqslant N}$.

Definition 4.2 (Generalized Eigenvalue Problem for the upper bound) *For each subdomain $1 \leqslant i \leqslant N$, we introduce the generalized eigenvalue problem*

$$\begin{array}{c} \text{Find } (\mathbf{U}_{ik}, \mu_{ik}) \in \mathbb{R}^{\#\mathcal{N}_i} \backslash \{0\} \times \mathbb{R} \text{ such that} \\ D_i R_i A R_i^T D_i \mathbf{U}_{ik} = \mu_{ik} B_i\, \mathbf{U}_{ik} \ . \end{array} \tag{21}$$

Let $\gamma > 0$ be a user-defined threshold, we define $V^\gamma_{i,geneo} \subset \mathbb{R}^{\#\mathcal{N}}$ as the vector space spanned by the family of vectors $(R_i^T D_i \mathbf{U}_{ik})_{\mu_{ik}>\gamma}$ corresponding to eigenvalues larger than γ. Let V^γ_{geneo} be the vector space spanned by the collection over all subdomains of vector spaces $(V^\gamma_{j,geneo})_{1\leqslant j\leqslant N}$.

Now, let ξ_i denote the B_i-orthogonal projection from $\mathbb{R}^{\#\mathcal{N}_i}$ on

$$V_{i\,\gamma} := \mathrm{Span}\,\{\mathbf{U}_{ik}\,|\,\gamma < \mu_{ik}\}$$

parallel to

$$W_{i\,\gamma} := \mathrm{Span}\,\{\mathbf{U}_{ik}\,|\,\gamma \geqslant \mu_{ik}\}\ .$$

By Lemma 7.6, page 167 in [2] , we have:

Lemma 4.1 (Intermediate Lemma for GenEO-2) *For all subdomains $1 \leqslant i \leqslant N$ and $\mathbf{U}_i \in \mathbb{R}^{\mathcal{N}_i}$, we have:*

$$\tau\ ((I_d - \tilde{\pi}_i)\mathbf{U}_i)^T B_j (I_d - \tilde{\pi}_i)\mathbf{U}_i \leqslant\ \mathbf{U}_i^T \widetilde{A}^i \mathbf{U}_i\,, \tag{22}$$

and

$$\left(R_i^T D_i (I_d - \xi_i)\mathbf{U}_i\right)^T A\, R_i^T D_i (I_d - \xi_i)\mathbf{U}_i) \leqslant \gamma\ (B_i (I_d - \xi_i)\mathbf{U}_i, (I_d - \xi_i)\mathbf{U}_i)\,. \tag{23}$$

Let b_i be the bilinear form related to B_i (i.e. $b_i(\mathbf{U}_i, \mathbf{V}_i) := (B_i\mathbf{U}_i, \mathbf{V}_i)$), we note that ξ_i is actually a b_i-orthogonal projection.

The coarse space V_0 built from the above generalized eigenvalues:

$$V_0 := V_{geneo}^{\tau} \bigoplus V_{geneo}^{\gamma}\,.$$

is spanned by the columns of a full rank rectangular matrix $Z = R_0^T$ with $\#\mathcal{N}_0$ columns. Projection P_0 and its approximation $\tilde{P}_0$ are defined by the same formula as above, see (7).

In addition to Lemma 4.1, we have the following

Lemma 4.2 *For $1 \leqslant j \leqslant N$, let us introduce the B_j-orthogonal projection p_j from $\mathbb{R}^{\#\mathcal{N}_j}$ on*

$$V_{j,\tau\gamma} := V_{j,\tau} \oplus V_{j,\gamma}\,.$$

Then for all $\mathbf{U}_j \in \mathbb{R}^{\#\mathcal{N}_j}$, we have:

$$\tau\,(B_j(I_d - p_j)\mathbf{U}_j, (I_d - p_j)\mathbf{U}_j) \leqslant (\tilde{A}_j\mathbf{U}_j, \mathbf{U}_j)\,.$$

Moreover, for all $\mathbf{U} \in \mathbb{R}^{\#\mathcal{N}}$, we have:

$$\tau \sum_{j=1}^{N} (B_j(I_d - p_j)R_j\mathbf{U}, (I_d - p_j)R_j\mathbf{U}) \leqslant k_1\, a(\mathbf{U}, \mathbf{U})\,.$$

Proof Let $\mathbf{U}_j \in \mathbb{R}^{\#\mathcal{N}_j}$, we have:

$$\begin{array}{rcl}
(B_j(I_d - \tilde{\pi}_j)\mathbf{U}_j, (I_d - \tilde{\pi}_j)\mathbf{U}_j) & = & (B_j(I_d - p_j + (p_j - \tilde{\pi}_j))\mathbf{U}_j, (I_d - p_j + (p_j - \tilde{\pi}_j))\mathbf{U}_j) \\
& = & \|(I_d - p_j)\mathbf{U}_j\|_{B_j}^2 + \|(p_j - \tilde{\pi}_j)\mathbf{U}_j\|_{B_j}^2 \\
& & +2\underbrace{(B_j(I_d - p_j)\mathbf{U}_j, (p_j - \tilde{\pi}_j)\mathbf{U}_j)}_{=0 \text{ since } \tilde{\pi}_j\mathbf{U}_j \in V_{j,\tau} \subset V_{j,\tau\gamma}} \\
& \geqslant & \|(I_d - p_j)\mathbf{U}_j\|_{B_j}^2 = (B_j(I_d - p_j)\mathbf{U}_j, (I_d - p_j)\mathbf{U}_j)\,.
\end{array}$$

Since we have (22):

$$\tau\,(B_j(I_d - \tilde{\pi}_j)\mathbf{U}_j, (I_d - \tilde{\pi}_j)\mathbf{U}_j) \leqslant (\tilde{A}_j\mathbf{U}_j, \mathbf{U}_j)\,,$$

the conclusion follows by summation over all subdomains. ■

The definition of the stable preconditioner is based on a pseudo inverse of B_i that we introduce now. Let b_{W_i} denote the restriction of b_i to $W_{i\,\gamma} \times W_{i\,\gamma}$ where $W_{i\,\gamma}$ is endowed with the Euclidean scalar product:

$$\begin{aligned} b_{W_i} :& W_{i\,\gamma} \times W_{i\,\gamma} \longrightarrow \mathbb{R} \\ & (\mathbf{U}_i\,,\, \mathbf{V}_i) \mapsto b_i(\mathbf{U}_i\,,\, \mathbf{V}_i)\,. \end{aligned} \tag{24}$$

By Riesz representation theorem, there exists a unique isomorphism $B_{W_i} : W_{i\,\gamma} \longrightarrow W_{i\,\gamma}$ into itself so that for all $\mathbf{U}_i\,,\, \mathbf{V}_i \in W_{i\,\gamma}$, we have:

$$b_{W_i}(\mathbf{U}_i\,,\, \mathbf{V}_i) = (B_{W_i}\,\mathbf{U}_i\,,\, \mathbf{V}_i)\,.$$

The inverse of B_{W_i} will be denoted by $B_i^\dagger$ and is given by the following formula

$$B_i^\dagger = (I_d - \xi_i) B_i^{-1}\,. \tag{25}$$

In order to check this formula, we have to show that $B_{W_i}(I_d - \xi_i)B_i^{-1} y = y$ for all $y \in W_{i\,\gamma}$. Let $z \in W_{i\,\gamma}$, using the fact that $I_d - \xi_i$ is the b_i-orthogonal projection on $W_{i\,\gamma}$, we have:

$$(B_{W_i}(I_d - \xi_i)B_i^{-1} y, z) = b_i((I_d - \xi_i)B_i^{-1} y, z) = b_i(B_i^{-1} y, z) = (y, z)\,. \tag{26}$$

Since this equality holds for any $z \in W_{i\,\gamma}$, this proves that $B_{W_i}(I_d - \xi_i)B_i^{-1} y = y$.

We study now the preconditioner given by:

Definition 4.3 (Preconditioner $M^{-1}_{GenEO2ACS}$) *Let q_i denote the orthogonal projection from $\mathbb{R}^{\#\mathcal{N}_i}$ onto $W_{i\,\gamma}$. We define the preconditioner $M^{-1}_{GenEO2ACS}$ as follows:*

$$\begin{aligned} M^{-1}_{GenEO2ACS} :=& \; Z\,\tilde{E}^{-1}\,Z^T \\ & + (I_d - \tilde{P}_0)\left(\sum_{i=1}^N R_i^T\,D_i\,q_i\,B_i^\dagger\,q_i\,D_i\;R_i\right)(I_d - \tilde{P}_0^T)\,. \end{aligned} \tag{27}$$

Remark 4.1 *Note that $q_i\,B_i^\dagger$ is actually equal to $B_i^\dagger$ but its presence shows the symmetry of the preconditioner.*

We can now define the abstract framework for the preconditioner. Let H_D be defined by

$$H_D := \mathbb{R}^{\#\mathcal{N}_0} \times \Pi_{i=1}^N W_{i\,\gamma}$$

endowed with the following bilinear form arising from local SPD matrices $(B_i)_{1 \leqslant i \leqslant N}$

$$\begin{array}{rcl} \tilde{b} : H_D \times H_D & \longrightarrow & \mathbb{R} \\ (\mathcal{U}, \mathcal{V}) & \longmapsto & b(\mathcal{U}, \mathcal{V}) := (\tilde{E}\mathbf{U}_0,\, \mathbf{V}_0) + \sum_{i=1}^N (B_i\,\mathbf{U}_i,\, \mathbf{V}_i) \end{array} \tag{28}$$

and $\widetilde{\mathcal{R}} : H_D \longrightarrow H$ is defined using operator $\tilde{P}_0$ (see eq. (7)):

$$\widetilde{\mathcal{R}}(\mathcal{U}) := Z\,\mathbf{U}_0 + (I_d - \tilde{P}_0)\sum_{i=1}^N R_i^T D_i \mathbf{U}_i\,. \tag{29}$$

Recall that had we had used an exact coarse space solve, we would have introduced:

$$\mathcal{R}(\mathcal{U}) := Z\,\mathbf{U}_0 + (I_d - P_0)\,\sum_{i=1}^{N} R_i^T D_i \mathbf{U}_i\,. \tag{30}$$

Note that we have

$$\widetilde{\mathcal{R}}(\mathcal{U}) = \mathcal{R}(\mathcal{U}) + (P_0 - \tilde{P}_0)\,\sum_{i=1}^{N} R_i^T D_i U_i\,.$$

It can be checked that the resulting preconditioner with approximate coarse solve $M^{-1}_{GenEO2ACS} := \widetilde{\mathcal{R}}\,\widetilde{B}^{-1}\,\widetilde{\mathcal{R}}^T$ is actually $M^{-1}_{GenEO2ACS}$ defined in (27). Indeed, we have:

$$\widetilde{\mathcal{R}}^T\mathbf{V} = (Z^T\mathbf{V}, (q_i D_i R_i (I_d - \tilde{P}_0^T)\mathbf{V})_{1\leqslant i\leqslant N})$$

Auxiliary results on GEVP Beware, in this paragraph, A and B have nothing to do with the global problem to be solved:

Lemma 4.3 *Let A be a symmetric positive semi definite matrix and B be a symmetric positive definite matrix. We consider the generalized eigenvalue problem:*

$$A\mathbf{U} = \lambda B\mathbf{U}\,.$$

The generalized eigenvectors and eigenvalues are denoted by $(\mathbf{U}_k, \lambda_k)_{k\geqslant 1}$. Let τ be a positive number. We define

$$V_\tau := \mathit{Span}\{\mathbf{U}_k\,|\lambda_k < \tau\}\,.$$

Let W be any linear subspace. We denote by p the B-orthogonal projection on $V_\tau + W$.
Then, for all $\mathbf{U}$ we have the following estimate:

$$\tau\,(B\,(I_d - p)\mathbf{U}\,,\,(I_d - p)\mathbf{U})\;\leqslant (A(I_d - p)\mathbf{U}, (I_d - p)\mathbf{U})\,. \tag{31}$$

Similarly, let γ be a positive number. We define

$$V_\gamma := \mathit{Span}\{\mathbf{U}_k\,|\,\lambda_k > \gamma\}\,.$$

Let W be any linear subspace. We denote by q the B-orthogonal projection on $V_\gamma + W$.
Then, for all $\mathbf{U}$ we have the following estimate:

$$(A(I_d - q)\mathbf{U}, (I_d - q)\mathbf{U}) \leqslant \gamma\,(B\,(I_d - q)\mathbf{U}\,,\,(I_d - q)\mathbf{U})\,. \tag{32}$$

Proof We have using $V_\tau \subset V_\tau + W$:

$$\tau \leqslant \min_{\mathbf{U}\in V_\tau^{B\perp}} \frac{(A\mathbf{U},\mathbf{U})}{(B\mathbf{U},\mathbf{U})} \leqslant \min_{\mathbf{U}\in (V_\tau + W)^{B\perp}} \frac{(A\mathbf{U},\mathbf{U})}{(B\mathbf{U},\mathbf{U})}\,.$$

For all $\mathbf{U}$, the vector $(I_d - p)\mathbf{U}$ is B-orthogonal to $V_\tau + W$ and this ends the proof of (31). The proof of (32) follows similarily from

$$\gamma \geqslant \max_{\mathbf{U}\in V_\gamma^{B\perp}} \frac{(A\mathbf{U},\mathbf{U})}{(B\mathbf{U},\mathbf{U})}\,.$$

■

In order to apply the fictitious space Lemma to the study of the preconditioner (27), three assumptions have to be checked.

• $\widetilde{\mathcal{R}}$ is onto.
Let $\mathbf{U} \in H$, we have

$$\begin{aligned}
\mathbf{U} &= \tilde{P}_0\mathbf{U} + (I_d - \tilde{P}_0)\,\mathbf{U} \\
&= \tilde{P}_0\mathbf{U} + (I_d - \tilde{P}_0)\,\textstyle\sum_{i=1}^N R_i^T D_i R_i \mathbf{U} \\
&= \tilde{P}_0\mathbf{U} + (I_d - \tilde{P}_0)\,\textstyle\sum_{i=1}^N R_i^T D_i \xi_i R_i \mathbf{U} + (I_d - \tilde{P}_0)\,\textstyle\sum_{i=1}^N R_i^T D_i (I_d - \xi_i) R_i \mathbf{U} \\
&= \underbrace{\tilde{P}_0\mathbf{U} + (P_0 - \tilde{P}_0)\,\sum_{i=1}^N R_i^T D_i \xi_i R_i \mathbf{U}}_{:=F\mathbf{U}} + \underbrace{(I_d - P_0)\,\sum_{i=1}^N R_i^T D_i \xi_i R_i \mathbf{U}}_{=0} \\
&\quad + (I_d - \tilde{P}_0)\,\textstyle\sum_{i=1}^N R_i^T D_i (I_d - \xi_i) R_i \mathbf{U}\,.
\end{aligned}$$

Let us consider the last equality. Since $F\mathbf{U}$ is the sum two terms that belong to $\mathbf{V}_0$ there exists $\mathbf{U}_0$ such that $Z\mathbf{U}_0 = F\mathbf{U}$. The third term is zero since $\sum_{i=1}^N R_i^T D_i \xi_i R_i \mathbf{U} \in V_0$. Note also that $(I_d - \xi_i) R_i \mathbf{U} \in W_{i\,\gamma}$. Therefore, we have

$$\mathbf{U} = \widetilde{\mathcal{R}}(\mathbf{U}_0, ((I_d - \xi_i)\, R_i \mathbf{U})_{1 \leqslant i \leqslant N}))\,.$$

• Continuity of $\widetilde{\mathcal{R}}$

We have to estimate a constant c_R such that for all $\mathcal{U} = (\mathbf{U}_0\,,\, (\mathbf{U}_i)_{1 \leqslant i \leqslant N}) \in H_D$ we have:

$$\begin{aligned}
a(\widetilde{\mathcal{R}}(\mathcal{U}), \widetilde{\mathcal{R}}(\mathcal{U})) &\leqslant c_R\, \tilde{b}(\mathcal{U}\,, \mathcal{U}) \\
&= c_R[(\tilde{E}\mathbf{U}_0, \mathbf{U}_0) + \textstyle\sum_{i=1}^N (B_i \mathbf{U}_i, \mathbf{U}_i)]\,.
\end{aligned}$$

Note that using $(I_d - \xi_i)\mathbf{U}_i = \mathbf{U}_i$ (recall that $\mathbf{U}_i \in W_{i\gamma}$), we have:

$$\begin{aligned}
\tilde{\mathcal{R}}(\mathcal{U}) &= Z\mathbf{U}_0 + (I_d - \tilde{P}_0)\,\textstyle\sum_{i=1}^N R_i^T\, D_i\, \mathbf{U}_i \\
&= Z\mathbf{U}_0 + (P_0 - \tilde{P}_0)\,\textstyle\sum_{i=1}^N R_i^T\, D_i\, \mathbf{U}_i + (I_d - P_0)\,\textstyle\sum_{i=1}^N R_i^T\, D_i\, \mathbf{U}_i \\
&= \underbrace{Z\mathbf{U}_0 + (P_0 - \tilde{P}_0)\,\sum_{i=1}^N R_i^T\, D_i\, (I_d - \xi_i)\mathbf{U}_i}_{\in V_0} + (I_d - P_0)\,\textstyle\sum_{i=1}^N R_i^T\, D_i\, (I_d - \xi_i)\mathbf{U}_i
\end{aligned}$$

We have thus the following estimate using the A-orthogonality of $I_d - P_0$:

$$\begin{aligned}
a(\widetilde{\mathcal{R}}(\mathcal{U}),\widetilde{\mathcal{R}}(\mathcal{U})) &= \|Z\mathbf{U}_0 + (P_0-\tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i \\
&\quad +(I_d-P_0)\textstyle\sum_{i=1}^N R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&= \|Z\mathbf{U}_0 + (P_0-\tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&\quad +\|(I_d-P_0)\textstyle\sum_{i=1}^N R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&\leqslant (1+\delta)\|Z\mathbf{U}_0\|_A^2 + (1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\textstyle\sum_{i=1}^N R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&\quad +\|\textstyle\sum_{i=1}^N R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&\leqslant (1+\delta)(E\,\mathbf{U}_0\,,\,\mathbf{U}_0) + k_0\textstyle\sum_{i=1}^N \|R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2 k_0\textstyle\sum_{i=1}^N \|R_i^T D_i\,(I_d-\xi_i)\mathbf{U}_i\|_A^2 \\
&\leqslant (1+\delta)\lambda_{max}(E\tilde{E}^{-1})(\tilde{E}\,\mathbf{U}_0\,,\,\mathbf{U}_0) \\
&\quad +k_0\,\gamma(1+(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2)\textstyle\sum_{i=1}^N (B_i\,(I_d-\xi_i)\mathbf{U}_i\,,\,(I_d-\xi_i)\mathbf{U}_i) \\
&\leqslant \max((1+\delta)\,\lambda_{max}(E\tilde{E}^{-1})\,,\,k_0\,\gamma\,(1+(1+\tfrac{1}{\delta})\,\epsilon_A^2)\;\tilde{b}(\mathcal{U},\mathcal{U})\,.
\end{aligned}$$

Based on Lemma 3.1, we can optimize the value of δ and take

$$c_R := \frac{k_0\,\gamma\,(1+\epsilon_A^2) + \lambda_{max}(E\tilde{E}^{-1}) + \sqrt{(k_0\,\gamma\,(1+\epsilon_A^2) - \lambda_{max}(E\tilde{E}^{-1}))^2 + 4\lambda_{max}(E\tilde{E}^{-1})k_0\gamma(\epsilon_A^2+1)}}{2}\,. \tag{33}$$

• Stable decomposition

The stable decomposition estimate is based on using projections p_j defined in Lemma 4.2. Let $\mathbf{U}\in H$ be decomposed as follows:

$$\begin{aligned}
\mathbf{U} &= P_0\mathbf{U} + (I_d-P_0)\textstyle\sum_{j=1}^N R_j^T D_j(I_d-p_j)R_j\mathbf{U} + \underbrace{(I_d-P_0)\sum_{j=1}^N R_j^T D_j p_j R_j\mathbf{U}}_{=0} \\
&= \underbrace{P_0\mathbf{U} + (\tilde{P}_0-P_0)\sum_{j=1}^N R_j^T D_j(I_d-p_j)R_j\mathbf{U}}_{:=F\mathbf{U}\in V_0} + (I_d-\tilde{P}_0)\textstyle\sum_{j=1}^N R_j^T D_j(I_d-p_j)R_j\mathbf{U}\,.
\end{aligned}$$

We define $\mathbf{U}_0$ be such that $Z\mathbf{U}_0 = F\mathbf{U}$. We have that $(I_d-p_j)R_j\mathbf{U}$ is B_j-orthogonal to $V_{\gamma\,j} \dotplus V_{\tau\,j}$ and thus to $V_{\gamma\,j}$. This means that $(I_d-p_j)R_j\mathbf{U}\in W_{\gamma\,j}$ and that we can choose the following decomposition:

$$\mathbf{U} = \widetilde{\mathcal{R}}(\mathbf{U}_0, ((I_d-p_j)R_j\mathbf{U})_{1\leqslant j\leqslant N})\,.$$

The stability of the decomposition consists in estimating a constant $c_T>0$ such that :

$$c_T\,[(\tilde{E}\,\mathbf{U}_0,\mathbf{U}_0) + \sum_{j=1}^N (B_j(I_d-p_j)R_j\mathbf{U},(I_d-p_j)R_j\mathbf{U})] \leqslant a(\mathbf{U},\mathbf{U})\,. \tag{34}$$

Using Lemma 4.2, we have

$$\tau\,\sum_{j=1}^N (B_j(I_d-p_j)R_j\mathbf{U},(I_d-p_j)R_j\mathbf{U}) \leqslant k_1\,a(\mathbf{U},\mathbf{U})\,. \tag{35}$$

We now focus on the first term of the left hand side of (34). Let δ be some positive number, the following auxiliary result will be useful:

$$\begin{aligned}
\|F\mathbf{U}\|_A^2 &\leqslant (1+\delta)\|P_0\mathbf{U}, P_0\mathbf{U}\|_A^2 \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\textstyle\sum_{j=1}^N R_j^T D_j(I_d-p_j)R_j\mathbf{U}\|_A^2 \\
&\leqslant (1+\delta)(A\mathbf{U},\mathbf{U}) \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2\|\textstyle\sum_{j=1}^N R_j^T D_j(I_d-p_j)R_j\mathbf{U}\|_A^2 \\
&\leqslant (1+\delta)a(\mathbf{U},\mathbf{U}) \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2\, k_0\textstyle\sum_{j=1}^N \|R_j^T D_j(I_d-p_j)R_j\mathbf{U}\|_A^2 \\
&\leqslant (1+\delta)a(\mathbf{U},\mathbf{U}) \\
&\quad +(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2\, k_0\gamma\textstyle\sum_{j=1}^N (B_j(I_d-p_j)\,R_j\,\mathbf{U},(I_d-p_j)\,R_j\,\mathbf{U}) \\
&\leqslant ((1+\delta)+(1+\tfrac{1}{\delta})\|(P_0-\tilde{P}_0)\|_A^2\, k_0\gamma\,\tau^{-1}k_1)\,a(\mathbf{U},\mathbf{U})
\end{aligned}$$

where we have used Lemma 4.3 (32) (applied with A replaced by $D_j\,R_j\,A\,R_j^T\,D_j$ and B by B_j) for the one before last estimate and Lemma 4.2 for the last estimate.

The optimal value for δ yields:

$$\|F\mathbf{U}\|_A^2 \leqslant (1+\epsilon_A\sqrt{k_0\,k_1\gamma\,\tau^{-1}})^2 a(\mathbf{U},\mathbf{U})\,. \tag{36}$$

We have

$$\begin{aligned}
(\tilde{E}\mathbf{U}_0,\mathbf{U}_0) &\leqslant \lambda_{max}(E^{-1}\tilde{E})(E\mathbf{U}_0,\mathbf{U}_0) = \lambda_{max}(E^{-1}\tilde{E})(A\,Z\mathbf{U}_0,Z\mathbf{U}_0) \\
&= \lambda_{max}(E^{-1}\tilde{E})\|F\mathbf{U}\|_A^2\,.
\end{aligned}$$

so that with (36), this yields:

$$(\tilde{E}\mathbf{U}_0\,,\,\mathbf{U}_0) \leqslant \lambda_{max}(E^{-1}\tilde{E})\,(1+\epsilon_A\sqrt{k_0\,k_1\,\gamma\,\tau^{-1}})^2\,a(\mathbf{U},\mathbf{U})\,.$$

Finally, in (34) we can take :

$$c_T := \frac{1}{\lambda_{max}(E^{-1}\tilde{E})\,(1+\epsilon_A\sqrt{k_0\,k_1\,\gamma\,\tau^{-1}})^2+k_1\,\tau^{-1}}\,. \tag{37}$$

Overall, with c_T given by (37) and c_R by (33), we have proved the following spectral estimate:

$$c_T \leqslant \lambda(M_{GenEO2ACS}^{-1}\,A) \leqslant c_R\,. \tag{38}$$

Constants c_T and c_R are stable with respect to ϵ_A and the spectrum of $E\tilde{E}^{-1}$ so that (38) proves the stability of preconditioner $M_{GenEO2ACS}^{-1}$ (27) w.r.t. approximate solves.

Remark 4.2 *Had we taken the GenEO-2 algorithm introduced in [3] and modified only the coarse space solves:*

$$\widetilde{M}_{GenEO,2}^{-1} = Z\,\tilde{E}^{-1}\,Z^T + (I_d-\tilde{P}_0)\,(\sum_{i=1}^N R_i^T\,D_i\,B_i^{-1}\,D_i\,R_i)\,(I_d-\tilde{P}_0^T)\,, \tag{39}$$

the estimate for the upper bound of the preconditioned system would be for arbitray $\delta > 0$

$$\lambda_{max} \leqslant \max(1+\delta\,,\, k_0\gamma + (1+\frac{1}{\delta})\epsilon_A^2 k_0 \max_{1\leqslant i\leqslant N} \|B_i^{-1}\, D_i\, R_i\, A\, R_i^T\, D_i\|_2^2)$$

and would depend on the product of ϵ_A *with the largest eigenvalue of the local operators* $B_i^{-1}\, D_i\, R_i\, A\, R_i^T\, D_i$. *This last term can be very large and we were not able to guarantee robustness with respect to approximte coarse solves.*

Remark 4.3 *If for some subdomain* i, $1 \leqslant i \leqslant N$, $B_i = \tilde{A}_i$ *and* $\tilde{A}_i$ *is symmetric positive semi-definite and* $D_i R_i A R_i^T D_i$ *is SPD, the eigenvalue problem* (20) *will not contribute to the coarse space. More precisely, the contribution of the subdomain to the coarse space will be* $R_i^T\, D_i\, \ker(\tilde{A}_i) \bigoplus V_{i,geneo}^{\gamma}$. *Also in Definition 4.3,* $B_i^{\dagger}$ *is the pseudo inverse of* B_i *where* $W_{i\gamma}$ *is the image of* B_i *which is orthogonal to* $\ker(\tilde{A}_i)$ *and* ξ_i *is the orthogonal projection on* $\ker(\tilde{A}_i)$ *parallel to* $W_{i\gamma}$.

5 Conclusion

We have proved the robustness of GenEO methods with respect to approximate coarse solves. It paves the way to three or more level methods in a multigrid fashion.

References

[1] Markus Blatt, Ansgar Burchardt, Andreas Dedner, Christian Engwer, Jorrit Fahlke, Bernd Flemisch, Christoph Gersbacher, Carsten Gräser, Felix Gruber, Christoph Grüninger, et al. The distributed and unified numerics environment, version 2.4. *Archive of Numerical Software*, 4(100):13–29, 2016.

[2] Victorita Dolean, Pierre Jolivet, and Frédéric Nataf. *An Introduction to Domain Decomposition Methods: algorithms, theory and parallel implementation.* SIAM, 2015.

[3] R. Haferssas, P. Jolivet, and F. Nataf. An Additive Schwarz Method Type Theory for Lions's Algorithm and a Symmetrized Optimized Restricted Additive Schwarz Method. *SIAM J. Sci. Comput.*, 39(4):A1345–A1365, 2017.

[4] Ryadh Haferssas, Pierre Jolivet, and Frédéric Nataf. A robust coarse space for optimized Schwarz methods: SORAS-GenEO-2. *C. R. Math. Acad. Sci. Paris*, 353(10):959–963, 2015.

[5] F. Hecht. New development in Freefem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[6] Pierre Jolivet and Frédéric Nataf. Hpddm: High-Performance Unified framework for Domain Decomposition methods, MPI-C++ library. https://github.com/hpddm/hpddm, 2014.

[7] Jan Mandel. Balancing domain decomposition. *Comm. on Applied Numerical Methods*, 9:233–241, 1992.

[8] Jan Mandel, Bedřich Sousedík, and Clark R. Dohrmann. *On Multilevel BDDC*, pages 287–294. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[9] Roy A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM J. Numer. Anal.*, 24(2):355–365, 1987.

[10] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[11] Clemens Pechstein and Clark R. Dohrmann. A unified framework for adaptive BDDC. *Electron. Trans. Numer. Anal.*, 46:273–336, 2017.

[12] C. Prud'homme. A Domain Specific Embedded Language in c++ for automatic differentiation, projection, integration and variational formulations. *Scientific Programming*, 14(2):81–110, 2006.

[13] Nicole Spillane, Victorita Dolean, Patrice Hauret, Frédéric Nataf, Clemens Pechstein, and Robert Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numer. Math.*, 126(4):741–770, 2014.

[14] Amik St-Cyr, Martin J. Gander, and Stephen J. Thomas. Optimized Multiplicative, Additive, and Restricted Additive Schwarz Preconditioning. *SIAM J. Sci. Comput.*, 29(6):2402–2425 (electronic), 2007.

[15] Andrea Toselli and Olof Widlund. *Domain Decomposition Methods - Algorithms and Theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer, 2005.

[16] Xuemin Tu. Three-level BDDC in three dimensions. *SIAM J. Sci. Comput.*, 29(4):1759–1780, 2007.

[17] Xuemin Tu. A three-level BDDC algorithm for a saddle point problem. *Numer. Math.*, 119(1):189–217, 2011.