Robust domain decomposition methods for symmetric positive definite problems

Nicole Spillane

École Doctorale Paris Centre

Thèse de doctorat

Discipline : Mathématiques Appliquées

présentée par

Nicole Spillane

Méthodes de décomposition de domaine robustes pour les problèmes symétriques définis positifs

---

dirigée par Frédéric Nataf et Patrice Hauret

Soutenue le 22 janvier 2014 devant le jury composé de :

Victorita Dolean	Examinateur
Patrice Hauret	Directeur
Frédéric Hecht	Examinateur
Patrick Le Tallec	Rapporteur
Frédéric Nataf	Directeur
Christian Rey	Examinateur
Daniel J. Rixen	Examinateur
Olof Widlund	Rapporteur

Laboratoire Jacques Louis Lions
Boite courrier 187
4, place Jussieu
75252 Paris cedex 05

École doctorale Paris centre
Boite courrier 188
4, place Jussieu
75252 Paris cedex 05

Remerciements

Quel plaisir de repenser à ces trois années et à toutes les personnes qui m'ont accompagnées tout au long. Naturellement mes premiers remerciements sont pour mes directeurs de thèse Frédéric Nataf et Patrice Hauret. Frédéric je te remercie pour ta gentillesse, ta disponibilité, toutes les connaissances mathématiques que tu as partagées avec moi mais aussi la liberté que tu m'as accordée et qui m'a permis de m'épanouir dans mes débuts en recherche. Patrice, même si Ladoux n'est pas la porte à côté je n'ai jamais hésité à te contacter et je te suis sincèrement reconnaissante pour cette présence. J'ai appris beaucoup de nos conversations et elles continuent à me laisser entrevoir des pistes de réflexion.

Non contente d'avoir deux directeurs de thèse j'ai eu la chance d'être suivie très régulièrement par Victorita Dolean, Daniel Rixen et Pascal Tremblay. Je les remercie pour le temps qu'ils m'ont accordé et leur aide fréquente et toujours bienvenue. Un grand merci en particulier à Daniel pour m'avoir accueillie à Delft puis à Munich et m'avoir patiemment expliqué les ficelles de FETI.

Les rapporteurs de cette thèse, Patrick Le Tallec et Olof Widlund, sont parmi les pionniers de la décomposition de domaine et je suis flattée qu'ils aient accepté d'évaluer mon travail et d'assister à la soutenance. Frédéric Hecht et Christian Rey ont aussi accepté de participer au jury et j'en suis honorée.

Le laboratoire Jacques Louis Lions a été l'endroit idéal pour préparer cette thèse. Je profite de cette occasion pour remercier Catherine, Nadine, Salima, Madame Boullic, Madame Ruprecht ainsi que Khashayar et Altair sans qui tout serait infiniment plus compliqué !

De mes séjours chez Michelin je garde un très bon souvenir grâce à l'ensemble de l'équipe ER. C'était toujours un plaisir de vous retrouver et ce n'est pas peu dire vu mon amour, disons partagé, pour l'Auvergne.

J'ai aussi une pensée pour l'ensemble de l'équipe INRIA Alpines à commencer par Frédéric et Laura Grigori grâce à qui j'en fais partie. J'espère pouvoir contribuer de nouveau à ce travail. Merci Laurence d'être toujours si efficace et agréable. Et puis un merci particulier à Pierre Jolivet pour nos travaux communs et les 1001 dépannages informatiques !

Parmi les personnes avec qui j'ai eu la chance de travailler et que je n'ai pas encore citées figurent Robert Scheichl et Clemens Pechstein : c'était une chance de travailler avec vous, j'ai beaucoup appris et l'atmosphère amicale dans laquelle vous m'avez intégrée m'y a beaucoup aidé.

Je n'oublie pas bien sur tous les autres doctorants du laboratoire : en premier lieu les mousquetaires mais aussi l'équipe du GTT, les habitants successifs du bureau 15-16 301, les compagnons de crous et bien évidemment ceux du Mayflower... Je suis heureuse de vous compter désormais parmi mes amis.

Enfin il y a la vraie vie, mes amis et ma famille et mes derniers remerciements sont bien évidemment pour Ann, Alice, Juliette, Mamadou, Marc, Olivier, Paul, Sonia, Sophie, Tony et Yann.

Contents

Chapitre 1

Introduction : Version française

Non french speakers may go directly to the next chapter where the same introduction is given in english.

Contents

Il existe deux grandes familles de méthodes qui permettent de résoudre un système linéaire sur une architecture parallèle et pour lesquelles des implémentations utilisables en boite noire sont disponibles. Il s'agit des solveurs directs et des solveurs itératifs. Les solveurs directs sont robustes dans le sens où on peut garantir théoriquement, aux arrondis près, qu'ils vont trouver la solution en un certain nombre d'opérations peu importe la difficulté du problème. L'inconvénient est que si les problèmes sont trop gros alors la mémoire requise devient limitante. Les solveurs itératifs quant à eux sont naturellement parallèles et ne rencontrent pas de problèmes de mémoires puisqu'ils exploitent surtout des produits matrice-vecteur. Par contre ils manquent souvent de robustesse : pour des problèmes mal conditionnés le fait de préconditionner le problème devient incontournable et le choix du bon préconditionneur est un art à part entière.

Les méthodes de décomposition de domaine peuvent être vues comme des solveurs hybrides : elles résolvent le système avec une méthode itérative tout en exploitant des solveurs directs sur des sous problèmes pour réécrire le problème ou pour appliquer le

préconditionneur (ou parfois les deux). L'idée derrière cette hybridation est de tirer le meilleur parti de chaque famille de solveur et ainsi d'allier la robustesse à la parallélisation.

Dans cette section nous présentons trois des méthodes de décomposition de domaine les plus populaires et nous illustrons le fait qu'elles peuvent manquer de robustesse quand elles sont confrontées à des problèmes particulièrement difficiles. Puis nous décrivons ce qu'est une méthode de décomposition de domaines à deux niveaux et comment l'ajout du second niveau peut contribuer à la robustesse. Finalement, nous présentons les contributions principales de cette thèse qui sont développées dans les chapitres suivants. La motivation qui lie l'ensemble des travaux est de fabriquer des méthodes de décomposition de domaines pour lesquelles on sait prouver qu'elles vont converger même pour des problèmes très difficiles et qui peuvent être implémentées en boite noire et donc utilisées sans connaissance particulière du problème sous-jacent au système linéaire.

1.1 Méthodes de décomposition de domaine

Nous introduisons deux familles de méthodes de décomposition de domaine pour lesquelles nous développerons des améliorations dans le cœur de cette thèse. La première est la méthode de Schwarz. L'un de ces principaux avantages est qu'elle peut être implémentée de manière purement algébrique : aucune connaissance du problème autre que sa formulation $A\mathbf{x} = \mathbf{b}$ n'est nécessaire. La seconde famille de méthodes est constituée des méthodes dites de sous-structuration. Elles sont plus sophistiquées puisqu'elles requièrent l'accès aux matrices élémentaires afin de pouvoir assembler les matrices du problème restreint à certains sous domaines. Dans un cadre industriel ce sont souvent les méthodes de sous-structuration qui sont mises en œuvre.

1.1.1 Méthodes de Schwarz

Une présentation historique détaillée est donnée dans [41] avec les références bibliographiques complètes. Les méthodes de décomposition de domaine de Schwarz sont nommées après H. A. Schwarz qui en 1870 [101] a proposé l'algorithme de Schwarz alterné afin d'étudier l'existence d'une solution au problème de Poisson homogène avec conditions aux limites imposées (1.1) :

$$\left\{ \begin{array}{rcl} -\Delta u & = & 0 \text{ dans } \Omega, \\ u & = & g \text{ sur } \partial\Omega, \end{array} \right. \tag{1.1}$$

où $\Omega = \Omega_1 \cup \Omega_2$ est dessiné dans la Figure 1.1. Partant de l'existence d'une solution sur des domaines réguliers (cercles, carrés...) l'idée de Schwarz est de démontrer par un argument de construction l'existence d'une solution sur un domaine Ω non régulier mais constitué d'éléments réguliers (comme celui de la Figure 1.1). Pour cela il propose de résoudre en alternance le problème dans chacun des sous domaines avec des conditions de transmission basées sur la solution qui vient d'être calculée dans le sous domaine voisin. Plus précisément Schwarz démontre que l'algorithme d'*alternance de Schwarz* initialisé par u_2^0 et mis à jour selon

$$\begin{array}{ll} -\Delta u_1^{n+1} = 0 \text{ dans } \Omega_1 & -\Delta u_2^{n+1} = 0 \text{ dans } \Omega_2 \\ u_1^{n+1} = g \text{ sur } \partial\Omega_1 \cap \partial\Omega & u_2^{n+1} = g \text{ sur } \partial\Omega_2 \cap \partial\Omega \\ u_1^{n+1} = u_2^n \text{ dans } \Omega \setminus \Omega_1, & u_2^{n+1} = u_1^{n+1} \text{ dans } \Omega \setminus \Omega_2, \end{array} \tag{1.2}$$

converge vers la solution du problème de Poisson (1.1) et donc que cette solution existe.

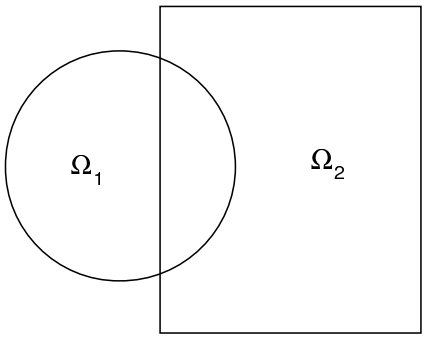


Figure 1.1 – Le domaine Ω consiste d'un rectangle et d'un disque avec une partie superposée.

Les travaux [3] proposent d'adapter cet algorithme pour l'utiliser en tant que solveur itératif. L'adaptation la plus immédiate est la méthode de Schwarz multiplicative [11, 10, 102]. Son inconvénient est qu'elle est séquentielle par nature puisque le problème à résoudre dans chaque sous domaine dépend de la solution qui vient d'être obtenue dans les sous domaines voisins *via* la condition aux limites. L'adaptation parallèle est le préconditionneur de Schwarz additif auquel Matsokin et Nepomnyaschikh [72] ont largement contribué et que nous présentons à présent en nous basant sur la description que l'on trouve dans [112].

Si la discrétisation par éléments finis de (1.1) s'écrit $A\mathbf{u} = \mathbf{b}$, et que $R_1^\top$ (respectivement $R_2^\top$) est la matrice d'interpolation (booléenne) qui prolonge une fonction éléments finis définie sur Ω_1 (respectivement Ω_2) à Ω tout entier par zéro alors on peut définir les opérateurs locaux $A_1 := R_1 A R_1^\top$, $A_2 := R_2 A R_2^\top$ ainsi que le préconditionneur de Schwarz additif :

$$M^{-1} := R_1^\top A_1^{-1} R_1 + R_2^\top A_2^{-1} R_2. \tag{1.3}$$

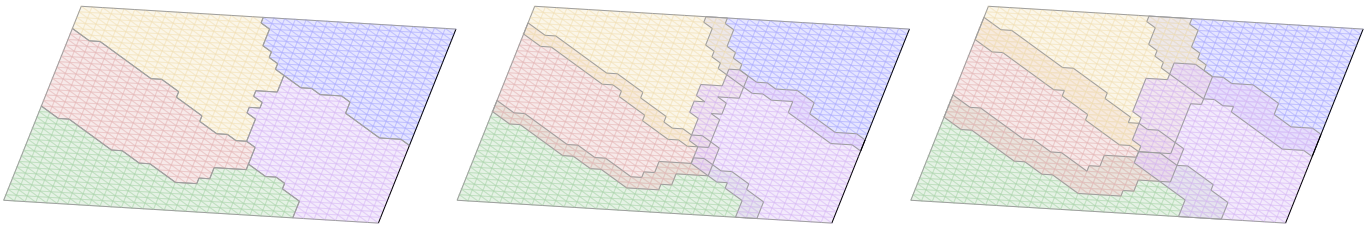
Il est assez intuitif de voir que c'est un bon préconditionneur pour A. En effet il approche l'inverse de A par une somme d'inverses sur chacun des deux sous domaines. Ceci se généralise aisément au cas de N sous domaines et d'une matrice A symétrique définie positive quelconque. Tout ce dont on a besoin est un ensemble de sous espaces V_i de l'espace V des degrés de liberté et d'opérateurs d'interpolation $R_i^\top : V_i \to V$ qui vérifient

$$V = \sum_{i=1}^{N} R_i^\top V_i.$$

Le préconditionneur de Schwarz additif s'écrit alors comme la somme des N inverses locaux

$$M^{-1} := \sum_{i=1}^{N} R_i^\top A_i^{-1} R_i, \text{ où } A_i := R_i A R_i^\top. \tag{1.4}$$

Une manière courante d'appliquer ce préconditionneur est de découper l'espace V des degrés de liberté en sous domaines deux à deux disjoints puis de rajouter l couches de recouvrement à chaque sous domaines comme c'est illustré dans la Figure 1.2. Dans ce cas les opérateurs d'interpolation $R_i^\top$ sont booléens et les matrices A_i sont simplement des sous matrices extraites de A. La partition de départ peut être obtenue à la main en se basant sur la géométrie ou grâce à un partitionneur de graphe automatique tel que Metis [54] ou Scotch [13].



(a) $l = 0$ (b) $l = 1$ (c) $l = 2$

FIGURE 1.2 – Partition de $\Omega = [0; 1]^2$ en $N = 5$ sous domaines avec différentes valeurs du paramètre de recouvrement.

Comme nous le verrons plus tard le fait qu'il y ait du recouvrement est essentiel pour assurer la convergence avec le préconditionneur de Schwarz additif. Il y a deux inconvénients principaux à ce recouvrement. Le premier, et peut être le plus évident, est qu'il faut assumer le coût de résoudre plusieurs fois le problème dans la partie du domaine qui est dédoublée. Le second est que lorsqu'on simule un objet qui est constitué de plusieurs matériaux il est naturel de construire les sous domaines de manière à ce qu'un seul matériau soit présent dans chaque sous domaine, avec le recouvrement ce n'est pas possible.

1.1.2 Méthodes de sous-structuration

Nous présentons ici deux méthodes très populaires : BDD (Balancing Domain Decomposition) et FETI (Finite Element Tearing and Interconnecting). La méthode BDD est due à Mandel [67] et est basée sur la méthode Neumann-Neumann [14] par De Roeck et Le Tallec. La méthode FETI a été écrite par Farhat et Roux [36]. Notre objectif dans ce chapitre d'introduction est d'illustrer les idées qui sont à la base de ces méthodes. Afin de gagner en clarté nous nous limitons à un problème d'élasticité linéaire. Des présentations rigoureuses pour un problème symétrique défini positif général seront données au Chapitre 5. Les travaux [58, 112, 45] proposent des présentations très complètes des méthodes de sous-structuration. En particulier on trouve dans [45] leur interprétation mécanique ainsi que des techniques d'implémentation.

Soit Ω une partie ouverte de $\mathbb{R}^d$ pour $d = 2$ ou $d = 3$. Soit $\partial\Omega$ le bord du domaine Ω et $\partial\Omega_D \subset \partial\Omega$ une partie de ce bord où une condition de Dirichlet homogène est imposée. On introduit l'espace $\mathcal{V} := \{v \in H^1(\Omega) : v_{|\Omega_D} = 0\}$. Pour une force extérieure $f \in \mathcal{V}'$, la formulation variationnelle du problème d'élasticité linéaire est : Trouver le champ des déplacements $v \in \mathcal{V}$ tel que

$$2\int_\Omega \mu\, \epsilon(u) : \epsilon(v)dx + \int_\Omega \lambda(\nabla \cdot u)(\nabla \cdot v)dx = \int_\Omega \langle f, v\rangle dx \quad \forall\, v \in \mathcal{V}, \tag{1.5}$$

où

$$\epsilon(u) : \epsilon(v) := \sum_{i=1}^{d}\sum_{j=1}^{d} \epsilon_{ij}(u)\epsilon_{ij}(v); \quad \epsilon_{ij}(u) := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

$$\langle f, v\rangle := \sum_{i=1}^{d} f_i v_i,$$

et μ et λ sont deux paramètres appelés les coefficients de Lamé qui décrivent les propriétés du matériel. Ils s'écrivent en fonction du module de Young E et du coefficient de Poisson

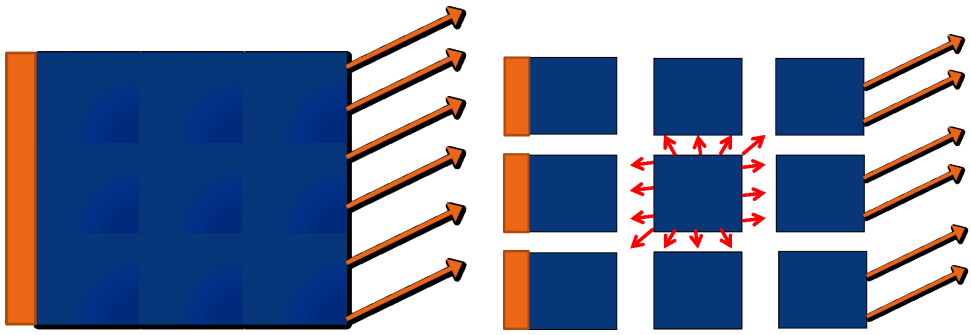


Figure 1.3 – Partition de Ω en sous domaines réguliers (gauche : domaine de départ – droite : domaine partitionné). Le domaine est encastré à gauche et soumis à une force surfacique sur le bord de droite. Les flèches rouges entre les sous domaines correspondent aux forces surfaciques qui résultent de l'interaction du sous domaine du milieu avec ses voisins.

ν $(0 < \nu < 0.5)$ du matériau selon

$$\lambda := \frac{E\nu}{(1+\nu)(1-2\nu)},\ \mu := \frac{E}{2(1+\nu)}.$$

L'élasticité linéaire est une approximation, en petite déformation, des équations de l'elasticité [62]. Supposons que l'on a discrétisé (1.5) avec des éléments finis linéaires ($\mathbb{P}_1$) et que dans cette base le problème discrétisé s'écrit : Trouver $\hat{\mathbf{u}} \in \mathbb{R}^n$ tel que

$$\hat{K}\hat{\mathbf{u}} = \hat{\mathbf{f}}.$$

Nous pouvons maintenant introduire les composantes locales des méthodes de sous-structuration. Supposons que Ω a été partitionné en sous domaines disjoints deux à deux :

$$\Omega = \bigcup_{i=1}^{N} \overline{\Omega}_i; \quad \Omega_i \cap \Omega_j = \emptyset \quad \forall\, i \neq j.$$

On dénote par K_i et $\mathbf{f}_i$ la matrice du problème local et le terme de force extérieure local qui correspondent à la discrétisation de $2\int_{\Omega_i} \mu\, \epsilon(u) : \epsilon(v)dx + \int_{\Omega_i} \lambda(\nabla \cdot u)(\nabla \cdot v)dx$ et de $\int_{\Omega_i} \langle f, v\rangle dx$ pour les fonctions $\{u_{|\Omega_i}; u \in \mathbb{P}_1\}$. L'équation d'équilibre du sous domaine Ω_i s'écrit alors

$$K_i\mathbf{u}_i = \mathbf{f}_i + \mathbf{g}_i, \text{ où } \mathbf{g}_i \text{ sont les forces surfaciques.} \tag{1.6}$$

Nous constatons que dans l'équation d'équilibre mécanique une inconnue supplémentaire est apparue : le terme de force surfacique $\mathbf{g}_i$ qui correspond à la pression exercée par les sous domaines voisins. C'est ce que nous illustrons avec les flèches rouges dans la Figure 1.3.

Nous introduisons à présent une partition des degrés de liberté en degrés de liberté à l'interface, qui sont partagés par deux sous domaines au moins et qui forment l'ensemble

$$\Gamma := \bigcup_{i,j=1,\ldots,N;\, i\neq j} \left(\partial\Omega_i \cap \partial\Omega_j\right),$$

et tous les autres degrés de liberté dénotés avec un I (pour Intérieur). Avec des notations évidentes, l'équation d'équilibre local (1.6) admet la formulation par blocs suivantes

$$\begin{pmatrix} K_i^{II} & K_i^{I\Gamma} \\ K_i^{\Gamma I} & K_i^{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_i^I \\ \mathbf{u}_i^\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{f}_i^I \\ \mathbf{f}_i^\Gamma \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{g}_i^\Gamma \end{pmatrix}. \tag{1.7}$$

Par définition les forces d'interfaces sont nulles pour les degrés de liberté à l'intérieur de Ω_i et en utilisant un complément de Schur on peut aussi éliminer les inconnues de déplacements à l'intérieur du sous domaines. Sous la forme de système (1.7) s'écrit

$$\begin{cases} K_i^{II}\mathbf{u}_i^I + K_i^{I\Gamma}\mathbf{u}_i^\Gamma = \mathbf{f}_i^I \\ K_i^{\Gamma I}\mathbf{u}_i^I + K_i^{\Gamma\Gamma}\mathbf{u}_i^\Gamma = \mathbf{f}_i^\Gamma + \mathbf{g}_i^\Gamma \end{cases}$$

ce qui est équivalent à

$$\begin{cases} \mathbf{u}_i^I = {K_i^{II}}^{-1}\left(\mathbf{f}_i^I - K_i^{I\Gamma}\mathbf{u}_i^\Gamma\right) \\ \underbrace{\left[K_i^{\Gamma\Gamma} - K_i^{\Gamma I}{K_i^{II}}^{-1}K_i^{I\Gamma}\right]}_{:=S_i}\mathbf{u}_i^\Gamma = \underbrace{\left[\mathbf{f}_i^\Gamma - K_i^{\Gamma I}{K_i^{II}}^{-1}\mathbf{f}_i^I\right]}_{:=\tilde{\mathbf{f}}_i} + \mathbf{g}_i^\Gamma. \end{cases} \tag{1.8}$$

En plus de l'équation d'équilibre mécanique $S_i\mathbf{u}_i^\Gamma = \tilde{\mathbf{f}}_i + \mathbf{g}_i^\Gamma$, chaque sous domaine doit satisfaire à des conditions de continuité et de compatibilité avec ces voisins. Ces conditions s'écrivent avec deux types opérateurs d'interpolation :

- Les **opérateurs d'assemblage** $R_i^\top$ sont des matrices booléennes : étant donné un vecteur $\mathbf{u}_i^\Gamma$ correspondant aux degrés de liberté de $\partial\Omega_i \cap \Gamma$, $R_i^\top\mathbf{u}_i$ est un vecteur dont les entrées sur le bord Γ tout entier ont les mêmes valeurs que $\mathbf{u}_i^\Gamma$ pour les degrés de liberté sur $\partial\Omega_i \cap \Gamma$ et valent 0 partout ailleurs.
- Les **opérateurs de saut** B_i sont des matrices booléennes signées où chaque ligne de B_i correspond à un degré de liberté x de l'interface Γ et à deux sous domaines Ω_k et Ω_l tels que $x \in \partial\Omega_k \cap \partial\Omega_l$. Si $i = k$ ou l alors, à l'entrée de la ligne qui correspond à la numérotation locale de x, est assignée la valeur -1 si $i = \min(k, l)$ et $+1$ autrement.

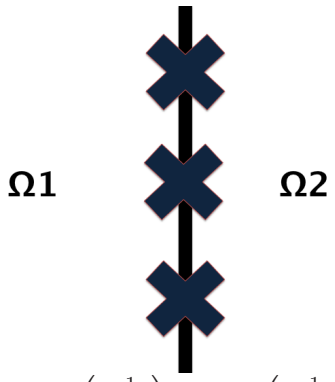
L'action des opérateurs de saut et d'assemblage est illustrée sur un exemple simplifié dans la Figure 1.4. Avec ces opérateurs le problème d'élasticité global $\hat{K}\hat{\mathbf{u}} = \hat{\mathbf{f}}$ se réécrit : Pour chaque sous domaine $i = 1, \ldots, N$ trouver le champ des déplacements $\mathbf{u}_i^\Gamma$ et le champ des forces d'interface $\mathbf{g}_i^\Gamma$ tel que

$$\begin{cases} S_i\mathbf{u}_i^\Gamma = \tilde{\mathbf{f}}_i + \mathbf{g}_i^\Gamma, \quad \forall\, i = 1, \ldots, N & \text{[Équilibre Local]} \\ \sum_{i=1}^N B_i\mathbf{u}_i^\Gamma = \mathbf{0} & \text{[Continuité à l'interface]} \\ \sum_{i=1}^N R_i^\top\mathbf{g}_i^\Gamma = \mathbf{0} & \text{[Équilibre de l'interface].} \end{cases} \tag{1.9}$$

Les formulations FETI et BDD du problème d'élasticité sont toutes deux basées sur (1.9).

Formulation BDD La première étape est d'éliminer la condition de continuité à l'interface en cherchant les déplacements dans l'ensemble réduit $\{(\mathbf{u}_1^\Gamma, \ldots, \mathbf{u}_N^\Gamma); \sum_{i=1}^N B_i\mathbf{u}_i^\Gamma = \mathbf{0}\}$. Par définition des opérateurs d'assemblage et de saut cela revient à chercher un vecteur $\hat{\mathbf{u}}^\Gamma$ défini sur l'interface globale entre les sous domaines et à choisir $\mathbf{u}_i^\Gamma = R_i\hat{\mathbf{u}}^\Gamma$ dans chaque sous domaine. Dans ce cas (1.9) est équivalent à : Trouver $\hat{\mathbf{u}}^\Gamma$ et $\mathbf{g}_1^\Gamma, \ldots, \mathbf{g}_N^\Gamma$ tels que

$$\begin{cases} S_iR_i\hat{\mathbf{u}}^\Gamma = \tilde{\mathbf{f}}_i + \mathbf{g}_i^\Gamma, \quad \forall\, i = 1, \ldots, N \\ \sum_{i=1}^N R_i^\top\mathbf{g}_i^\Gamma = \mathbf{0}. \end{cases} \tag{1.10}$$



$$\textbf{Assemblage :}\quad R_1^\top \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} + R_2^\top \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} x_1^1 + x_2^1 \\ x_1^2 + x_2^2 \\ x_1^3 + x_2^3 \end{pmatrix}$$

$$\textbf{Saut :}\ B_1 \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} + B_2 \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} x_1^1 - x_2^1 \\ x_1^2 - x_2^2 \\ x_1^3 - x_2^3 \end{pmatrix}$$

Figure 1.4 – Illustration de l'action des opérateurs de saut B_i et des opérateurs d'assemblage $R_i^\top$ sur un cas à deux sous domaines, où l'interface est constituée de trois degrés de liberté.

Finalement on injecte la première ligne de (1.10) dans la seconde pour trouver la formulation BDD du problème :

$$\left(\sum_{i=1}^{N} R_i^\top S_i R_i\right) \hat{\mathbf{u}}^\Gamma = \sum_{i=1}^{N} R_i^\top \tilde{\mathbf{f}}_i. \tag{1.11}$$

Une fois que l'on a résolu le problème à l'interface on peut calculer les déplacements à l'intérieur des sous domaines *via* la première équation de (1.8). Puisque l'opérateur BDD est une somme de compléments de Schur il est naturel de le préconditionner avec une somme d'inverses de compléments de Schur. Plus précisément dans le cas où les matrices S_i sont inversibles le préconditionneur pour BDD est

$$M^{-1} := \sum_{i=1}^{N} \widetilde{R_i^\top} S_i^{-1} \widetilde{R_i}, \tag{1.12}$$

où $\widetilde{R_i^\top}$ et $\widetilde{R_i}$ sont les mêmes opérateurs que $R_i^\top$ and R_i mais pondérés par une partition de l'unité.

Formulation FETI Cette fois c'est l'équation d'équilibre de l'interface que l'on élimine en cherchant le champ des forces d'interface dans le sous espace $\{(\mathbf{g}_1^\Gamma, \ldots, \mathbf{g}_N^\Gamma); \sum_{i=1}^N R_i^\top \mathbf{g}_i^\Gamma = \mathbf{0}\}$. Par définition des opérateurs de saut cela revient au même que de chercher un vecteur $\boldsymbol{\lambda} \in \mathrm{Im}(\sum_{i=1}^N B_i)$ et de choisir $\mathbf{g}_i^\Gamma = -B_i^\top \boldsymbol{\lambda}$ dans chaque sous domaine. De cette manière (1.9) est équivalente à : Trouver $\boldsymbol{\lambda} \in \mathrm{Im}(\sum_{i=1}^N B_i)$ tel que

$$\left\{ \begin{array}{lcll} S_i \mathbf{u}_i^\Gamma & = & \tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda}, & \forall\, i = 1, \ldots, N \\ \sum_{i=1}^N B_i \mathbf{u}_i^\Gamma & = & \mathbf{0}. & \end{array} \right. \tag{1.13}$$

Supposons que l'on est dans le cas très particulier où les matrices S_i sont inversibles. Dans ce cas on peut écrire

$$\left\{ \begin{array}{lcll} \mathbf{u}_i^\Gamma & = & S_i^{-1}(\tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda}), & \forall\, i = 1, \ldots, N \\ \sum_{i=1}^N B_i \mathbf{u}_i^\Gamma & = & \mathbf{0}, & \end{array} \right. \tag{1.14}$$

et finalement la formulation FETI du problème s'obtient en injectant la première ligne dans la seconde :

$$\left(\sum_{i=1}^N B_i S_i^{-1} B_i^\top \right) \boldsymbol{\lambda} = \sum_{i=1}^N B_i S_i^{-1} \tilde{\mathbf{f}}_i. \tag{1.15}$$

Une fois que l'on a trouvé l'inconnue $\boldsymbol{\lambda}$ le champ des déplacements est calculé via (1.14) pour les degrés de liberté à l'interface et (1.8) pour les degrés de liberté internes. Puisque l'opérateur FETI est une somme d'inverse de compléments de Schur il est naturel de le préconditionner par une somme de compléments de Schur. Plus précisément le préconditionneur pour FETI est

$$M^{-1} = \sum_{i=1}^N \widetilde{B_i} S_i \widetilde{B_i}, \tag{1.16}$$

où $\widetilde{B_i^\top}$ et $\widetilde{B_i}$ sont les mêmes opérateurs que $B_i^\top$ and B_i mais pondérés par une partition de l'unité.

Remarque 1.1. Nous avons supposé que les opérateurs S_i sont inversibles. C'est loin d'être le cas général. En fait, pour le système de l'élasticité linéaire, dès que le sous domaine n'est pas concerné par la condition de Dirichlet du problème global le noyau de S_i est la trace des modes rigides sur le bord des sous domaines (où les modes rigides sont les déplacements de Ω_i qui ne déforment pas le sous domaine). En dimension 2 l'espace des modes rigides est engendré par les deux translations et la rotation du plan. En dimension 3 l'espace des modes rigides est engendré par les trois rotation et les trois translations. Dans le Chapitre 5 nous considèrerons le cas général de matrices S_i symétriques et positives et nous donnons dès la prochaine section une manière de contourner le problème. Le fait qu'il faut réserver un traitement particulier au noyau de S_i pour FETI et BDD est bien connu depuis longtemps [37, 67].

1.1.3 Défaut de robustesse : une première illustration

Puisque nous nous concentrons sur les problèmes symétriques et les préconditionneurs symétriques le solveur itératif naturel est le Gradient Conjugué Préconditionné (PCG) que nous présentons dans l'algorithme 1.1 (voir [60, 51] pour les premières introductions et [95] pour une présentation moderne).

Une manière d'utiliser la robustesse d'un solveur basé sur PCG est d'utiliser le résultat de convergence suivant qui remonte à [73, 53] (voir aussi [95](Théorème 6.29) pour une preuve) :

$$\|\mathbf{x}_* - \mathbf{x}_m\|_A \leq \frac{\|\mathbf{x}_* - \mathbf{x}_0\|_A}{C_m\left(\frac{\lambda_{max} + \lambda_{min}}{\lambda_{max} - \lambda_{min}}\right)}, \tag{1.17}$$

où

C_m est le polynôme de Tchebyshev de degré m de la première espèce,

$\mathbf{x}_*$ est la solution exacte,

Algorithm 1.1 Gradient Conjugué Préconditionné pour $A\mathbf{x}_* = \mathbf{b}$ préconditionné par M^{-1} et initialisé avec $\mathbf{x}_0$.

$\mathbf{r}_0 := \mathbf{b} - A\mathbf{x}_0$; $\mathbf{z}_0 := M^{-1}\mathbf{r}_0$ et $\mathbf{p}_0 = \mathbf{z}_0$
for $j = 0, 1, \ldots$ jusque convergence **do**
  $\alpha_j := \langle \mathbf{r}_j, \mathbf{z}_j \rangle / \langle A\mathbf{p}_j, \mathbf{p}_j \rangle$
  $\mathbf{x}_{j+1} := \mathbf{x}_j + \alpha_j \mathbf{p}_j$
  $\mathbf{r}_{j+1} := \mathbf{r}_j - \alpha_j A\mathbf{p}_j$
  $\mathbf{z}_{j+1} := M^{-1}\mathbf{r}_{j+1}$
  $\beta_j := \langle \mathbf{r}_{j+1}, \mathbf{z}_{j+1} \rangle / \langle \mathbf{r}_j, \mathbf{z}_j \rangle$
  $\mathbf{p}_{j+1} := \mathbf{z}_{j+1} + \beta_j \mathbf{p}_j$
end for

$\mathbf{x}_m$ est la solution approchée donnée par l'itération m de l'algorithme 1.1,
λ_{max} et λ_{min} sont les valeurs propres extrêmes de l'opérateur préconditionné $M^{-1}A$,
$\|\cdot\|_A$ est la norme induite par A.
Une simplification est le résultat de convergence linéaire suivant :

$$\|\mathbf{x}_* - \mathbf{x}_m\|_A \leq 2 \left[\frac{\sqrt{\lambda_{max}/\lambda_{min}} - 1}{\sqrt{\lambda_{max}/\lambda_{min}} + 1} \right]^m \|\mathbf{x}_* - \mathbf{x}_0\|_A. \tag{1.18}$$

Malgré le fait que ces bornes sont en général pessimistes, elles nous apprennent que tant que l'on peut borner le spectre de l'opérateur préconditionné, on peut aussi majorer l'erreur relative à l'itération m par une quantité qui ne dépend que de ces bornes.

Notre ambition ici est de montrer que dès que l'on considère des simulations en milieu hétérogène il est assez facile de construire un cas test pour lequel le solveur itératif devient inefficace. Nous considérons le préconditionneur de Schwarz additif (1.4) appliqué à la discrétisation du problème scalaire elliptique (que l'on appelle aussi l'équation de Darcy)

$$\begin{cases} -\nabla \cdot (\alpha \nabla u) & = 1, & \text{dans } \Omega, \\ u(x,y) & = 0, & \text{si } x = 0, \\ \frac{\partial u}{\partial \mathbf{n}}(x,y) & = 0, & \text{sur le reste de } \partial\Omega. \end{cases} \tag{1.19}$$

où $\Omega = [0; N] \times [0; 1]$. On discrétise le problème par des éléments finis de Lagrange $\mathbb{P}_1$ sur un maillage régulier à $(20N+1) \times 21$ nœuds ($N \in \mathbb{N}$). Le coefficient α est une fonction à valeurs réelles $\alpha : \Omega \to \mathbb{R}^+$. Le domaine simulé est constitué de deux matériaux (caractérisés par deux valeurs de α : α_1 et α_2) répartis en sept couches successives comme illustré dans la Figure 1.5. Afin de construire le découpage en sous domaines on partage Ω en N carrés unitaires puis on rajoute deux épaisseurs de maille à chacun.

Dans le tableau 1.1 on présente les résultats de notre test de convergence. Nous donnons le nombre d'itérations nécessaires pour que le solveur converge ainsi que l'estimation du conditionnement de l'opérateur préconditionné $M^{-1}A$ basée sur les valeurs de Ritz à la dernière itération du gradient conjugué (voir par exemple [15]). Le critère d'arrêt est basé sur l'erreur relative à l'itération m

$$\frac{\|x_* - x_m\|_\infty}{\|x_*\|_\infty} < 10^{-6}.$$

Nous observons que le nombre d'itération croit avec le nombre de sous domaines et l'ampleur du saut dans le coefficient à l'exception du cas $\alpha_2 = 10^6$ qui demande moins d'itérations que le cas $\alpha_2 = 10^4$. Le fait que dans le tableau l'estimation du conditionnement ne dépend que du nombre de sous domaines n'est pas une faute de frappe.

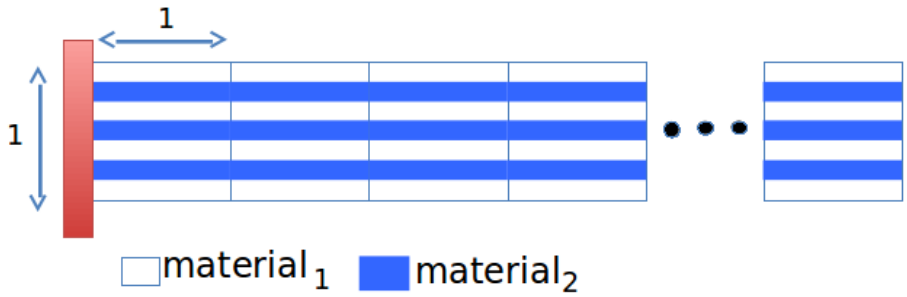


FIGURE 1.5 – Géométrie pour le test de robustesse – le domaine est composé de sept couches de deux différents matériaux. Des conditions de Dirichlet homogènes sont imposées sur le bord gauche et des conditions de Neumann homogènes partout ailleurs. Le nombre de sous domaines N et donc la longueur du domaine varient.

Nombre d'itérations :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	18	33	62	120
$\alpha_2 = 10^2$	24	37	64	117
$\alpha_2 = 10^4$	32	63	117	187
$\alpha_2 = 10^6$	21	51	107	208

Estimation du Conditionnement :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$
$\alpha_2 = 10^2$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$
$\alpha_2 = 10^4$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$
$\alpha_2 = 10^6$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$

TABLE 1.1 – Résultats de convergence pour le **problème scalaire elliptique** (1.19) discrétisé par des éléments finis $\mathbb{P}_1$ avec le **préconditionneur de Schwarz additif** (1.4). La géométrie est présentée dans la Figure 1.5. Deux couches d'éléments sont ajoutées à chaque sous domaine. On fait varier le nombre de sous domaines et le paramètre α_2 dans le matériau 2. On présente ici le nombre d'itérations nécessaire pour converger (en haut) et l'estimation du conditionnement de la matrice préconditionné basée sur les valeurs de Ritz (en bas).

Le préconditionneur de Schwarz que nous avons présenté est loin d'être la version la plus évoluée. Dans la prochaine section nous présenterons un moyen simple et bien connu d'améliorer la convergence et en particulier de récupérer la robustesse dans les cas à coefficients constants. Même avec cette amélioration, le défaut de robustesse à l'égard des hétérogénéités dans les coefficients posera problème et sera une parfaite illustration de la famille de problèmes auxquels nous nous attaquons dans les prochains chapitres de ce manuscrit. Mais d'abord nous discutons le lien entre la robustesse et le choix de la partition en sous domaines.

1.1.4 Agir sur la partition en sous domaines pour améliorer la robustesse

Il est bien connu que les méthodes de décomposition de domaine sont robustes si la partition en sous domaines est choisie d'une certaine manière, voir par exemple [23, 22, 68]. Des généralisations de ces résultats existent (cf. [86, 85, 98]...) et confirment qu'agir sur la partition n'aide pas seulement les analyses théoriques mais accélère aussi la mise en pratique.

FETI et BDD Comme nous l'avons déjà mentionné un avantage significatif des méthodes de décomposition de domaine sans recouvrement est que lorsque l'on effectue des simulations dans des domaines constitués de plusieurs matériaux on peut faire en sorte que les bords des sous domaines coïncident avec les bords des différents matériaux. En d'autres termes la partition en sous domaines accommode les sauts dans les coefficients. En agissant sur la partition de l'unité dans les préconditionneurs FETI et BDD il est alors possible de retrouver une convergence tout aussi bonne que dans le cas à coefficients constants. Pour FETI l'idée remonte à [93] où des poids basés sur les valeurs diagonales de la matrice de rigidité sont introduits et interprétés mécaniquement. La subtilité repose dans le fait que les poids pour les inconnues de déplacements et les poids pour les inconnues de forces à l'interface sont liés (mais différents). Dans [58] une formulation mathématique abstraite de ces poids est introduite qui permet d'écrire l'analyse théorique de la méthode. Les résultats qui correspondent pour BDD peuvent aussi être trouvés dans cet article.

Plus récemment, pour FETI, les auteurs de [87, 84, 83] démontrent que certaines configurations particulières d'hétérogénéités qui ne sont pas *accommodées* par la partition en sous domaine ne nuisent pas non plus à la convergence mais cela reste loin d'être le cas général.

Qu'en est-il de Schwarz ? C'est Pierre-Louis Lions qui en 1990 introduisit la première version de l'algorithme de Schwarz sans recouvrement [66]. L'astuce est de remplacer la condition de transmission de Dirichlet dans (1.2) par une condition de Robin pour un paramètre β :

$$u_1^{n+1} + \beta \frac{\partial}{\partial \mathbf{n}_1} u_1^{n+1} = u_2^n + \beta \frac{\partial}{\partial \mathbf{n}_1} u_2^n \text{ sur } \partial\Omega_2 \cap \partial\Omega_1,$$

dans le premier pas de l'itération et

$$u_2^{n+1} + \beta \frac{\partial}{\partial \mathbf{n}_2} u_2^{n+1} = u_1^{n+1} + \beta \frac{\partial}{\partial \mathbf{n}_2} u_1^{n+1} \text{ sur } \partial\Omega_2 \cap \partial\Omega_1,$$

dans le second pas de l'itération ($\mathbf{n}_1$ et $\mathbf{n}_2$ sont les normales unitaires pour les sous domaines Ω_1 et Ω_2). Lions prouve que l'algorithme appliqué au problème de Poisson converge sans recouvrement quel que soit le nombre de sous domaines. L'idée qui consiste à changer les conditions de transmission a encore été généralisée pour trouver les conditions de

transmissions optimales parmi tous les opérateurs linéaires. Il s'avère que les conditions optimales ne sont pas locales ce qui signifie que le solveur qui en résulte est coûteux à appliquer. Des développements de Taylor tronqués peuvent donner de bons résultats. Parmi la vaste littérature sur le sujet nous renvoyons à [77, 16, 42, 40] et aux références qui y sont présentées.

AGMG Les méthodes multigrilles [48] sont étroitement liées à la décomposition de domaine. Du point de vue de la décomposition de domaine une méthode multigrille est une application récursive de la décomposition de domaine : un domaine global est divisé en sous domaines qui sont à leur tour divisés en sous domaines et ainsi de suite jusqu'à arriver à l'échelle du maillage. Du point de vue des méthodes multigrilles, la décomposition de domaine est une méthode multigrille où, en partant du maillage, une seule étape de déraffinement a été exécutée.

Les méthodes multigrilles algébriques [5, 94] sont des variantes de la méthode de départ où la connaissance des matrices élémentaires n'est pas requise. Ceci convient particulièrement aux cas où la matrice ne découle pas d'un problème discrétisé sur un maillage ou alors aux cas où le maillage est non structuré. Plus généralement ces méthodes sont intéressantes car elles s'implémentent en boite noire.

Pour les problèmes à coefficients hétérogènes les méthodes de multigrille algébrique risquent de construire des *agrégats* (la contrepartie des sous domaines) qui intersectent les hétérogénéités et donc qui nuisent à la convergence. C'est pourquoi dans [75] les auteurs proposent une méthode multigrille algébrique où la convergence est garantie a priori. C'est la première analyse complète d'une méthode multigrille algébrique basée sur l'agrégation simple. Elle repose sur leurs travaux précédents [74]. Le résultat est prouvé pour le cas de M-matrices dont la somme des coefficients sur chaque ligne est positive ou nulle. L'idée est d'agir sur la manière dont on forme les agrégats : partant du résultat de convergence requis défini par l'utilisateur, un critère de qualité pour les agrégats est formulé. Puis les agrégats sont construits de manière adaptative en veillant à ce que le critère de qualité soit toujours satisfait : tout comme avec les méthodes de décomposition de domaine on peut agir sur la partition pour gagner en robustesse.

Pourquoi on ne mise pas sur cette stratégie Dans cette thèse l'un des objectifs est de ne jamais utiliser l'hypothèse que la partition en sous domaines exploite la connaissance des hétérogénéités. Notre objectif est de résoudre des problèmes industriels et plus particulièrement les problèmes Michelin (voir Figure 1.6 pour un exemple simplifié). Sachant cela les raisons pour lesquelles nous avons décidé de ne pas reposer sur une hypothèse concernant la partition sont les suivantes :

- La manière dont les matériaux sont distribués dans la Figure 1.6 suggère que si les bords des sous domaines suivent les hétérogénéités alors ils auront de mauvais aspects de forme. Ceci signifieraient que les problèmes locaux pourraient devenir très mal conditionnés, difficiles à résoudre et exiger des méthodes fines et potentiellement coûteuse (adaptées aux plaques et coques par exemple).
- Si la partition en sous domaines est liée aux matériaux alors elle doit être implémentée dans une partie du code ou les matériaux sont connus. Ceci s'oppose à l'objectif d'avoir un solveur en boite noire qui interfère le moins possible avec les codes existants et futurs.
- L'argument qui est peut être le plus décisif est que les hétérogénéités dans les matériaux ne constituent que l'un des paramètres qui nuisent à la convergence et dont

Figure 1.6 – Une illustration du type de problème auquel est confronté Michelin

on veut s'affranchir. En fait on cherche à construire un solveur qui s'adapte automatiquement à nombreux types de difficultés.

Pour remettre les choses en perspective, rappelons que dans un contexte industriel la question n'est pas seulement : "existe-t-il une partition en sous domaines qui est compatible avec les hétérogénéités et qui conduit à des problèmes locaux bien posés ?" mais plutôt "combien de temps faudrait-il à un ingénieur pour trouver cette partition". S'il est possible de partitionner le domaine en utilisant un outil automatique comme Metis [54] ou Scotch [13] puis de laisser le solveur trouver et contourner les difficultés alors cette possibilité semble très attrayante. C'est avec cet objectif en tête que nous travaillerons à construire des solveurs plus robustes dans les chapitres suivants.

1.2 Méthodes de décomposition de domaine à deux niveaux

Le défaut de robustesse que nous avons mis en évidence dans la section précédente peut s'expliquer par un manque de communication globale entre les sous domaines : au cours d'une itération un sous domaine échange de l'information seulement avec ses voisins ou dans certains cas (FETI et BDD préconditionnés) avec les voisins de ses voisins. Pour cette raison une amélioration possible est d'ajouter un mécanisme de communication globale. C'est ce qu'on appelle une méthode à deux niveaux. L'idée est d'utiliser un solveur direct non seulement dans les sous domaines locaux mais aussi sur un problème qui est commun à tous les sous domaines : le problème grossier. Ce problème grossier est une approximation de A et la manière de le choisir occupera une place importante dans la suite de ce manuscrit. Avant de s'y plonger introduisons ce qu'on appelle la théorie de Schwarz abstraite [61, 112] : un cadre théorique adapté à la formulation et l'étude des méthodes de décomposition de domaine.

1.2.1 Théorie de Schwarz abstraite

Ce qui suit est adapté du livre de Toselli et Widlund [112](Chapitre 2). Nous renvoyons à ce livre pour les détails bibliographiques sur l'émergence de la théorie de Schwarz. Nous mentionnons tout de même les contributions [80] et [117] qui sont souvent jugées importantes. Certains éléments de notation ont déjà été introduits mais ce n'est pas un problème car dans ce cas il s'agit de la généralisation des mêmes notions. Soit un espace de Hilbert V de dimension finie, soit une forme bilinéaire symétrique et coercive

$$a(\cdot,\cdot) : V \times V \to \mathbb{R},$$

et un élément $f \in V'$, considérons le problème de trouver $u \in V$, tel que

$$a(u,v) = f(v),\ v \in V. \tag{1.20}$$

Si A est la matrice de rigidité associée à la forme bilinéaire $a(\cdot,\cdot)$ dans une certaine base de V, si $\mathbf{f}$ est le vecteur associé à f dans la même base alors le problème (1.20) est équivalent au système linéaire

$$A\mathbf{u} = \mathbf{f}, \tag{1.21}$$

où A est symétrique, définie et positive. On considère à présent une famille d'espaces $\{V_i, i = 0,\dots,N\}$ et on suppose qu'il existe des opérateurs d'interpolation

$$R_i^\top : V_i \to V.$$

Supposons aussi que V s'écrit de la manière suivante (la somme n'étant pas nécessairement directe)

$$V = R_0^\top V_0 + \sum_{i=1}^{N} R_i^\top V_i. \tag{1.22}$$

Remarquons que les sous espaces sont désormais numérotés de 0 à N. Ceci ne change pas la définition au niveau abstrait mais dans de nombreux cas V_0 sera un espace bien particulier : l'espace grossier, tandis que les N autres espaces V_i seront les sous domaines habituels basés sur la géométrie.

Introduisons des formes bilinéaires locales et supposons qu'elles sont symétriques et coercives aussi

$$\tilde{a}_i(\cdot,\cdot) : V_i \times V_i \to \mathbb{R}, \quad i = 0,\dots,N,$$

et que les matrices de rigidité qui leurs sont associées sont les matrices

$$\tilde{A}_i : V_i \to V_i.$$

Les opérateurs de Schwarz sont définis à partir des opérateurs suivants :

$$P_i = R_i^\top \tilde{P}_i : V \to R_i^\top V_i \subset V, i = 0,\dots,N,$$

où $\tilde{P}_i : V \to V_i$, est défini par

$$\tilde{a}_i(\tilde{P}_i u, v_i) = a(u, R_i^\top v_i), \quad v_i \in V_i. \tag{1.23}$$

On remarque que $\tilde{P}_i$ est bien défini puisque les formes bilinéaires sont coercives.

Remarque 1.2. Dans le cas où on choisit d'utiliser la forme bilinéaire de départ sur un des sous espaces V_i elle s'écrit

$$\tilde{a}_i(u_i, v_i) = a(R_i^\top u_i, R_i^\top v_i), \quad u_i, v_i \in V_i \tag{1.24}$$

et on trouve que

$$\tilde{A}_i = R_i A R_i^\top = A_i. \tag{1.25}$$

Dans ce cas on dit qu'on utilise un *solveur exact sur* V_i.

Le lemme suivant se démontre aisément.

Lemma 1.1. Les P_i s'écrivent avec la formulation

$$P_i = R_i^\top \tilde{A}_i^{-1} R_i A, \quad 0 \leq i \leq N. \tag{1.26}$$

De plus les matrices P_i sont auto-adjointes pour le produit scalaire induit par $a(\cdot, \cdot)$ et elles sont positives. Si on a choisi la forme bilinéaire $\tilde{a}_i$ suivant (1.24) alors P_i est une projection c'est à dire

$$P_i^2 = P_i. \tag{1.27}$$

A partir de maintenant on fait l'hypothèse suivante.

Hypothèse 1.3. *Un solveur exact est utilisé sur l'espace grossier* V_0. *(Et donc* P_0 *est une projection A-orthogonale).*

A partir des opérateurs P_i on peut définir trois familles d'opérateurs de Schwarz

1. Opérateurs Additifs :

$$P_{ad} := \sum_{i=0}^{N} P_i. \tag{1.28}$$

2. Opérateurs Multiplicatifs :

$$P_{mu} := I - (I - P_N)(I - P_{N-1}) \ \ldots (I - P_0). \tag{1.29}$$

3. Opérateurs Hybrides :

$$P_{hy} := P_0 + (I - P_0) \sum_{i=1}^{N} P_i (I - P_0). \tag{1.30}$$

Les preuves de convergence dans la théorie de Schwarz abstraite reposent sur l'hypothèse 1.3 et trois hypothèses supplémentaires.

Hypothèse 1.4 (Inégalités de Cauchy-Schwarz généralisées)**.** *Il existe des constantes* $0 \leq \epsilon_{ij} \leq 1$, $1 \leq i, j \leq N$, *telles que*

$$|a(R_i^\top u_i, R_j^\top u_j)| \leq \epsilon_{ij} a(R_i^\top u_i, R_i^\top u_i)^{1/2} a(R_j^\top u_j, R_j^\top u_j)^{1/2},$$

pour $u_i \in V_i$ *et* $u_j \in V_j$. *Nous dénotons le rayon spectral de* $\epsilon = \{\epsilon_{ij}\}$ *par* $\rho(\epsilon)$.

Hypothèse 1.5 (Stabilité locale des solveurs)**.** *Il existe* $\omega > 0$ *tel que*

$$a(R_i^\top u_i, R_i^\top u_i) \leq \omega \tilde{a}_i(u_i, u_i), \quad u_i \in Im(\tilde{P}_i) \subset V_i, \, 1 \leq i \leq N. \tag{1.31}$$

Hypothèse 1.6 (Existence d'une décomposition stable)**.** *Il existe une constante C_0 telle que chaque $u \in V$ admette une décomposition*

$$u = \sum_{i=0}^{N} R_i^\top u_i, \quad \{u_i \in V_i, 0 \leq i \leq N\}$$

qui vérifie

$$\sum_{i=0}^{N} \tilde{a}_i(u_i, u_i) \leq C_0^2 a(u, u).$$

Le théorème suivant donne des résultats de convergence pour les méthodes de Schwarz

Théorème 1.7. *Sous les hypothèses 1.3, 1.4, 1.5 et 1.6 les opérateurs définis par* (1.28)*,* (1.29) *et* (1.30) *vérifient, pour tout $u \in V$,*

$$\begin{aligned} C_0^{-2} a(u,u) \leq \quad & a(P_{ad}u, u) \quad \leq \omega(\rho(\epsilon)+1) a(u,u), \\ \max(1, C_0^2)^{-1} a(u,u) \leq \quad & a(P_{hy}u, u) \quad \leq \max(1, \omega\rho(\epsilon)) a(u,u). \end{aligned}$$

et, sous l'hypothèse que $\omega < 2$,

$$\|I - P_{mu}\|_A \leq 1 - \frac{2-\omega}{(2\max(1,\omega)^2 \rho(\epsilon)^2 + 1) C_0^2} < 1,$$

Plus précisément, pour l'opérateur hybride il est suffisant de démontrer l'hypothèse 1.6 sur $\mathrm{Im}(I - P_0)$. Les opérateurs additifs et hybrides P_{ad} et P_{hy} sont le produit d'une matrice symétrique par un préconditionneur symétrique et donc ils seront résolus avec le gradient conjugué préconditionné. Comme on l'a déjà exhibé dans (1.17) le taux de convergence du gradient conjugué préconditionné est borné par une quantité qui ne dépend que des valeurs extrêmes du spectre de l'opérateur préconditionné. Ces quantités sont à leur tour liées aux quotients de Rayleigh du résultat du théorème et donc les deux premiers résultats dans le Théorème 1.7 sont bien des résultats de convergence. La variante multiplicative P_{mu} n'est pas symétrique. Plutôt qu'un préconditionneur pour une méthode itérative P_{mu} servirait plutôt dans un algorithme de type Richardson. La norme dans le théorème est alors la norme (induite par A) du propagateur d'erreur et elle est bien liée à un résultat de convergence.

En général l'hypothèse 1.4 est démontrée en appliquant le lemme suivant. On y apprend qu'une borne pour la constante $\rho(\epsilon)$ dans le résultat de convergence dépend seulement de la géométrie du découpage en sous domaines (mais pas du nombre de sous domaines).

Lemme 1.8. *Supposons que les sous espaces locaux V_i, $i = 1, \ldots, N$ ont été coloriés de manière à ce que deux sous domaines V_k et V_l qui ont la même couleur soient orthogonaux*

$$P_k P_l = P_l P_k = 0$$

et que N^C couleurs ont été nécessaires. Alors l'hypothèse 1.4 est satisfaite et $\rho(\epsilon) \leq N^C$.

Pour la preuve de tous ces résultats et plus de détails sur les méthodes de Schwarz nous référons de nouveau le lecteur à [112](Chapitre 2).

Le préconditionneur de Schwarz additif introduit dans la sous section 1.1.1 est P_{ad} pour un espace grossier vide $V_0 = \emptyset$ et des solveurs locaux exacts : les $\tilde{a}_i$ sont définies selon (1.24). En supposant que l'espace grossier V_0 est non vide le préconditionneur de Schwarz additif à deux niveaux est défini comme suit.

Définition 1.9. Le préconditionneur de Schwarz à deux niveaux est

$$M^{-1} := \sum_{i=0}^{N} R_i^\top A_i^{-1} R_i, A_i := R_i A R_i^\top, \text{ pour } i = 0, \ldots, N. \tag{1.32}$$

Remarque 1.10. Puisque les solveurs locaux sont tous des solveurs exact l'hypothèse 1.5 est automatiquement vérifiée pour $\omega = 1$. Dans ce cas, selon le lemme 1.8 et le théorème 1.7 le conditionnement de $M^{-1}A$ dépend seulement de l'existence d'une décomposition stable dans le sens donné par l'hypothèse 1.6 et si elle est vérifiée le conditionnement de $M^{-1}A$ est borné par $C_0^{-2}(N^C + 1)$. Dans cette expression le nombre de couleurs N^C peut être remplacé par le plus grand nombre de sous domaines auxquels appartient un élément du maillage [25, Section 4].

1.2.2 Espaces grossiers basés sur les noyaux des opérateurs

Nous introduisons la famille la plus simple d'espaces grossiers qui est constituée des noyaux de certains opérateurs locaux. Le fait qu'un bon espace grossier doit contenir au moins ces vecteurs est maintenant bien connu dans la littérature. Dans certains cas (FETI ou BDD avec un préconditionneur) c'est même indispensable pour que les opérateurs soient bien définis.

Le préconditionneur de Schwarz additif et les espaces grossiers Nicolaides et Partition de l'unité Considérant le problème de Poisson sur un domaine Ω, Nicolaides a proposé en 1987 [80] d'accélérer la convergence d'un solveur itératif en partitionnant le domaine Ω en sous domaines disjoints Ω_i^*, $i = 1, \ldots, N$ puis en utilisant l'espace des fonctions constantes par sous domaine comme un espace grossier (puisqu'il fait ceci hors du cadre de la décomposition de domaine on parle plutôt de déflation) :

$$V_0^{NICO} = \text{span}(\mathbf{1}_{\Omega_1^*}, \ldots, \mathbf{1}_{\Omega_N^*}),$$

où $\mathbf{1}_{\Omega_i^*}$ est la fonction indicatrice de Ω_i^*.

Il y a un inconvénient important à l'espace grossier de Nicolaides : les fonctions de base ont une énergie qui est de l'ordre de H/h où H est la taille d'un sous domaine et h est le pas du maillage. Pour cette raison on ne peut pas s'attendre à ce que la convergence de la méthode à deux niveaux correspondante soit indépendante du nombre de mailles. La solution est de remplacer les fonctions indicatrices par des fonctions plus régulières.

Dans [97] Sarkis introduit et analyse un espace grossier engendré par une fonction de base par sous domaine et dont l'ensemble constitue une partition de l'unité sur Ω. Il prouve que pour le problème de Poisson la méthode à deux niveaux converge indépendamment de la taille des sous domaines et du pas du maillage. Plus précisémment l'estimation du conditionnement varie linéairement avec la fraction du volume d'un sous domaine qui est recouverte par ses voisins :

$$\kappa(M^{-1}A) \leq C\left(1 + \frac{H}{\delta}\right), \tag{1.33}$$

où H est la taille du sous domaine, δ est la largeur du recouvrement et C est une constante qui dépend de la géométrie du partitionnement mais pas de H, δ ou le pas du maillage h. Certaines hypothèses sur la régularité des sous domaines sont aussi requises.

Nous renvoyons vers [97] (ou [112](Lemme 3.24)) pour la construction précise de l'espace grossier basé sur une partition de l'unité et pour la preuve du résultat de convergence. Dans le cas où le maillage est régulier, pour chaque $i = 1, \ldots, N$ la fonction de base prend

Nombre d'itérations :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	18	24	24	23
$\alpha_2 = 10^2$	22	25	25	24
$\alpha_2 = 10^4$	36	62	95	128
$\alpha_2 = 10^6$	31	51	89	154

Estimation du Conditionnement :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	28.0	28.2	28.1	28.1
$\alpha_2 = 10^2$	28.0	28.2	28.1	28.1
$\alpha_2 = 10^4$	415	$1.29 \cdot 10^3$	$2.39 \cdot 10^3$	$1.36 \cdot 10^3$
$\alpha_2 = 10^6$	479	$2.02 \cdot 10^3$	$7.96 \cdot 10^3$	$2.76 \cdot 10^4$

TABLE 1.2 – Résultats de convergence pour le **problème scalaire elliptique** (1.19) discrétisé par des éléments finis $\mathbb{P}_1$ avec le préconditionneur de Schwarz à deux niveaux (1.32) et V_0 est l'espace grossier **Partition de l'unité** (fonctions constantes à l'intérieur d'un sous domaine, nulles en dehors du sous domaine et qui décroissent linéairement dans le recouvrement). La géométrie est présentée dans la Figure 1.5. Deux couches d'éléments sont ajoutées à chaque sous domaine. On fait varier le nombre de sous domaines et le paramètre α_2 dans le matériau 2. On présente ici le nombre d'itérations nécessaire pour converger (en haut) et l'estimation du conditionnement de la matrice préconditionné basée sur les valeurs de Ritz (en bas).

la valeur 1 dans la partie de Ω_i qui n'est pas recouverte par les sous domaines voisins, 0 en dehors de Ω_i et décroit linéairement de 1 vers 0 dans le recouvrement. La raison principale pour laquelle les fonctions constantes par sous domaines ont besoin d'être dans l'espace grossier est ce qu'on appelle dans [112] l'argument de la topologie quotient. Plus précisément, un outil très important est l'inégalité de Poincaré : supposons que $1 \leq p \leq \infty$ et que $\tilde{\Omega}$ est un domaine ouvert, connexe et lipschitzien de $\mathbb{R}^n$. Dans ce cas il existe une constante C qui dépend seulement de $\tilde{\Omega}$ et de p telle que chaque fonction u de l'espace de Sobolev $W^{1,p}(\tilde{\Omega})$ vérifie

$$\|u - u_{\tilde{\Omega}}\|_{L^p(\tilde{\Omega})} \leq C\|\nabla u\|_{L^p(\tilde{\Omega})}; \quad u_{\tilde{\Omega}} = \frac{1}{|\tilde{\Omega}|}u(y)dy.$$

Grâce au choix particulier de l'espace grossier cette inégalité (pour $p = 2$) peut être appliquée localement à des fonctions locales à moyenne nulle.

On illustre à présent numériquement l'efficacité de l'espace grossier basé sur une partition de l'unité sur le même test de robustesse que pour la méthode à un niveau (pour lesquels les résultats sont dans le tableau 1.1). Cette fois on reporte les résultats dans le tableau 1.2, et on remarque que si le coefficient est constant ($\alpha_2 = \alpha_1 = 1$) ou varie peu ($\alpha_2 = 100$) la convergence n'est plus détériorée par l'augmentation du nombre de sous domaines. Par contre l'espace grossier basé sur une partition de l'unité ne suffit pas à assurer la robustesse lorsque les sauts dans les coefficients deviennent très significatifs. Le cas des coefficients très hétérogènes est un bon exemple des problèmes auxquels on s'intéressera dans les chapitres suivants.

BDD, FETI et l'espace des modes rigides Rappelons que selon (1.11) et (1.12) la formulation BDD du problème d'élasticité est

$$\left(\sum_{i=1}^{N} R_i^\top S_i R_i\right) \hat{\mathbf{u}}^\Gamma = \sum_{i=1}^{N} R_i^\top \tilde{\mathbf{f}}_i \text{ préconditionné par } M^{-1} = \sum_{i=1}^{N} \widetilde{R_i^\top} S_i^{-1} \widetilde{R_i}.$$

Nous avons jusqu'à présent supposé que les inverses S_i^{-1} sont définis. Ceci n'est généralement pas le cas et un préconditionneur pour BDD s'écrirait plutôt $\left(\sum_{i=1}^{N} \widetilde{R_i^\top} S_i^\dagger \widetilde{R_i}\right)$ où $S_i^\dagger$ est un pseudo inverse de S_i. Puisque l'action de $S_i^\dagger$ est définie seulement sur $\mathrm{Im}(S_i)$ il est nécessaire d'introduire des opérateurs de projection qui permettent de s'assurer que le résidu vit toujours dans cet espace. On fait cela en utilisant un préconditionneur de type hybride (comme ceux définis par (1.30)). Si les poids dans les opérateurs $\widetilde{R_i^\top}$ s'écrivent $R_i^\top D_i^{-1}$ pour une matrice diagonale D_i alors l'opérateur BDD [67] est

$$P_{bdd} = P_0 + (I - P_0)\left(\sum_{i=1}^{N} \widetilde{R_i^\top} S_i^\dagger \widetilde{R_i}\right) \underbrace{\left(\sum_{i=1}^{N} R_i^\top S_i R_i\right)}_{:=\hat{S}} (I - P_0), \tag{1.34}$$

où le projecteur grossier P_0, l'opérateur d'interpolation $R_0^\top : V_0 \to V$ et l'espace grossier V_0 sont définis par

$$P_0 := R_0^\top S_0^{-1} R_0 \hat{S},\ S_0 := R_0 \hat{S} R_0^\top,\ V_0 := \mathrm{Im}(R_0^\top) = \sum_{i=1}^{N} R_i^\top D_i \left(\mathrm{Ker}(S_i)\right).$$

On appelle V_0 l'espace des modes rigides car dans le cas de l'élasticité le noyau de S_i est la trace des modes rigides sur le bord du sous domaine.

Pour FETI les choses sont un peu plus compliquées. En effet, selon (1.15) et (1.16) la formulation FETI du problème d'élasticité est

$$\left(\sum_{i=1}^{N} B_i S_i^{-1} B_i^\top\right) \boldsymbol{\lambda} = \sum_{i=1}^{N} B_i S_i^{-1} \tilde{\mathbf{f}}_i \text{ préconditionné par } M^{-1} = \sum_{i=1}^{N} \widetilde{B_i} S_i \widetilde{B_i}.$$

Dès que l'un des S_i est non inversible la reformulation du problème doit être adaptée. En effet, si $\mathcal{R}_i^\top$ est un interpolateur de $\mathbb{R}^{\dim(\mathrm{Ker}(S_i))}$ dans le noyau de S_i alors $\mathbf{u}_i^\Gamma = S_i^{-1}(\tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda})$ dans (1.14) doit être remplacé par

$$\mathbf{u}_i^\Gamma = S_i^\dagger(\tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda}) + \mathcal{R}_i^\top \boldsymbol{\alpha_i}, \quad \boldsymbol{\alpha_i} \in \mathbb{R}^{\dim(\mathrm{Ker}(S_i))}.$$

Nous ne souhaitons pas rentrer dans les détails ici, ils seront présentés dans le Chapitre 5. Ce qui est important est que même dans le cas où S est non inversible il est possible de réécrire le problème d'élasticité en fonction des forces d'interface. La différence majeure est que les forces, au lieu d'appartenir simplement à $U = \mathrm{Im}\left(\sum_{i=1}^{N} B_i\right)$, doivent être dans l'espace des contraintes admissibles

$$\{\boldsymbol{\lambda} \in U; G^\top \boldsymbol{\lambda} = \sum_{i=1}^{N} \mathcal{R}_i^\top \tilde{\mathbf{f}}_i\} \text{ où } G := \sum_{i=1}^{N} B_i \mathcal{R}_i^\top.$$

Pour cette raison le solveur itératif pour FETI est le Gradient Conjugué Préconditionné et Projeté (PPCG) : il est initialisé avec $\boldsymbol{\lambda}_0$ qui satisfait la contrainte ($G^\top \boldsymbol{\lambda}_0 = \sum_{i=1}^{N} \mathcal{R}_i^\top \tilde{\mathbf{f}}_i$)

puis toutes les directions de recherche sont projetées dans le noyau de $G^\top$ afin que les approximations successives retournée par PPCG satisfasse la contrainte. Si P est un opérateur de projection dont l'image est $\mathrm{Ker}(G^\top)$ alors l'opérateur de FETI préconditionné s'écrit

$$P_{FETI} = P \underbrace{\left(\sum_{i=1}^{N} \widetilde{B_i} S_i \widetilde{B_i} \right)}_{M^{-1}} P^\top \underbrace{\left(\sum_{i=1}^{N} B_i S_i^{-1} B_i^\top \right)}_{:=F}.$$

Une autre grande différence avec la méthode de Schwarz additive et BDD est qu'on ne peut pas utiliser un solveur exact pour définir la projection P afin d'obtenir une projection F-orthogonale. En effet la raison pour laquelle on utilise P est justement de s'occuper d'un espace où F n'est pas défini. A la place on utilise la meilleure approximation de F dont on dispose : $(M^{-1})^{-1}$. Plus précisément l'opérateur de projection est

$$P = I - M^{-1} G \left(G^\top M^{-1} G \right)^{-1} G^\top,$$

et il est $(M^{-1})^{-1}$-orthogonal.

L'algorithme PPCG appliqué à ce problème est présenté dans l'algorithme 1.2. Il s'applique bien sûr à tous les problèmes de type Schwarz hybride (1.30). Pour une étude détaillée des différentes alternatives qui permettent de résoudre ce problème voir [110, 56] et les références qui y sont présentées.

Algorithm 1.2 PPCG : Algorithme de Gradient Conjugué Projeté et Préconditionné pour résoudre $PM^{-1}P^\top F\boldsymbol{\lambda} = PM^{-1}P^\top \mathbf{d}$ (où $\mathbf{d} := BS^\dagger \mathbf{f}$ est le membre de droite pour FETI)

$\boldsymbol{\lambda}_0 := M^{-1}G \left(G^\top M^{-1} G \right)^{-1} \sum_{i=1}^{N} \mathcal{R}_i^\top \tilde{\mathbf{f}}_i$
$\mathbf{r}_0 := P^\top (\mathbf{d} - F\boldsymbol{\lambda}_0)$; $\mathbf{z}_0 := M^{-1}\mathbf{r}_0$; $\mathbf{p}_0 := \mathbf{z}_0$
for $j = 0, 1, \ldots$ jusque convergence **do**
  $\mathbf{p}_j = P\mathbf{p}_j$
  $\alpha_j := \langle \mathbf{r}_j, \mathbf{z}_j \rangle / \langle F\mathbf{p}_j, \mathbf{p}_j \rangle$
  $\boldsymbol{\lambda}_{j+1} := \boldsymbol{\lambda}_j + \alpha_j \mathbf{p}_j$
  $\mathbf{r}_{j+1} := \mathbf{r}_j - \alpha_j P^\top F\mathbf{p}_j$
  $\mathbf{z}_{j+1} := M^{-1}\mathbf{r}_{j+1}$
  $\beta_j := \langle \mathbf{r}_{j+1}, \mathbf{z}_{j+1} \rangle / \langle \mathbf{r}_j, \mathbf{z}_j \rangle$
  $\mathbf{p}_{j+1} := \mathbf{z}_{j+1} + \beta_j \mathbf{p}_j$
end for

Rappelons que pour le préconditionneur de Schwarz avec l'espace grossier basé sur une partition de l'unité il est possible de montrer que la convergence sur le problème de Poisson ne dépend pas du nombre de sous domaines (voir (1.33)), mais qu'à la place elle dépend de la taille relative du recouvrement H/δ. Pour le problème scalaire elliptique, FETI est par construction équipé d'un second niveau lié au noyau de S qui se trouve être la trace des fonctions constantes sur le bord de chaque sous domaine. Pour cette raison il est assez naturel qu'une fois de plus un résultat de convergence qui ne dépend que faiblement du nombre de sous domaines (à travers le nombre de mailles par sous domaine) soit disponible [58]

$$\kappa(PM^{-1}P^\top F_{|\mathrm{range}(P)}) \leq C \left(1 + \log\left(\frac{H}{h} \right) \right)^2, \tag{1.35}$$

où H est la taille d'un sous domaine et h est la taille d'une maille. Des hypothèses de régularité sur les sous domaines sont requises.

1.2.3 Espace grossiers analytiques

Il existe des problèmes pour lesquels un bon espace grossier, voire même l'espace grossier optimal est connu dans la littérature. Ici on présente certaines de ces contributions.

Espaces grossiers basés sur le coefficient Pour le problème scalaire elliptique en milieu binaire (où deux matériaux coexistent) les auteurs de [115] considèrent le même problème que celui de la Figure 1.5 (des couches de matériaux). Leur objectif est d'appliquer PPCG au problème de Poisson en choisissant astucieusement l'espace projeté. Ils utilisent une factorisation de Cholesky incomplète dont le comportement est comparable à un préconditionneur de type Schwarz puisqu'il traite les grandes valeurs propres et laisse à régler le problème des plus petites valeurs propres. Leur conclusion est que l'espace grossier doit être constitué d'autant de vecteurs qu'il y a de couches à fort coefficient. Dans [47] la distribution des coefficients est constitué de nombreuses petites inclusions à fort coefficient qui n'intersectent le bord de chaque sous domaine qu'au plus une fois. Il est démontré qu'une seule fonction de base par sous domaine suffit à assurer la robustesse.

Autres espaces grossiers Pour FETI le premier espace grossier est introduit par [32] pour des problèmes d'élasticité instationnaires. Ces problèmes ont un terme d'ordre zéro qui les rend non singuliers et donc il n'y a pas d'espace grossier naturel lié aux modes rigides. Les auteurs de [32] proposent d'utiliser quand même un espace grossier avec les modes rigides.

Un autre problème pour lequel la recherche d'un espace grossier a été très active est celui des plaques et des coques. Dans [35] un espace grossier est proposé pour la résolution des problèmes de plaques avec FETI puis il est adapté dans [33] aux coques. Pour BDD un espace grossier pour les problèmes de plaques et de coques est introduit, analysé et testé numériquement par [63].

Finalement, nous notons que dans le cadre de l'élasticité linéaire les auteurs de [17, 18] proposent puis améliorent un espace grossier pour la méthode de Schwarz à deux niveaux pour lequel ils prouvent des résultats de convergence indépendants des propriétés du matériau et donc de la possibilité qu'il soit quasiment incompressible. Pour FETI un seul vecteur par sous domaine est nécessaire [114]. Plus récemment [43] arrive à la même conclusion et analyse théoriquement le comportement de cet espace grossier dans la limite incompressible.

Notre objectif Les espaces grossiers que nous venons de mentionner sont optimaux dans le sens où, étant donné un découpage en sous domaines, on ne peut pas espérer trouver un espace grossier plus petit qui assure la robustesse. L'inconvénient est que ces espaces ne peuvent pas être construit automatiquement sans connaitre *a priori* le type de difficultés que l'on trouve dans le problème. Notre ambition dans ce manuscrit est de définir des espaces grossiers qui peuvent gérer tous les types d'hétérogénéités dans les coefficients ainsi que d'autres difficultés et qui sont construits de manière automatique et implémentés en boite noire. Dans les cas où des espace grossiers optimaux sont connus ils constituent de bons points de comparaison pour nos nouvelles méthodes. En particulier nous les testerons sur des cas à coefficients discontinus ainsi que sur le système de l'élasticité linéaire dans la limite incompressible. Les problèmes aux valeurs propres généralisés constitueront un un outil fondamental.

1.2.4 Espaces grossiers qui utilisent des problèmes aux valeurs propres

Ce qu'apportent les problèmes aux valeurs propres généralisés Une fois qu'une méthode de décomposition de domaine a été écrite dans le formalisme de Schwarz afin de prouver qu'elle va converger il suffit de montre que les solveurs locaux sont stables (Hypothèse 1.5) et que chaque vecteur admet une décomposition stable (Hypothèse 1.6). Ces conditions sont des inégalités entre des normes induites par différents produits scalaires et la robustesse repose sur le fait qu'elles soient vérifiées avec des constantes qui ne dépendent pas de certains paramètres. En gardant cet objectif à l'esprit nous introduisons les problèmes aux valeurs propres généralisés ainsi que les propriétés des spectres qu'ils induisent dont nous aurons besoin par la suite.

Définition 1.11 (Problèmes aux valeurs propres généralisés)**.** Soit $\widetilde{A}$ et $\widetilde{B}$ deux matrices symétriques de l'espace $\mathbb{R}^{n\times n}$. Les valeurs propres généralisées associées au couple $(\widetilde{A},\widetilde{B})$ sont $\lambda \in \mathbb{R}\cup\{+\infty\}$ tels que :

– $\lambda \in \mathbb{R}$ et il existe $\mathbf{x}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$ tel que

$$\widetilde{A}\mathbf{x} = \lambda\,\widetilde{B}\mathbf{x}, \tag{1.36}$$

– ou alors $\lambda = +\infty$ et il existe $\mathbf{x}\in\mathbb{R}^n\backslash\{\mathbf{0}\}$ tel que

$$\widetilde{B}\mathbf{x} = \mathbf{0},\ \text{et}\ \widetilde{A}\mathbf{x}\neq\mathbf{0}.$$

Dans les deux cas $\mathbf{x}$ est un vecteur propre généralisé associé à la valeur propre λ pour le couple $(\tilde{A},\tilde{B})$.

La définition ci-dessus prévoit l'existence de valeurs propres généralisées infinies. Une manière de comprendre pourquoi c'est tout à fait naturel est de se rendre compte que si $(+\infty,\mathbf{x})$ est un couple (valeur propre, vecteur propre) pour $(\widetilde{A},\widetilde{B})$ alors $(0,\mathbf{x})$ est un couple (valeur propre, vecteur propre) pour $(\widetilde{B},\widetilde{A})$ et il n'y a aucune raison d'introduire une discrimination entre ces deux formulations. Si la matrice $\tilde{B}$ est définie alors par définition toutes les valeurs propres sont finies et le lemme suivant donne une propriété fondamentale du spectre.

Lemme 1.12. *Soit $\tilde{A}\in\mathbb{R}^{n\times n}$ une matrice symétrique et $\tilde{B}\in\mathbb{R}^{n\times n}$ une matrice symétrique définie positive. L'ensemble des vecteurs propres généralisés $\{\mathbf{x}^k\}_{k=1,\dots,n}$ associé au couple $(\tilde{A},\tilde{B})$ peut être choisi de manière à former une base $\tilde{B}$-orthonormale de $\mathbb{R}^n$:*

$$\langle\mathbf{x}^k,\tilde{B}\mathbf{x}^k\rangle = 1,\ \textit{pour tout}\ k=1,\dots,n\ \textit{et}\ \langle\mathbf{x}^k,\tilde{B}\mathbf{x}^l\rangle = 0,\ \textit{pour tout}\ k,l=1,\dots,n;\ k\neq l.$$

On a alors aussi pour tout $k=1,\dots,n$

$$\langle\mathbf{x}^k,\tilde{A}\mathbf{x}^k\rangle = \lambda^k,\ \textit{et}\ \langle\mathbf{x}^k,\tilde{A}\mathbf{x}^l\rangle = 0,\ \textit{si}\ l=1,\dots,n;\ l\neq k.$$

Démonstration. Cette preuve est en majeure partie la réécriture de la preuve dans [64]. Un résultat bien connu est que pour une matrice réelle symétrique $\tilde{M}\in\mathbb{R}^{n\times n}$, il existe une base orthonormale $\{\mathbf{y}^k\}_{k=1,\dots,N}$ de $\mathbb{R}^n$ qui est constituée des vecteurs propres $\mathbf{y}^k$ de $\tilde{M}$. Ceci se réécrit :

$$\tilde{M}\mathbf{y}^k = \lambda^k\mathbf{x}^k;\quad \langle\mathbf{y}^k,\mathbf{y}^k\rangle = 1;\quad \text{et}\quad \langle\mathbf{y}^k,\mathbf{y}^l\rangle = 0,\ \text{si}\ k\neq l. \tag{1.37}$$

La manière de prouver le lemme est de réduire le problème aux valeurs propres généralisé (1.36) en un problème aux valeurs propres classique. Tout d'abord remarquons que le fait

d'écrire le problème sous la forme $\tilde{B}^{-1}\tilde{A}\mathbf{x}^k = \lambda^k\mathbf{x}^k$ n'apporte rien car le produit $\tilde{B}^{-1}\tilde{A}$ n'est en général pas symétrique. A la place on utilise le fait que, $\tilde{B}$ étant une matrice symétrique elle admet une factorisation de Cholesky :

$$\tilde{B} = LL^\top; \text{ où } L \text{ et une matrice triangulaire inférieure inversible.}$$

Avec cela (1.36) se réécrit $\tilde{M}\mathbf{y}^k = \lambda^k\mathbf{y}^k$ pour $\tilde{M} = L^{-1}\tilde{A}L^{\top-1}$, $\mathbf{y}^k = L^\top\mathbf{x}^k$. Supposons que les $\mathbf{y}^k$ ont été choisis de manière à ce que (1.37) soit vérifiée, alors l'ensemble des vecteurs $\mathbf{x}^k = (L^\top)^{-1}\mathbf{y}^k$ constitue la base de $\mathbb{R}^n$ que nous cherchons puisqu'elle satisfait les conditions suivantes :

$$\tilde{A}\mathbf{x}^k = \lambda^k\tilde{B}\mathbf{x}^k; \quad \langle\mathbf{x}^k, \tilde{B}\mathbf{x}^k\rangle = 1; \quad \text{et} \quad \langle\mathbf{x}^k, \tilde{B}\mathbf{x}^l\rangle = 0,\ k \neq l.$$

Maintenant le fait que $\langle\mathbf{x}^k, \tilde{A}\mathbf{x}^k\rangle = \lambda^k$ est évident. Quant à la dernière propriété dans le lemme, soient $(\lambda^k, \mathbf{x}^k)$ et $(\lambda^l, \mathbf{x}^l)$ deux couples (valeur propre, vecteur propre) généralisés avec $k \neq l$. Supposons que $\lambda^l \neq 0$, dans ce cas

$$\left(\langle\mathbf{x}^k, \tilde{B}\mathbf{x}^l\rangle = 0 \text{ and } \tilde{A}\mathbf{x}^l = \lambda^l\tilde{B}\mathbf{x}^l\right) \Rightarrow \frac{1}{\lambda^l}\langle\mathbf{x}^k, \tilde{A}\mathbf{x}^l\rangle = 0 \Rightarrow \langle\mathbf{x}^k, \tilde{A}\mathbf{x}^l\rangle = 0.$$

Si $\lambda^l = 0$ alors $\mathbf{x}^l \in \mathrm{Ker}(\tilde{A})$ donc $\langle\mathbf{x}^k, \tilde{A}\mathbf{x}^l\rangle = 0$ dans ce cas aussi. □

Une conséquence directe est le lemme suivant qui nous laisse entrevoir comment, avec un problème aux valeurs propres généralisé, on identifie l'espace où des inégalités de la forme donnée dans les Hypothèses 1.5 et 1.6 sont satisfaites.

Lemme 1.13. *Avec les notations introduites dans le lemme 1.12 et étant donné un critère $\tau \in \mathbb{R}^+$ on définit les espaces*

$$E_1 = \mathit{Vect}\{\mathbf{x}^k; \lambda_k < \tau\} \text{ et } E_2 = \mathit{Vect}\{\mathbf{x}^k; \lambda_k \geq \tau\}.$$

On a alors

$$\begin{cases} \langle\mathbf{x}, \tilde{A}\mathbf{x}\rangle < \tau\langle\mathbf{x}, \tilde{B}\mathbf{x}\rangle, & \textit{pour tout } \mathbf{x} \in E_1, \\ \langle\mathbf{x}, \tilde{A}\mathbf{x}\rangle \geq \tau\langle\mathbf{x}, \tilde{B}\mathbf{x}\rangle, & \textit{pour tout } \mathbf{x} \in E_2. \end{cases}$$

État de l'art En pratique la résolution d'un problème aux valeurs propres généralisés qui implique la matrice globale est plus coûteux que la résolution du système linéaire. Pour cette raison avant d'utiliser un problème aux valeurs propres il est nécessaire de réécrire l'estimation que l'on veut satisfaire sous une forme locale. On aura alors un problème aux valeurs propres généralisés par sous domaine et ils pourront être résolus en parallèle. Il nous semble que cette stratégie remonte à [7] où elle a été appliquée pour construire une méthode multigrille algébrique avec agrégation basée sur les éléments (AMGe) pour laquelle on peut choisir a priori la vitesse de convergence que l'on souhaite atteindre. Les mêmes idées sont au fondement de la méthode *spectral AMG* [12]. Depuis, cette stratégie a été une direction de recherche prolifique et en particulier les méthodes introduites dans les chapitres suivant s'appuient sur ces idées fondamentales.

Plus récemment, de nombreuses contributions ont proposé de construire des espaces grossiers pour des problèmes à coefficients fortement hétérogènes en résolvant des problèmes aux valeurs propres dans les sous domaines. Comparé aux premiers travaux sur le méthode AMG cette nouvelle vague d'articles se différencie en utilisant des problèmes aux valeurs propres généralisés. On distingue trois familles de méthodes qui se distinguent

par le choix de la matrice dans l'un des termes du problème aux valeurs propres généralisés. Dans [38, 39], cette matrice est la matrice de masse du problème aux éléments finis ou une version "homogénéisée" de la matrice de masse qui s'obtient en utilisant des fonctions de partitions de l'unité venant de la théorie multiéchelle. Dans [78, 79, 21] la matrice correspond à un produit scalaire L_2 sur le bord du sous domaine. C'est la méthode que nous présentons et analysons dans le chapitre 3. Les deux familles d'espaces grossiers que l'on trouve dans [38, 39, 78, 79, 21] sont taillées pour le problème scalaire elliptique $(-\nabla \cdot \alpha \nabla u = f)$. La dernière vague de méthodes [99, 26, 105, 106, 70, 103, 109], utilise une nouvelle forme de problème aux valeurs propres généralisés ou le matrice est choisie au travers de l'analyse théorique de convergence. Ce choix est particulièrement intéressant car il s'applique à la plupart des systèmes d'équations aux dérivées partielles discrétisés par éléments finis et en particulier aux problèmes d'élasticité linéaire ou aux systèmes issus de la linéarisation d'un problème d'élasticité. Dans le chapitre 4 nous présentons notre contribution à cette famille de méthodes pour le préconditionneur de Schwarz puis pour les solveurs FETI et BDD dans le chapitre 5.

Mentionnons aussi que des extensions à plus de deux niveaux de ces techniques existent (voir [28, 27, 116, 100, 70, 103]).

1.3 Contributions de cette thèse

Le sujet de cette thèse a été défini en concertation avec l'entreprise Michelin à la suite des travaux [78]. En milieu industriel la robustesse d'un solveur figure parmi les propriétés les plus importantes : il faut être capable de garantir qu'une fois le calcul lancé il va effectivement converger. Pour cette raison la stratégie développée dans cette thèse prend place au niveau le plus algébrique possible : on ne fait presque aucune hypothèse sur le système linéaire symétrique défini positif que l'on résout. Ainsi on est paré pour faire face à un large champ de difficultés (en particulier la présence de discontinuités dans les coefficients).

Le chapitre 3 de cette thèse se concentre sur les problèmes scalaires du type $-\nabla \cdot (\alpha \nabla u) = f$. L'espace grossier, pour Schwarz, est construit à partir des modes à basse fréquence de l'opérateur Dirichlet-to-Neumann défini sur le bord de chaque sous domaine. Dans le chapitre 4 nous restons dans le cadre de Schwarz et nous proposons et analysons un espace grossier qui s'applique cette fois aux systèmes d'équations aux dérivées partielles. C'est celui-ci que nous appelons GenEO pour Generalized Eigenproblems in the Overlaps. Puis, le chapitre 5 propose d'appliquer la stratégie GenEO aux méthodes BDD et FETI. Si l'idée de départ est similaire la mise en œuvre est très différente. Dans chaque cas on démontre que la convergence ne dépend pas des difficultés spécifiques à chaque problème et on illustre cela par des résultats numériques. L'objectif du chapitre 6 est de tester nos méthodes sur des problèmes d'elasticité dans la limite quasi incompressible. Enfin, dans le chapitre 7 nous proposons quelques pistes et travaux en cours pour améliorer la méthode GenEO et un premier cas test industriel.

L'idée fondamentale sur laquelle est basée l'ensemble de ce travail est qu'au sein d'un solveur itératif on peut, grâce à des projections bien choisies, séparer le problème en deux parties : la première est résolue avec le solveur itératif et on réserve un traitement particulier à la seconde (une résolution avec un solveur direct). Dans la suite de ce manuscrit l'enjeu sera d'identifier quelle est la partie de l'espace solution sur laquelle le solveur itératif est efficace. Le complémentaire de cet espace, qui ralentit la convergence, servira d'espace grossier et par ce biais c'est à lui qu'on lui appliquera un solveur direct.

Nous cherchons des espaces grossiers qui sont
- engendrés par des vecteurs locaux (pour que la matrice du problème grossier soit creuse),
- calculés de manière automatique,
- en nombre raisonnable,
- tels que la méthode à deux niveaux soit robuste.

La stratégie pour le choix de l'espace grossier est toujours la même : grâce à la théorie de Schwarz abstraite on trouve quel est le point bloquant pour garantir la convergence de la méthode de décomposition de domaine. On définit ensuite un problème aux valeurs propres généralisés qui identifie quels vecteurs ont besoin d'être dans l'espace grossier. De cette manière on peut garantir la convergence théoriquement.

1.3.1 DtN : un espace grossier pour un problème scalaire

Nous présentons ici le chapitre 3 dans lequel nous nous concentrons sur un problème scalaire elliptique. En nous appuyant sur une idée d'espace grossier antérieure à cette thèse [78] nous présentons l'heuristique derrière sa construction ainsi que son analyse théorique qui prouve qu'elle est robuste et des résultats numériques. Il s'agit de la refonte des articles [79, 21, 52].

Étant donné un membre de droite f, le problème scalaire elliptique s'écrit : Trouver u^* tel que

$$-\nabla \cdot (\alpha \nabla u^*) = f,$$

où $\alpha : \Omega \to \mathbb{R}^+$ est un coefficient dont la valeur varie au sein du domaine. Puisqu'il s'agit pour le moment d'introduire des idées on passe sous silence les questions de conditions aux limites pour le problème global.

Heuristique Considérons le cas où un domaine est découpé en tranches (dans une seule direction). Dans la Figure 1.7 on présente cette géométrie en se concentrant sur trois des sous domaines. Appliquons l'algorithme de Schwarz alterné (1.2) à ce problème. Il est facile de vérifier que les mises à jour de l'erreur $e^n = |u^n - u^*|$ obéissent au même algorithme mais pour le problème homogène. En particulier chaque mise à jour de l'erreur e_2 dans le sous domaine Ω_2 vérifie (en utilisant les notations de la Figure 1.7) :

$$\begin{aligned} -\nabla \cdot (\alpha \nabla e_2^{n+1}) &= 0, \\ e_2^{n+1}(A, y) &= e_1^{n+1}(A, y), \\ e_2^{n+1}(D, y) &= e_3^{n}(D, y). \end{aligned} \tag{1.38}$$

La figure 1.8 présente les mises à jour successives de l'erreur dans Ω_2 et dans les sous domaines voisins. On connait le comportement général de ces mises à jour par le principe du maximum : e_2^{n+1} va décroitre à l'intérieur du sous domaine en partant des conditions aux limites en $x = A$ et $x = D$ données par les voisins. Comme le montre la figure, la convergence peut être rapide ou lente selon que les solutions des problèmes locaux décroissent rapidement ou non à l'intérieur de la zone de recouvrement (rappelons que puisqu'on regarde le comportement de l'erreur l'objectif est de la mener à zéro). C'est exactement sur ce constat que s'appuie la construction de l'espace grossier DtN : on souhaite isoler les composantes de la solution qui, étant donné une condition de Dirichlet, décroissent lentement à l'intérieur du sous domaine, et transmettent donc une condition aux limites peu améliorée à leurs voisins.

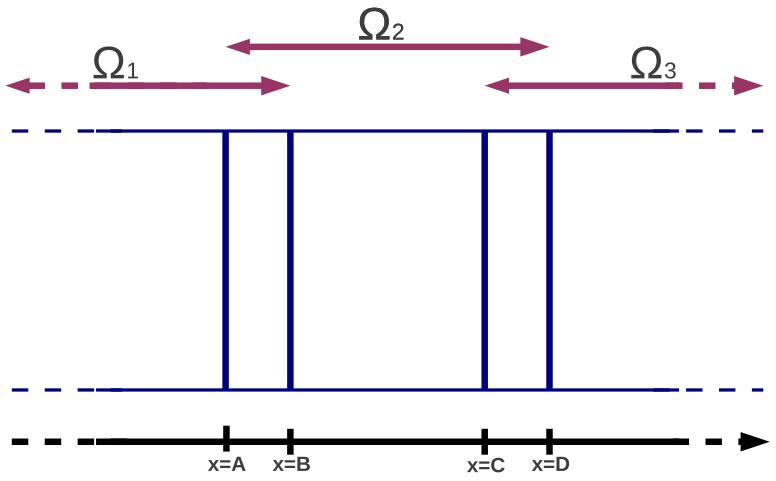


FIGURE 1.7 – Géométrie sur laquelle est basée l'heuristique derrière le choix de l'espace grossier DtN (voir aussi la Figure 1.8)

Sous l'hypothèse que la zone de recouvrement est étroite on peut estimer que si la dérivée normale de l'erreur au bord du sous domaine est grande alors on transmet au voisin une condition aux limites qui est plus proche de zéro.

Définition de l'espace grossier DtN L'opérateur Dirichlet-to-Neumann permet exactement d'évaluer cela. En effet, pour un domaine Ω_j il est défini ainsi.

Définition 1.14. Soit $\mathrm{tr}_j\alpha$ la trace du coefficient α sur le bord $\Gamma := \partial\Omega_j$ du sous domaine Ω_j en venant de l'intérieur et $\boldsymbol{n}_j$ la normale unitaire extérieure de Ω_j sur Γ. Pour toute fonction $v_\Gamma : \Gamma \to \mathbb{R}$ telle que $v_\Gamma|_{\partial\Omega} = 0$ si $\Gamma \cap \partial\Omega \neq \emptyset$ on définit

$$\mathrm{DtN}_j(v_\Gamma) \;:=\; \mathrm{tr}_j\alpha \frac{\partial v}{\partial \boldsymbol{n}_j}\bigg|_\Gamma \;, \text{ où } v \text{ est la solution de } \begin{cases} -\nabla\cdot(\alpha\nabla v) &= 0 & \text{in } \Omega_j \\ v &= v_\Gamma & \text{on } \Gamma \end{cases}. \tag{1.39}$$

En d'autres termes l'opérateur DtN prend une fonction définie sur Γ, calcule son extension harmonique à l'intérieur du sous domaine et retourne la dérivée normale de celle-ci sur le bord.

En s'appuyant dessus la procédure pour construire l'espace grossier est la suivante :

1. Calculer (en parallèle sur les sous domaines) les valeurs propres généralisées λ et les vecteurs propres généralisés v_Γ de

$$\mathrm{DtN}_j(v_\Gamma) \;=\; \lambda\,\mathrm{tr}_j\alpha\, v_\Gamma.$$

2. Sélectionner les vecteurs propres qui correspondent à une valeur propre plus petite que $1/\mathrm{diam}(\Omega_j)$ (l'inverse de la taille du sous domaine).

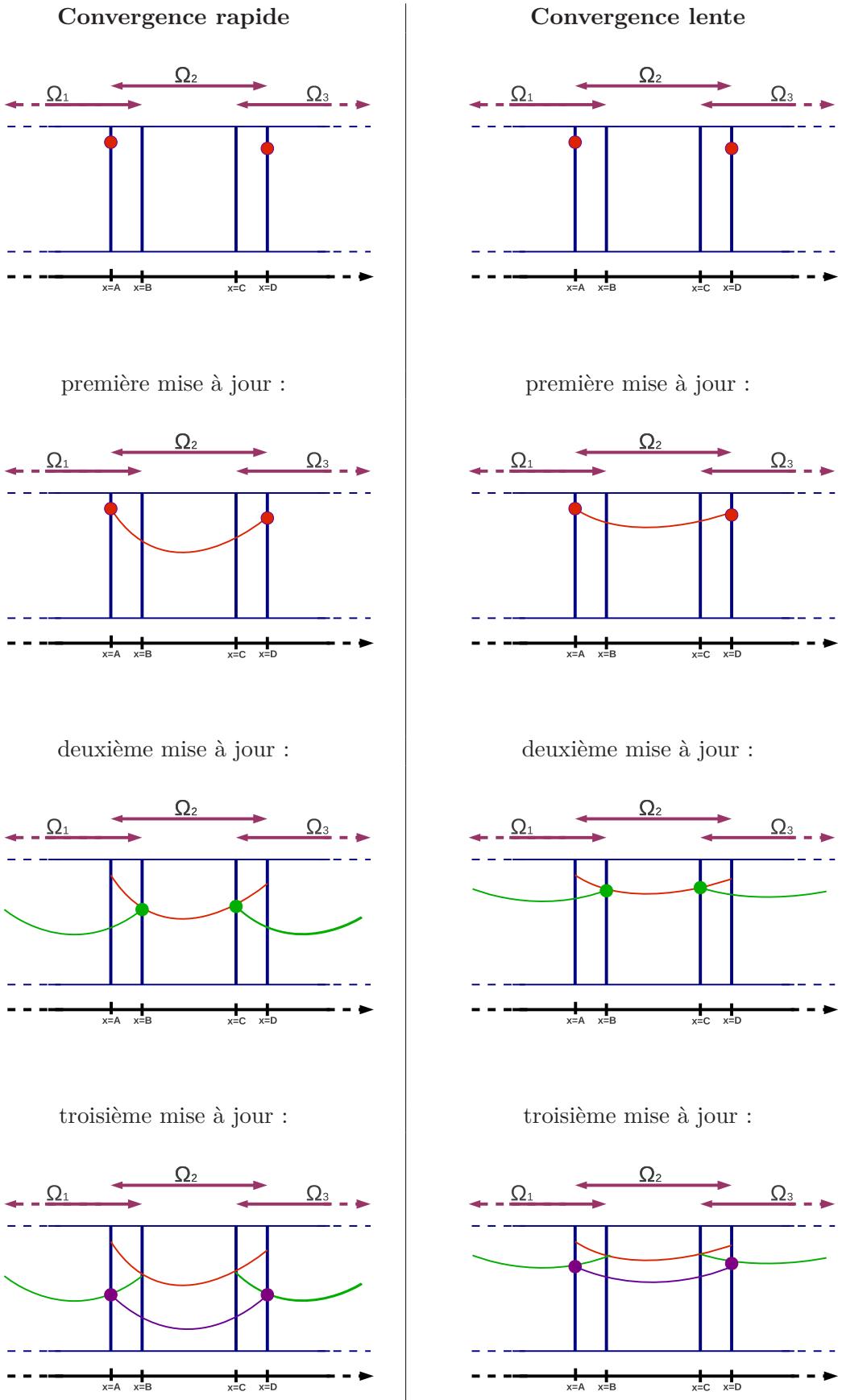


Figure 1.8 – Illustration de l'intérêt de l'opérateur DtN pour prédire la vitesse de convergence. (Puisqu'on regarde les mises à jour de l'erreur on a envie qu'elle décroisse le plus rapidement possible.) On constate que si le composante locale de l'erreur décroit rapidement dans le recouvrement on donne au sous domaine voisin une "bonne" valeur.

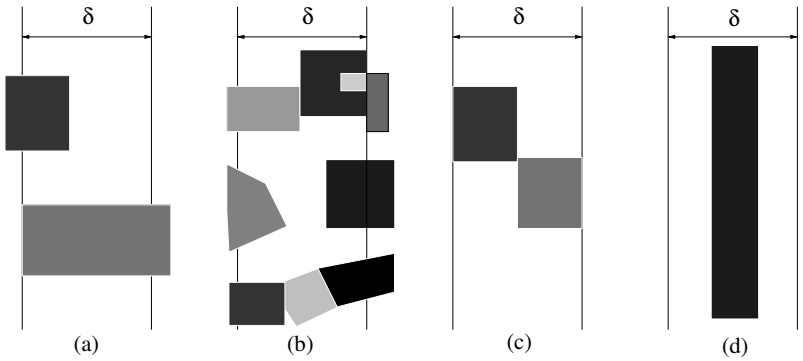


FIGURE 1.9 – Zone de recouvrement entre deux sous domaines (plus le niveau de gris est foncé plus la valeur de α correspondante est grande). Dans les deux premiers cas en partant de la gauche l'hypothèse sur les coefficients est vérifiée et $C_P = \mathcal{O}(1)$ dans le théorème (a & b), dans le troisième cas l'hypothèse est vérifiée et $C_P = \mathcal{O}(\log(\delta_j/h))$ dans le théorème (c), dans le dernier cas l'hypothèse n'est pas vérifiée (d).

3. Étendre ces vecteurs propres harmoniquement à l'intérieur du sous domaine.
4. Leur appliquer une partition de l'unité, interpoler dans l'espace éléments finis et prolonger par 0 à Ω tout entier.

Résultat théorique En construisant l'espace grossier de cette manière on peut garantir que la méthode à deux niveaux correspondantes convergera indépendamment de presque tous les paramètres du problème.

Théorème 1.15. *Sous une certaine hypothèse sur α le conditionnement de A préconditionné par Schwarz additif à deux niveaux avec l'espace grossier DtN satisfait*

$$\kappa(M_{AS,2}^{-1}A) \lesssim \left(C_P^2 + \max_{j=1}^{N} \frac{diam(\Omega_j)}{\delta_j}\right).$$

La constante qui est cachée par le symbole $\lesssim$ ne dépend ni de la taille du maillage h, ni de celle du recouvrement δ_j, ni de celle du sous domaine $\mathrm{diam}(\Omega_j)$*, ni du choix de α. Quelques détails sur C_P et sur l'hypothèse sous-jacente au théorème sont données dans la Figure 1.9.*

L'hypothèse sur les coefficients est requise pour pouvoir appliquer des inégalités de Poincaré pondérées [89]. Elle n'est pas très restrictive et est toujours vérifiée dans le cas où le recouvrement est minimal (la taille δ_j du recouvrement est égale à la taille d'une maille). Le cas typique où elle n'est pas vérifiée est celui où une zone où α est très élevé sert de séparation à deux zones de recouvrement où α prend une valeur plus faible. C'est ce qui est illustré dans la Figure 1.9. Le cas où l'inégalité ne s'applique pas correspond exactement au cas où, à cause de α, la valeur de la dérivée normale de l'erreur sur le bord du sous domaine n'est pas corrélée à la décroissance de l'erreur à l'intérieur du recouvrement et on ne peut donc pas garantir quelle sera la condition de raccordement transmise au sous domaine voisin.

D'un point de vue de la décomposition de domaine, comme toujours avec le préconditionneur de Schwarz additif et comme on l'a déjà remarqué (Remarque 1.10) l'essentiel de la preuve consiste à montrer qu'il existe une décomposition stable de n'importe quel vecteur sur les sous domaines locaux et l'espace grossier (hypothèse 1.6).

Nombre d'itérations :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	18	25	25	25
$\alpha_2 = 10^2$	21	26	27	26
$\alpha_2 = 10^4$	22	28	28	26
$\alpha_2 = 10^6$	17	25	25	25

Estimation du Conditionnement :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	22.4	25.3	26.1	26.2
$\alpha_2 = 10^2$	22.4	25.3	26.1	26.2
$\alpha_2 = 10^4$	22.4	25.3	26.1	26.2
$\alpha_2 = 10^6$	22.4	25.3	26.0	26.2

Table 1.3 – Résultats de convergence pour le **problème scalaire elliptique** (1.19) discrétisé par des éléments finis $\mathbb{P}_1$ avec le préconditionneur de Schwarz à deux niveaux (1.32) et V_0 l'espace grossier de **DtN**. La géométrie est présentée dans la Figure 1.5. Deux couches d'éléments sont ajoutées à chaque sous domaine. On fait varier le nombre de sous domaines et le paramètre α_2 dans le matériau 2. On présente ici le nombre d'itérations nécessaire pour converger (en haut) et l'estimation du conditionnement de la matrice préconditionné basée sur les valeurs de Ritz (en bas).

Résultat numérique A présent nous illustrons l'efficacité de l'espace grossier DtN. Le tableau 1.3 montre les résultats du test de robustesse que l'on a déjà fait passer au préconditionneur de Schwarz à un niveau et au préconditionneur à deux niveaux avec l'espace grossier basé sur une partition de l'unité. On observe une robustesse quasi parfaite. Le nombre de vecteurs qui est sélectionné pour l'espace grossier dans chaque sous domaine est égal au nombre de couches où α a une valeur élevée (donc trois dans notre cas) ce qui est la valeur optimale (voir [115, 47] ou bien la discussion dans la sous section 1.2.3). Dans le chapitre 3 on présente une série de tests plus complète.

1.3.2 GenEO : un espace grossier pour Schwarz

Dans le chapitre 4 nous construisons un espace grossier qui permet de garantir la robustesse dans le cas beaucoup plus général des matrices symétriques définies positives. Ces travaux ont fait l'objet des publications [105, 106].

Idées Nous expliquons ici les idées qui ont conduit à la construction de l'espace GenEO et nous renvoyons au chapitre 4 pour la présentation rigoureuse. La preuve de convergence pour l'espace grossier DtN repose sur deux arguments que l'on ne peut pas généraliser facilement à un système quelconque :

- des inégalités de Poincaré pondérées permettent d'obtenir une relation entre une norme sur le bord du sous domaine (qui est bornée par le problème aux valeurs propres) et une norme sur le recouvrement (dont on a besoin dans la preuve),
- une propriété de stabilité de l'interpolant qui envoie dans l'espace éléments finis (dans la norme euclidienne et la norme induite par A) intervient lorsque l'on applique la partition de l'unité à un vecteur propre car il faut alors interpoler le produit $\chi_j u_j$ dans l'espace des éléments finis.

Le problème aux valeurs propres que nous avons élaboré intègre ces deux difficultés dans sa définition même : l'une des formes bilinéaires dans le problème aux valeurs propre

généralisé est définie sur la zone de recouvrement entre les sous domaines et elle est pondérée par une partition de l'unité. De cette manière l'estimation qui résulte des propriétés du spectre du problème aux valeurs propres généralisé GenEO *via* le lemme 1.13 permet de contourner les arguments de la preuve du chapitre précédent que l'on ne sait pas démontrer dans le cas général.

Définition de l'espace grossier GenEO La définition de cet espace grossier est aussi présentée dans un poster en annexe de cette thèse. Ce poster a été présenté à la conférence *Special Semester on Multiscale Simulation and Analysis in Energy and the Environment* au RICAM à Linz (Autriche) en novembre 2011. Nous avons appelé l'espace grossier GenEO pour "Generalized Eigenvalues in the Overlaps". Ce mot apparait aussi dans le terme HeteroGenEOus ce qui est une coïncidence amusante. Afin de donner sa définition nous devons introduire quelques éléments de notation qui sont définis avec précision dans le corps du chapitre :

- pour tout domaine $D \subset \Omega$ qui est compatible avec le maillage, $V_h(D)$ est l'ensemble des restrictions à D des fonctions éléments finis,
- pour tout domaine $D \subset \Omega$ qui est compatible avec le maillage, $V_{h,0}(D)$ est l'ensemble des restrictions à D des fonctions éléments finis qui ont leur support dans $\bar{D}$,
- pour tout domaine $D \subset \Omega$ qui est compatible avec le maillage, la forme bilinéaire $a_D : D \times D \to \mathbb{R}^+$ est obtenue en assemblant seulement les matrices élémentaires des éléments qui composent D,
- Ω_j° est la partie de Ω_j qui est recouverte par au moins un autre sous domaine,
- pour $j = 1, \ldots, N$; $\Xi_j : V_h(\Omega_j) \to V_{h,0}(\Omega_j)$ est une famille de fonctions qui constituent une partition de l'unité subordonnée à la décomposition en sous domaines. Ces fonctions sont à valeurs directement dans un espace élément fini donc il n'y a pas besoin d'interpoler après les avoir appliquées.

Définition 1.16 (Espace grossier GenEO)**.** Pour chaque sous domaine $j = 1, \ldots, N$, on résout le problème aux valeurs propres généralisé suivant : trouver (λ, p) tel que

$$a_{\Omega_j}(p, v) = \lambda\, a_{\Omega_j^\circ}(\Xi_j(p), \Xi_j(v)), \quad \forall\, v \in V_h(\Omega_j). \tag{1.40}$$

Pour chaque $j = 1, \ldots, N$, soient $(p_j^k)_{k=1}^{m_j}$ les vecteurs propres associés au problème aux valeurs propre généralisé (1.40) et qui ont les m_j plus petites valeurs propres. On définit l'espace grossier ainsi

$$V_H := \mathrm{Vect}\{R_j^\top \Xi_j(p_j^k) : k = 1, \ldots, m_j;\ j = 1, \ldots, N\}.$$

Résultat théorique Avec cet espace grossier la méthode de Schwarz à deux niveaux converge indépendamment du nombre de sous domaines et des paramètres du problème comme l'indique le théorème suivant

Théorème 1.17. *Le conditionnement de la matrice A préconditionnée par Schwarz à deux niveaux avec l'espace grossier GenEO est borné par*

$$\kappa(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \ \leq\ (1 + k_0)\left[2 + k_0(2k_0 + 1)\max_{1 \leq j \leq N}\left(1 + \frac{1}{\lambda_j^{m_j+1}}\right)\right],$$

où la constante k_0 dépend de la géométrie du problème (mais pas du nombre de sous domaines) et est définie dans le lemme 1.8.

Nombre d'itérations :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	19	24	25	24
$\alpha_2 = 10^2$	23	26	27	26
$\alpha_2 = 10^4$	26	26	27	27
$\alpha_2 = 10^6$	17	21	22	25

Estimation du Conditionnement :

	8 sous domaines	16 sous domaines	32 sous domaines	64 sous domaines
$\alpha_2 = 1$	31.8	31.9	31.9	31.9
$\alpha_2 = 10^2$	31.8	31.9	31.9	31.9
$\alpha_2 = 10^4$	31.8	31.9	31.9	31.9
$\alpha_2 = 10^6$	31.8	31.9	31.9	31.9

Table 1.4 – Résultats de convergence pour le **problème scalaire elliptique** (1.19) discrétisé par des éléments finis $\mathbb{P}_1$ avec le préconditionneur de Schwarz à deux niveaux (1.32) et V_0 est l'espace grossier **GenEO** . La géométrie est présentée dans la Figure 1.5. Deux couches d'éléments sont ajoutées à chaque sous domaine. On fait varier le nombre de sous domaines et le paramètre α_2 dans le matériau 2. On présente ici le nombre d'itérations nécessaire pour converger (en haut) et l'estimation du conditionnement de la matrice préconditionné basée sur les valeurs de Ritz (en bas).

On renvoie au corps du chapitre pour les hypothèse exactes sous lesquelles ce résultat s'applique. Elles sont très peu restrictives. On remarque qu'apparait dans le résultat $\lambda_j^{m_j+1}$ qui, pour chaque sous domaine est la plus petite valeur propre qui n'a pas été sélectionnée par l'espace grossier. Une possibilité est d'utiliser le test $\lambda_j^k < \delta_j/\mathrm{diam}(\Omega_j)$ pour décider quels vecteurs on met dans l'espace grossier. C'est alors cette quantité qui apparait dans le théorème comme dans le résultat du chapitre précédent :

$$\kappa(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \ \leq \ (1+k_0)\left[2+k_0(2k_0+1)\max_{1\leq j\leq N}\left(1+\frac{\mathrm{diam}(\Omega_j)}{\delta_j}\right)\right],$$

.

Résultat Numérique Avec ce choix nous soumettons GenEO au test de robustesse que nous avons déjà effectué avec le préconditionneur à un niveau, et le préconditionneur à deux niveaux pour l'espace grossier basé sur une partition de l'unité et l'espace grossier DtN. Comme pour DtN on sélectionne autant de modes par sous domaine qu'il y a de couche ou α est élevé (trois) et grâce à cela le solveur est robuste comme le montrent les résultats du tableau 1.4.

1.3.3 Espace grossier GenEO pour FETI et BDD

Nous résumons ici les contribution du chapitre 5. Elles ont fait l'objet de la publication [109] ainsi que de la note [108].

Une nouvelle fois nous exploitons au maximum le formalisme de Schwarz pour identifier quelle est l'estimation qui est difficile à démontrer. La différence majeure avec la méthode GenEO pour Schwarz est que cette fois c'est l'hypothèse de stabilité des solveurs locaux (Hypothèse 1.5) qui imposera le choix du problème aux valeurs propres alors que l'existence d'une décomposition stable (Hypothèse 1.6) est triviale sur tout l'espace grâce à la présence d'opérateurs de partition de l'unité dans les préconditionneurs.

La majeure partie de l'étude théorique consiste à reformuler l'hypothèse 1.5 pour trouver le bon problème aux valeurs propres. Une fois trouvé ce problème et étant donné un critère τ on construit l'espace grossier en sélectionnant tous les vecteurs propres associés à une valeur propre plus petite que τ. Grâce à cela on démontre que les conditionnement des opérateurs BDD et FETI préconditionnés sont bornés par $\frac{\mathcal{N}}{\tau}$ où $\mathcal{N}$ mesure le nombre maximal de voisins qu'a un sous domaine.

Si pour BDD le problème aux valeurs propres apparait naturellement grâce à la formulation de BDD dans le formalisme de Schwarz, pour FETI la procédure a été plus complexe puisque c'est sur la transposée FM^{-1} de l'opérateur préconditionné, qui a le même spectre que $M^{-1}F$, que nous avons travaillé. Grâce à cela le résultat s'écrit non seulement pour le préconditionneur de Dirichlet que nous avons déjà présenté mais aussi pour le préconditionneur appelé Lumped qui est moins coûteux à appliquer.

Des résultats numériques illustrent le comportement de la méthode pour le cas de FETI.

1.3.4 Application à l'élasticité quasi-incompressible

Les résultats obtenus pour GenEO avec la méthode de Schwarz sont très satisfaisant pour de nombreux problèmes. Malheureusement il subsistait un obstacle : le préconditionneur de Schwarz est tellement mal adapté aux problèmes d'élasticité quasi incompressible qu'avec notre processus de sélection automatique on construit un espace grossier constitué de tous les champs de déplacements dans le recouvrement. Puisque nos méthodes doivent s'appliquer aux calculs qu'effectuent Michelin le cas de l'élasticité quasi incompressible est incontournable : le caoutchouc avec lequel on conçoit des pneus est un matériau quasi incompressible.

Dans le chapitre 6 nous montrerons numériquement que contrairement au cas de Schwarz, FETI GenEO convient parfaitement aux problèmes quasi incompressibles. Ici nous illustrons comment nous avons eu l'intuition qu'il était nécessaire de changer de solveur.

Analyse en Fourier de la méthode de Schwarz On considère le cas où $\Omega = \mathbb{R}^2$ et le domaine est partitionné en deux demi plans $\Omega_1 = \{(x, y);\, x < \delta\}$ et $\Omega_2 = \{(x, y);\, x > 0\}$. L'épaisseur du recouvrement est $\delta\,(> 0)$.

En deux dimensions l'équation de l'élasticité linéaire à coefficients constants s'écrit sous forme développée pour l'inconnue vectorielle $\boldsymbol{u} = (u, v)^{\mathrm{T}}$ et le membre de droite $\mathbf{f} = (f_1, f_2)^{\top}$ comme suit

$$\begin{cases} -\mu\Delta u - (\lambda+\mu)\partial_x \nabla\cdot(\boldsymbol{u}) = f_1, \\ -\mu\Delta v - (\lambda+\mu)\partial_y \nabla\cdot(\boldsymbol{u}) = f_2, \end{cases} \tag{1.41}$$

où λ et μ s'écrivent en fonction des coefficients de Lamé

$$\lambda := \frac{E\nu}{(1+\nu)(1-2\nu)},\ \mu := \frac{E}{2(1+\nu)}.$$

Étant donné la géométrie particulière du domaine on peut lui appliquer une transformée de Fourier dans la direction y, ce qui permet d'écrire le problème sous la forme suivante :

$$\begin{cases} -(2\mu+\lambda)\partial_{xx}\hat{u} + k^2\mu\hat{u} - ik(\lambda+\mu)\partial_x\hat{v} = \hat{f}_1, \\ k^2(2\mu+\lambda)\hat{v} - \mu\partial_{xx}\hat{v} - ik(\lambda+\mu)\partial_x\hat{u} = \hat{f}_2. \end{cases} \tag{1.42}$$

A $k > 0$ fixé, ceci est une équation différentielle ordinaire dont la solution (obtenue avec Maple) est

$$\left\{\begin{array}{rcl} \hat{u}\,(x) = & a_1\,\mathrm{e}^{-kx} + a_2\,\mathrm{e}^{-kx}x + a_3\,\mathrm{e}^{kx} + a_4\,\mathrm{e}^{kx}x, \\ \hat{v}\,(x) = & \frac{-i\left(\mu\,a_1\,k\mathrm{e}^{-kx}+\mu\,a_2\,k\mathrm{e}^{-kx}x-3\,\mu\,a_2\,\mathrm{e}^{-kx}-\mu\,a_3\,k\mathrm{e}^{kx}-\mu\,a_4\,k\mathrm{e}^{kx}x-3\,\mu\,a_4\,\mathrm{e}^{kx}\right)}{k(\mu+\lambda)} \\ + & \frac{-i\left(k\lambda\,a_1\,\mathrm{e}^{-kx}+k\lambda\,a_2\,\mathrm{e}^{-kx}x-\lambda\,a_2\,\mathrm{e}^{-kx}-k\lambda\,a_3\,\mathrm{e}^{kx}-k\lambda\,a_4\,\mathrm{e}^{kx}x-\lambda\,\mathrm{e}^{kx}a_4\right)}{k(\mu+\lambda)}, \end{array}\right. \tag{1.43}$$

a_1, a_2, a_3 et a_4 étant des constantes d'intégrations complexes.

Notons $(\hat{u}_1, \hat{v}_1)$ la solution de ce problème restreinte au sous domaine Ω_1 ($x < \delta$) et $(\hat{u}_2, \hat{v}_2)$ la solution de ce problème restreinte au sous domaine Ω_2 ($x > 0$). Par un argument classique on sait que $\hat{u}_1$ et $\hat{v}_1$ doivent être bornées en $-\infty$ et que $\hat{u}_2$ et $\hat{v}_2$ doivent être bornées en $+\infty$ ce qui nous permet d'éliminer la moitié des termes dans (1.43) :

$$\left\{\begin{array}{l} \hat{u}_1 = (a_1 + b_1 x)\mathrm{e}^{kx}, \\ \hat{v}_1 = \frac{i\,(a_1\,\mu k + b_1\,\mu k x + 3b_1\,\mu + a_1\,\lambda k + b_1\,\lambda k x + b_1\,\lambda)\mathrm{e}^{kx}}{k(\mu+\lambda)}, \\ \hat{u}_2 = (a_2 + b_2 x)\mathrm{e}^{kx}, \\ \hat{v}_2 = \frac{i\,(a_2\,\mu k + b_2\,\mu k x + 3b_2\,\mu + a_2\,\lambda k + b_2\,\lambda k x + b_2\,\lambda)\mathrm{e}^{kx}}{k(\mu+\lambda)}. \end{array}\right. \tag{1.44}$$

Comparaison entre quatre types de conditions de transmission Afin de trouver la solution du problème global nous utilisons l'algorithme de Schwarz alterné (1.2) : le problème est résolu tour à tour dans chaque sous domaine en utilisant la solution fournie par le voisin pour faire office de condition aux limites. En fait, nous généralisons l'algorithme classique (1.2) en considérant quatre type de conditions de transmission différentes : (A) continuité des déplacements (normal et tangentiel), (B) continuité des contraintes (normale et tangentielle), (C) continuité de la contrainte normale et du déplacement tangentiel, (D) continuité de la contrainte tangentielle et du déplacement normal.

$$\left\{\text{(A)}\ \begin{array}{c} u_1^{n+1}(\delta, y) = u_2^n(\delta, y), \\ v_1^{n+1}(\delta, y) = v_2^n(\delta, y), \\ u_2^{n+1}(0, y) = u_1^n(0, y), \\ v_2^{n+1}(0, y) = v_1^n(0, y). \end{array}\right. \quad \left\{\text{(B)}\ \begin{array}{c} {\sigma_1}_n^{n+1}(\delta, y) = {\sigma_2}_n^n(\delta, y), \\ {\sigma_1}_t^{n+1}(\delta, y) = {\sigma_2}_t^n(\delta, y), \\ {\sigma_2}_n^{n+1}(0, y) = {\sigma_1}_n^n(0, y), \\ {\sigma_2}_t^{n+1}(0, y) = {\sigma_1}_t^n(0, y). \end{array}\right.$$

$$\left\{\text{(C)}\ \begin{array}{c} v_1^{n+1}(\delta, y) = v_2^n(\delta, y), \\ {\sigma_1}_n^{n+1}(\delta, y) = {\sigma_2}_n^n(\delta, y), \\ v_2^{n+1}(0, y) = v_1^n(0, y), \\ {\sigma_2}_n^{n+1}(0, y) = {\sigma_1}_n^n(0, y). \end{array}\right. \quad \left\{\text{(D)}\ \begin{array}{c} u_1^{n+1}(\delta, y) = u_2^n(\delta, y), \\ {\sigma_1}_t^{n+1}(\delta, y) = {\sigma_2}_t^n(\delta, y), \\ u_2^{n+1}(0, y) = u_1^n(0, y), \\ {\sigma_2}_t^{n+1}(0, y) = {\sigma_1}_t^n(0, y). \end{array}\right.$$

où :

$$\sigma_n = (2\mu + \lambda)\frac{\partial u}{\partial x} + \lambda\frac{\partial v}{\partial y},\ \text{et}\ \sigma_t = \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right),$$

sont les composantes normales et tangentielles des forces à l'interface.

Après transformée de Fourier selon y les forces à l'interface s'écrivent

$$\hat{\sigma_n} = (2\mu + \lambda)\frac{\partial \hat{u}}{\partial x} + ik\lambda\hat{v},\ \text{et}\ \hat{\sigma_t} = \mu\left(\frac{\partial \hat{v}}{\partial x} + ik\hat{u}\right).$$

Par le même argument que dans l'analyse heuristique pour DtN, les mises à jour de l'erreur obéissent au même schéma itératif mais pour le problème homogène. Dans la suite un *bon* schéma est donc un schéma qui permet de converger vers zéro rapidement.

Grâce à des substitutions de variable astucieuses et à l'utilisation de Maple on peut trouver pour les coefficients a_1 et b_1 dans les expressions de nos inconnues une matrice d'itération qui lie leurs valeurs à l'itération $n-1$ à leurs valeurs à l'itération $n+1$:

$$\begin{pmatrix} a_1^{n+1} \\ b_1^{n+1} \end{pmatrix} = M \begin{pmatrix} a_1^{n-1} \\ b_1^{n-1} \end{pmatrix}. \tag{1.45}$$

Puisque Ω_1 et Ω_2 jouent des rôles symétriques les coefficients dans Ω_2 satisfont aussi à cette équation. Bien sûr M dépend du choix des conditions de transmission. Dans le cas où les conditions de transmissions sont mixtes ((C) et (D)) la matrice d'itération prend une forme très simple :

$$M_{C \text{ ou } D} = \begin{pmatrix} \mathrm{e}^{-2\,k\delta} & -2\,\delta\,\mathrm{e}^{-2\,k\delta} \\ 0 & \mathrm{e}^{-2\,k\delta} \end{pmatrix}$$

Dans les deux autres cas les matrices prennent des formes très compliquées et pour cette raison on s'intéresse désormais à leurs deux valeurs propres eig_1 et eig_2 indexées par A, B, C ou D en gardant à l'esprit que la convergence est bonne lorsque les valeurs propres sont petite et qu'elle est mauvaise lorsqu'elles approchent 1. On trouve

$$\begin{cases} eig_{A_1} &= \left[1 + 2\frac{(\delta k)^2}{(3-4\nu)^2} + 2\sqrt{\frac{(\delta k)^2}{(3-4\nu)^2} + \frac{(\delta k)^4}{(3-4\nu)^4}}\right] \mathrm{e}^{-2\,k\delta}, \\ eig_{A_2} &= \left[1 + 2\frac{(\delta k)^2}{(3-4\nu)^2} - 2\sqrt{\frac{(\delta k)^2}{(3-4\nu)^2} + \frac{(\delta k)^4}{(3-4\nu)^4}}\right] \mathrm{e}^{-2\,k\delta}. \end{cases} \tag{1.46}$$

$$\begin{cases} eig_{B_1} = \left(1 + 2\,\delta^2 k^2 + 2\,\sqrt{\delta^2 k^2 + \delta^4 k^4}\right) \mathrm{e}^{-2\,k\delta}, \\ eig_{B_2} = \left(1 + 2\,\delta^2 k^2 - 2\,\sqrt{\delta^2 k^2 + \delta^4 k^4}\right) \mathrm{e}^{-2\,k\delta}. \end{cases} \tag{1.47}$$

$$\left\{ eig_{C_1} = eig_{C_2} = eig_{D_1} = eig_{D_2} = \mathrm{e}^{-2\,k\delta} := eig_C := eig_D. \right. \tag{1.48}$$

On peut d'ores et déjà faire les remarques suivantes :

–

$$eig_{A_1} > eig_C = eig_D > eig_{A_2}, \tag{1.49}$$

et

$$eig_{B_1} > eig_C = eig_D > eig_{B_2}. \tag{1.50}$$

– $eig_{B_1}, eig_{B_2}, eig_C = eig_D$ ne dépendent pas des paramètres physiques λ et μ. Ils dépendent par contre de la taille du recouvrement δ et de la fréquence k.
– Dans l'ensemble des quatre cas, s'il n'y a pas de recouvrement l'algorithme ne converge pas et réciproquement si la taille du recouvrement est non nulle alors les valeurs propres sont < 1 et la convergence est garantie.
– La convergence est la plus mauvaise (les valeurs propres sont proches de 1) pour les basses fréquences. C'est ce qu'on attend d'une méthode de décomposition de domaine primale.

Étude de la limite incompressible Les paramètres du matériau n'ont d'influence que dans le cas (A) où les conditions de transmission sont des conditions purement sur le déplacement. Dans ce cas la limite incompressible est

$$\begin{cases} \lim\limits_{\nu \to 0.5} eig_{A_1} &= \left[1 + 2(\delta k)^2 + 2\sqrt{(\delta k)^2 + (\delta k)^4}\right] \mathrm{e}^{-2\,k\delta}, \\ \lim\limits_{\nu \to 0.5} eig_{A_2} &= \left[1 + 2(\delta k)^2 - 2\sqrt{(\delta k)^2 + (\delta k)^4}\right] \mathrm{e}^{-2\,k\delta}. \end{cases} \tag{1.51}$$

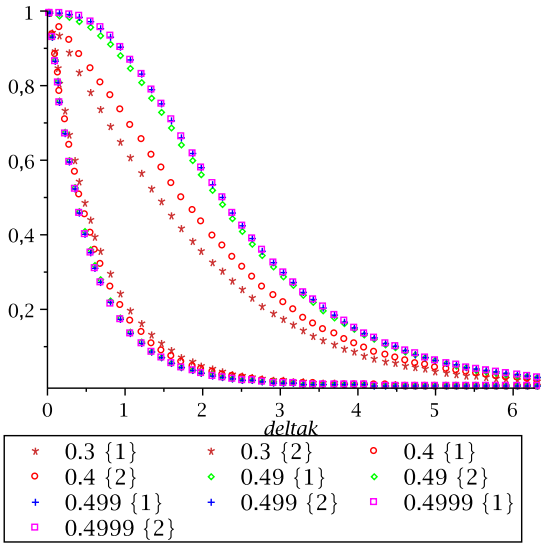


Figure 1.10 – Cas (A) : valeurs propres en fonction de δk pour différentes valeurs du coefficient de Poisson ν.

On observe que $\lim_{\nu \to 0.5} eig_{A1} = eig_{B1}$ et $\lim_{\nu \to 0.5} eig_{A2} = eig_{B2}$. Pour cette raison le fait que dans le cas (B) les valeurs propres ne dépendent pas des coefficients de Lamé n'est pas un avantage : avec les conditions de transmission en contrainte pure la convergence est toujours pire qu'avec les conditions de transmission en déplacement pur. Dans la figure 1.10 on trace les deux valeurs propres pour le cas (A) en fonction de δk et pour différentes valeurs du coefficient de Poisson ν. On observe effectivement un phénomène de convergence quand ν approche 0.5. Enfin, dans la Figure 1.11 on trace la plus grande valeur propre (et donc la plus mauvaise) en fonction du coefficient de Poisson ν pour différentes valeurs de δk. On se concentre sur les valeurs de $\delta k \leq 1$ car le problème qu'on résout est en fait un problème discrétisé pour lequel les fonctions ne peuvent pas "osciller" plus vite que le pas du maillage ce qui restreint le champ des fréquences à $k \leq \frac{1}{h}$. Si de plus le recouvrement est minimal ($\delta = h$) on a bien $\delta k \leq 1$.

Les conclusions que nous tirons de cette étude est que les conditions de Dirichlet qui sont celles mises en œuvre dans l'algorithme de Schwarz ne sont pas adaptées au problème d'élasticité quasi incompressible. L'étude en Fourier suggère que l'utilisation de conditions mixtes (une composante en déplacement, une composante en contrainte) conduirait à de bonnes performances même dans la limite quasi incompressible. Ce résultat est déjà connu [44, 45, 20, 82]. Nous n'avons pas poursuivi dans cette voie car il semblait difficile de tirer parti de ce type de conditions aux limites tout en raisonnant de manière algébrique. Pour cette raison nous nous sommes concentrés sur les formulations en sous-structuration (sans recouvrement) que sont FETI et BDD.

Nous tirons de [76] un argument supplémentaire en faveur d'une formulation en sous-structuration. Les résultats de [14] montrent que pour la géométrie que nous avons considérée ($\mathbb{R}^2$ divisé en deux demi plans), si on applique la formulation en sous-structuration pour itérer au sein d'une méthode de Richardson alors on obtient un solveur exact pour le problème de Poisson [14] : on trouve la solution en une itération. Dans le cas de l'équation de Stokes qui est très étroitement liée à une formulation mixte de l'élasticité quasi-incompressible les auteurs de [76] proposent un solveur exact et montrent aussi que la formulation BDD classique conduit à un algorithme qui converge indépendamment des

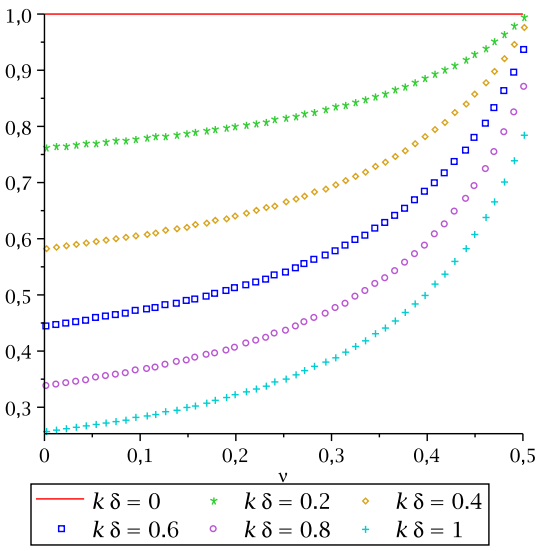


FIGURE 1.11 – Cas (A) : plus grande valeur propre en fonction du coefficient de Poisson ν pour différentes valeurs de δk.

paramètres de Lamé.

Dans le chapitre 6 nous évoquons certains problèmes liés à la discrétisation dans la limite quasi incompressible mais surtout nous illustrons le comportement de nos problèmes aux valeurs propres et en particulier le fait qu'avec FETI-GenEO un seul mode grossier par sous domaine suffit à gérer le comportement incompressible et notre objectif est donc atteint.

1.3.5 Perspectives

Pour terminer ce manuscrit nous présentons dans le chapitre 7 quelques pistes d'amélioration ou d'exploration de nos algorithmes. Il s'agit de trois directions de recherche en cours d'investigation et nous ne considérons pas ces travaux comme étant finis. Nous présentons d'abord comment, grâce à la formulation abstraite que nous avons employée, il est possible d'étendre l'idée des espaces grossiers GenEO à un algorithme multiniveaux. C'est un atout important dans les cas où le problème est très difficile et où la grille grossière devient très grande. Puis nous exposons une manière alternative de construire l'espace grossier où la sélection des vecteurs de base ne se fait plus *a priori* en résolvant des problèmes aux valeurs propres mais à la volée au sein même des itérations du gradient conjugué. Nous appelons cette méthode Frugal FETI car il s'agit d'être économe avec les moyens de calcul. Enfin, nous montrons un premier résultat obtenu sur un cas test de pneu avec la version de Frugal FETI qui a été implémentée au sein des codes de calcul Michelin au cours de cette thèse.

Chapter 2

Introduction

The same introduction is given in French in the previous chapter.

Contents

When faced with the problem of solving a large linear system on a parallel architecture two families of solvers are available with optimized black box implementations: direct solvers and iterative solvers. Direct solvers are robust in the sense that it is guaranteed that they will find the solution in a given number of operations no matter how hard the problem. Their memory requirements however are such that they can become unreliable when the problem becomes too large. On the other hand iterative solvers are naturally parallel since they mostly use matrix vector products. The drawback is that they often lack robustness: for ill conditioned problems the use of a preconditioner becomes essential in order for convergence to be achieved and choosing the right preconditioner is an art in itself.

Domain decomposition methods can be viewed as hybrid methods: they solve the problem with an iterative solver within which local direct solvers on some subproblems are used to reformulate the original problem or to define the preconditioner (or both). The rationale is to get the advantages out of both families of methods: robustness and parallelizability. In the next section we present three of the most popular domain decomposition methods and illustrate the possible lack of robustness when confronted to

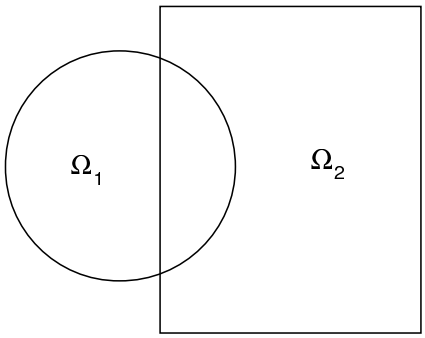


Figure 2.1: The domain Ω consists of a rectangle and a disk with an overlapping zone.

particularly hard problems. Then we will describe what a two level method is and how it may help with robustness. In the final section we describe the main contributions of this thesis. The unifying motivation behind this work is to design domain decomposition methods which are proved to converge even for very ill conditioned problems and which can be implemented as black box algorithms without any prior knowledge of the problem underlying the linear system being solved.

2.1 Domain Decomposition

We introduce two families of Domain Decomposition methods which we will work to improve in subsequent chapters. The first is the family of Schwarz methods. Their main advantage is that they are algebraic methods: they can be applied without any knowledge of the problem other than its formulation $A\mathbf{x} = \mathbf{f}$. The second family of methods consists of the substructuring methods. They are more sophisticated since they require access to the element matrices to assemble matrices of some local subproblems and are the solver of choice for many industrial applications.

2.1.1 Algebraic Domain Decomposition: the Schwarz method

A detailed historical approach is given in [41] with complete bibliographical references. The Schwarz domain decomposition methods are named after H. A. Schwarz who in 1870 [101] proposed the alternating Schwarz method in order to study the existence of a solution to the homogeneous Poisson problem with prescribed boundary conditions (2.1):

$$\left\{ \begin{array}{rcl} -\Delta u & = & 0 \text{ in } \Omega, \\ u & = & g \text{ on } \partial\Omega, \end{array} \right. \tag{2.1}$$

where $\Omega = \Omega_1 \cup \Omega_2$ is as in Figure 2.1. Given the existence of a solution in domains with simple geometries (disks, rectangles...) Schwarz's idea is to prove the existence of a solution on the more complex domain Ω by a constructive argument: he proposes to solve the problem alternately on each of the regular subdomains and to use transmission conditions coming from the solution just computed by the neighbour. More precisely Schwarz proves that the *Alternating Schwarz* algorithm initialized with u_2^0 and updated with:

$$\begin{array}{ll} -\Delta u_1^{n+1} = 0 \text{ in } \Omega_1 & -\Delta u_2^{n+1} = 0 \text{ in } \Omega_2 \\ u_1^{n+1} = g \text{ on } \partial\Omega_1 \cap \partial\Omega & u_2^{n+1} = g \text{ on } \partial\Omega_2 \cap \partial\Omega \\ u_1^{n+1} = u_2^n \text{ in } \Omega \setminus \Omega_1, & u_2^{n+1} = u_1^{n+1} \text{ in } \Omega \setminus \Omega_2, \end{array} \tag{2.2}$$

converges toward the solution of (2.1) and thus that the solution exists.

From a tool in functional analysis the alternating Schwarz method evolved into a solver for possibly complex domains [3]. The most immediate adaptation is what is now called multiplicative Schwarz [11, 10, 102]. The drawback of alternating Schwarz and of its discrete counterparts is that it is an inherently sequential approach to the problem. The parallel adaptation is the Additive Schwarz preconditioner which Matsokin and Nepomnyaschikh [72] significantly contributed to. We present it next based on the description given in [112].

If the finite element discretization of (2.1) reads $A\mathbf{u} = \mathbf{b}$, and $R_1^\top$ (respectively $R_2^\top$) is the (boolean) interpolation operator which prolongates a finite element function defined in Ω_1 (respectively Ω_2) to the whole of Ω by zero then we may define the local operators $A_1 := R_1 A R_1^\top$, $A_2 := R_2 A R_2^\top$ and the Additive Schwarz preconditioner

$$M^{-1} := R_1^\top A_1^{-1} R_1 + R_2^\top A_2^{-1} R_2. \tag{2.3}$$

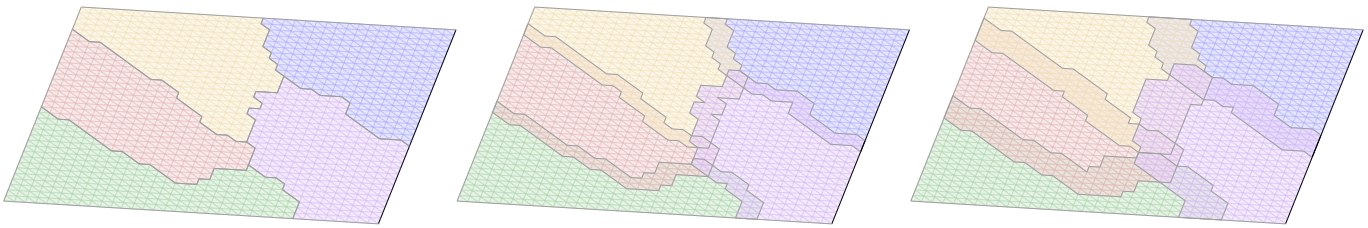
It is quite intuitive that this is a good preconditioner for A. Indeed it approximates the inverse of A by the sum of the inverses on each of the two subdomains. This generalizes easily to the case of more than two subdomains, and to general symmetric positive definite matrices A. All that is needed is a set of subspaces V_i of the space V of degrees of freedom and interpolation operators $R_i^\top : V_i \to V$ which satisfy

$$V = \sum_{i=1}^{N} R_i^\top V_i.$$

Then the additive Schwarz preconditioner is the sum of N local inverses

$$M^{-1} := \sum_{i=1}^{N} R_i^\top A_i^{-1} R_i, \text{ where } A_i := R_i A R_i^\top. \tag{2.4}$$

A common way to use the Additive Schwarz preconditioner is to start by splitting all the degrees of freedom in V into a non overlapping partition and then adding l layers of overlap to each subdomain as illustrated in Figure 2.2. In this case the interpolation operators $R_i^\top$ are boolean and the local matrices A_i are just extractions of the coefficients in A which correspond to the local degrees of freedom. The original splitting can be obtained either by hand based on the geometry of the underlying problem (if it is known) or with an automatic graph partitioner such as Metis [54] or Scotch [13] based on the graph of A.



(a) $l = 0$ (b) $l = 1$ (c) $l = 2$

Figure 2.2: Partition of $\Omega = [0; 1]^2$ into $N = 5$ subdomains with different values for the overlap parameter

As we will see later on the fact that the subdomains overlap is necessary in order to observe convergence with the Additive Schwarz preconditioner. There are two main

drawbacks to this overlapping setting. The first and perhaps the most obvious is that solving multiple times in the overlap requires more work. The second is that in cases where the computational domain consists of two different materials a natural way to divide Ω is to follow the material separation and this is not possible with overlapping subdomains. In the next subsection we present two non overlapping domain decomposition methods.

2.1.2 Substructuring methods

Here we present two popular substructuring methods: the Balancing Domain Decomposition (BDD) method and the Finite Element Tearing and Interconnecting (FETI) method. The BDD method is due to Mandel [67] based on the Neumann-Neumann method by De Roeck and Le Tallec [14]. The FETI method is the work of Farhat and Roux [36]. Our objective in this introductive chapter is to illustrate the ideas underlying substructuring methods and for sake of clarity we focus on the linear elasticity problem. Rigorous definitions of these methods for a general symmetric positive definite matrix are given in Chapter 5. Thorough presentations of the substructuring methods can also be found in [58, 112, 45]. In particular [45] gives an insight into mechanical interpretations and implementation techniques.

Let Ω be an open subset of $\mathbb{R}^d$ for $d = 2$ or $d = 3$. Let $\partial\Omega$ be the boundary of Ω and $\partial\Omega_D \subset \partial\Omega$ be a part of the boundary where a homogeneous Dirichlet boundary condition is imposed. Next, introduce the space $\mathcal{V} := \{v \in H^1(\Omega) : v_{|\Omega_D} = 0\}$. For a given body force $f \in \mathcal{V}'$, the variational formulation of the linear elasticity equations can be written as: find the set of displacements $v \in \mathcal{V}$ such that

$$2\int_\Omega \mu\,\epsilon(u) : \epsilon(v)dx + \int_\Omega \lambda(\nabla\cdot u)(\nabla\cdot v)dx = \int_\Omega \langle f, v\rangle dx \quad \forall\, v \in \mathcal{V}, \tag{2.5}$$

where the linear strain tensors terms are

$$\epsilon(u) : \epsilon(v) := \sum_{i=1}^{d}\sum_{j=1}^{d} \epsilon_{ij}(u)\epsilon_{ij}(v); \quad \epsilon_{ij}(u) := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

$$\langle f, v\rangle := \sum_{i=1}^{d} f_i v_i,$$

and μ and λ are two parameters called the Lamé parameters which describe the material and can be expressed in terms of Young's modulus E and Poisson's ratio ν ($0 < \nu < 0.5$) as

$$\lambda := \frac{E\nu}{(1+\nu)(1-2\nu)},\ \mu := \frac{E}{2(1+\nu)}.$$

The linear elasticity equations are an approximation, for small deformations, of the elasticity equations [62]. Lets assume that we've discretized (2.5) using piecewise linear ($\mathbb{P}_1$) Lagrange finite element functions and that in matrix formulation the problem can be written as: Find $\hat{\mathbf{u}} \in \mathbb{R}^n$ such that

$$\hat{K}\hat{\mathbf{u}} = \hat{\mathbf{f}}.$$

Let the original computational domain Ω be partitioned into a set of open non overlapping subdomains

$$\Omega = \bigcup_{i=1}^{N} \overline{\Omega}_i; \quad \Omega_i \cap \Omega_j = \emptyset \quad \forall\, i \neq j.$$

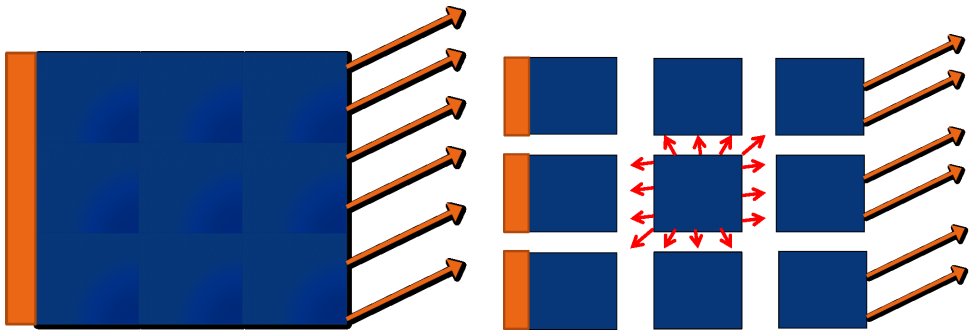


Figure 2.3: Partition of Ω into regular subdomains (left: original domain – right: partitioned domain). The domain is clamped on the left hand side and submitted to a surface force on the right hand side. The red arrows between subdomains in the partitioned domain represent the surface forces resulting from the interaction with that subdomain's neighbours.

We denote by K_i and $\mathbf{f}_i$ the local problem matrix and discrete body force corresponding to the discretization of $2\int_{\Omega_i}\mu\,\epsilon(u):\epsilon(v)dx+\int_{\Omega_i}\lambda(\nabla\cdot u)(\nabla\cdot v)dx$ and $\int_{\Omega_i}\langle f,v\rangle dx$ for the functions in $\{u_{|\Omega_i};u\in\mathbb{P}_1\}$. The equilibrium of subdomain Ω_i can be written as

$$K_i\mathbf{u}_i=\mathbf{f}_i+\mathbf{g}_i,\ \text{where } \mathbf{g}_i \text{ are surface forces.} \tag{2.6}$$

We notice that in this equilibrium equation an additional unknown has appeared: the surface force term $\mathbf{g}_i$ which represents the pressure exerted by neighbouring subdomains. This is what is illustrated with the red arrows in Figure 2.3.

Next we introduce a splitting of the degrees of freedom into boundary degrees of freedom which are shared by at least two subdomains and form the set

$$\Gamma:=\bigcup_{i,j=1,\dots,N;\,i\neq j}(\partial\Omega_i\cap\partial\Omega_j)\,,$$

and all other degrees of freedom denoted with I (for Interior). With obvious notation, the local equilibrium equation (2.6) can be rewritten in block formulation as

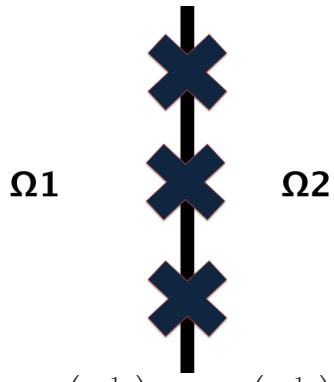
$$\begin{pmatrix}K_i^{II} & K_i^{I\Gamma}\\ K_i^{\Gamma I} & K_i^{\Gamma\Gamma}\end{pmatrix}\begin{pmatrix}\mathbf{u}_i^I\\ \mathbf{u}_i^\Gamma\end{pmatrix}=\begin{pmatrix}\mathbf{f}_i^I\\ \mathbf{f}_i^\Gamma\end{pmatrix}+\begin{pmatrix}\mathbf{0}\\ \mathbf{g}_i^\Gamma\end{pmatrix}. \tag{2.7}$$

By definition the interface forces are zero for degrees of freedom in the interior of Ω_i and using a Schur complement procedure we can also eliminate the interior displacement degrees of freedom. In system formulation (2.7) reads

$$\begin{cases}K_i^{II}\mathbf{u}_i^I+K_i^{I\Gamma}\mathbf{u}_i^\Gamma=\mathbf{f}_i^I,\\ K_i^{\Gamma I}\mathbf{u}_i^I+K_i^{\Gamma\Gamma}\mathbf{u}_i^\Gamma=\mathbf{f}_i^\Gamma+\mathbf{g}_i^\Gamma,\end{cases}$$

which is equivalent to

$$\begin{cases}\mathbf{u}_i^I=K_i^{II^{-1}}\left(\mathbf{f}_i^I-K_i^{I\Gamma}\mathbf{u}_i^\Gamma\right)\\ \underbrace{\left[K_i^{\Gamma\Gamma}-K_i^{\Gamma I}K_i^{II^{-1}}K_i^{I\Gamma}\right]}_{:=S_i}\mathbf{u}_i^\Gamma=\underbrace{\left[\mathbf{f}_i^\Gamma-K_i^{\Gamma I}K_i^{II^{-1}}\mathbf{f}_i^I\right]}_{:=\tilde{\mathbf{f}}_i}+\mathbf{g}_i^\Gamma.\end{cases} \tag{2.8}$$



$$\textbf{Assembly:} \quad R_1^\top \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} + R_2^\top \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} x_1^1 + x_2^1 \\ x_1^2 + x_2^2 \\ x_1^3 + x_2^3 \end{pmatrix}$$

$$\textbf{Jump:} \quad B_1 \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} + B_2 \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} x_1^1 - x_2^1 \\ x_1^2 - x_2^2 \\ x_1^3 - x_2^3 \end{pmatrix}$$

Figure 2.4: Illustration of the action of the jump operators B_i and assembly operators $R_i^\top$ on a two domain case where the interface consists of three degrees of freedom

As well as the local equilibrium equation $S_i \mathbf{u}_i^\Gamma = \tilde{\mathbf{f}}_i + \mathbf{g}_i^\Gamma$, each subdomain must satisfy continuity and compatibility constraints with its neighbours. These conditions are written using two interpolation operators as:

- The **assembly operators** $R_i^\top$ are boolean matrices: given a vector $\mathbf{u}_i^\Gamma$ of entries for the degrees of freedom on $\partial\Omega_i \cap \Gamma$, $R_i^\top \mathbf{u}_i$ is a vector of entries for the degrees of freedom on the whole of Γ which has the same values as $\mathbf{u}_i^\Gamma$ for the degrees of freedom on $\partial\Omega_i \cap \Gamma$ and is 0 everywhere else.
- The **jump operators** B_i are signed boolean matrices where each line in B_i corresponds to one degree of freedom x on Γ and two subdomains Ω_k and Ω_l such that $x \in \partial\Omega_k \cap \partial\Omega_l$. If $i = k$ or l then the entry in that line of B_i which corresponds to the local numbering of x is assigned -1 if $i = \min(k, l)$ and $+1$ otherwise.

The action of the assembly and jump operators is illustrated on a simplified example in Figure 2.4. With these operators the global elasticity problem $\hat{K}\hat{\mathbf{u}} = \hat{\mathbf{f}}$ can be rewritten as: For each subdomain $i = 1, \ldots, N$ find the set of displacements $\mathbf{u}_i^\Gamma$ and interface forces $\mathbf{g}_i^\Gamma$ such that

$$\begin{cases} S_i \mathbf{u}_i^\Gamma & = & \tilde{\mathbf{f}}_i + \mathbf{g}_i^\Gamma, \quad \forall i = 1, \ldots, N & \text{[Local Equilibrium]} \\ \sum_{i=1}^N B_i \mathbf{u}_i^\Gamma & = & \mathbf{0} & \text{[Interface compatibility]} \\ \sum_{i=1}^N R_i^\top \mathbf{g}_i^\Gamma & = & \mathbf{0} & \text{[Interface equilibrium].} \end{cases} \tag{2.9}$$

Both the FETI and the BDD formulation of the elasticity problem are based on (2.9).

BDD formulation The first step is to eliminate the interface compatibility constraint by searching for the displacements in the reduced set $\{(\mathbf{u}_1^\Gamma, \ldots, \mathbf{u}_N^\Gamma); \sum_{i=1}^N B_i \mathbf{u}_i^\Gamma = \mathbf{0}\}$ where it holds. By definition of the assembly and jump operators this is the same as looking for a vector $\hat{\mathbf{u}}^\Gamma$ defined on the global set of interfaces and then choosing $\mathbf{u}_i^\Gamma = R_i \hat{\mathbf{u}}^\Gamma$ in each

subdomain. With this (2.9) is equivalent to: Find $\hat{\mathbf{u}}^\Gamma$ and $\mathbf{g}_1^\Gamma, \ldots, \mathbf{g}_N^\Gamma$ such that

$$\left\{ \begin{array}{lcll} S_i R_i \hat{\mathbf{u}}^\Gamma & = & \tilde{\mathbf{f}}_i + \mathbf{g}_i^\Gamma, & \forall i = 1, \ldots, N \\ \sum_{i=1}^N R_i^\top \mathbf{g}_i^\Gamma & = & \mathbf{0}. & \end{array} \right. \tag{2.10}$$

Finally injecting the first line into the second we get the BDD formulation of the problem:

$$\left(\sum_{i=1}^N R_i^\top S_i R_i \right) \hat{\mathbf{u}}^\Gamma = \sum_{i=1}^N R_i^\top \tilde{\mathbf{f}}_i. \tag{2.11}$$

After solving this interface problem the interior degrees of freedom can be computed *via* the first equation in (2.8). Since the BDD operator is a sum of local Schur complements it is quite natural to precondition it by the sum of inverses of these local Schur complements, more precisely in the case where S_i is non singular the BDD preconditioner is

$$M^{-1} := \sum_{i=1}^N \widetilde{R_i^\top} S_i^{-1} \widetilde{R_i}, \tag{2.12}$$

where $\widetilde{R_i^\top}$ and $\widetilde{R_i}$ are the same operators as $R_i^\top$ and R_i but weighted by partitions of unity.

FETI formulation This time the interface equilibrium is eliminated by searching for the interface forces in the reduced set $\{(\mathbf{g}_1^\Gamma, \ldots, \mathbf{g}_N^\Gamma); \sum_{i=1}^N R_i^\top \mathbf{g}_i^\Gamma = \mathbf{0}\}$. By definition of the assembly and jump operators this is the same as looking for a vector $\boldsymbol{\lambda} \in \text{range}(\sum_{i=1}^N B_i)$ and choosing $\mathbf{g}_i^\Gamma = -B_i^\top \boldsymbol{\lambda}$ in each subdomain. With this (2.9) is equivalent to: Find $\boldsymbol{\lambda} \in \text{range}(\sum_{i=1}^N B_i)$ such that

$$\left\{ \begin{array}{lcll} S_i \mathbf{u}_i^\Gamma & = & \tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda}, & \forall i = 1, \ldots, N \\ \sum_{i=1}^N B_i \mathbf{u}_i^\Gamma & = & \mathbf{0}. & \end{array} \right. \tag{2.13}$$

Lets assume that we are in the highly unlikely case where S_i is non singular then we may write

$$\left\{ \begin{array}{lcll} \mathbf{u}_i^\Gamma & = & S_i^{-1}(\tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda}), & \forall i = 1, \ldots, N \\ \sum_{i=1}^N B_i \mathbf{u}_i^\Gamma & = & \mathbf{0}. & \end{array} \right. \tag{2.14}$$

and finally the FETI formulation of the problem is obtained by injecting the first line into the second:

$$\left(\sum_{i=1}^N B_i S_i^{-1} B_i^\top \right) \boldsymbol{\lambda} = \sum_{i=1}^N B_i S_i^{-1} \tilde{\mathbf{f}}_i. \tag{2.15}$$

After solving for $\boldsymbol{\lambda}$ the set of displacements can be computed via (2.14) for the interface degrees of freedom and (2.8) for the interior degrees of freedom. Since the FETI operator is a sum of inverses of the local Schur complements it is quite natural to precondition it by the sum of the Schur complements, more precisely the preconditioner for FETI is

$$M^{-1} = \sum_{i=1}^N \widetilde{B_i} S_i \widetilde{B_i}^\top, \tag{2.16}$$

where $\widetilde{B_i^\top}$ and $\widetilde{B_i}$ are the same operators as $B_i^\top$ and B_i but weighted by partitions of unity.

Remark 2.1. We have assumed that all the operators S_i are non singular. This is not the general case. How to handle singularities in S_i has been well understood since very early work on FETI and BDD [37, 67]. In the next section we describe one way to handle the kernels for elasticity and in Chapter 5 we consider the case of general symmetric positive semi-definite matrices S_i. .

2.1.3 Lack of robustness: a first illustration

Since we are concentrating on symmetric problems and symmetric preconditioners the iterative solver of choice is the Preconditioned Conjugate Gradient (PCG) algorithm which we present in Algorithm 2.1 (see [60, 51] for its first introduction and [95] for a modern presentation).

Algorithm 2.1 Preconditioned Conjugate Algorithm for $A\mathbf{x}_* = \mathbf{f}$ preconditioned by M^{-1} and initialized by $\mathbf{x}_0$.

$\mathbf{r}_0 := \mathbf{f} - A\mathbf{x}_0$; $\mathbf{z}_0 := M^{-1}\mathbf{r}_0$ and $\mathbf{p}_0 := \mathbf{z}_0$
for $j = 0, 1, \ldots$ until convergence **do**
  $\alpha_j := \langle \mathbf{r}_j, \mathbf{z}_j \rangle / \langle A\mathbf{p}_j, \mathbf{p}_j \rangle$
  $\mathbf{x}_{j+1} := \mathbf{x}_j + \alpha_j \mathbf{p}_j$
  $\mathbf{r}_{j+1} := \mathbf{r}_j - \alpha_j A\mathbf{p}_j$
  $\mathbf{z}_{j+1} := M^{-1}\mathbf{r}_{j+1}$
  $\beta_j := \langle \mathbf{r}_{j+1}, \mathbf{z}_{j+1} \rangle / \langle \mathbf{r}_j, \mathbf{z}_j \rangle$
  $\mathbf{p}_{j+1} := \mathbf{z}_{j+1} + \beta_j \mathbf{p}_j$
end for

One way to evaluate the robustness of a solver that is based on PCG is to use the following convergence result [73, 53] (see also [95](Theorem 6.29) for the proof):

$$\|\mathbf{x}_* - \mathbf{x}_m\|_A \leq \frac{\|\mathbf{x}_* - \mathbf{x}_0\|_A}{C_m\left(\frac{\lambda_{max} + \lambda_{min}}{\lambda_{max} - \lambda_{min}}\right)}, \tag{2.17}$$

in which

C_m is the Chebyshev polynomial of degree m of the first kind,

$\mathbf{x}_*$ is the exact solution,

$\mathbf{x}_m$ is the approximate solution returned by the m-th step of Algorithm 2.1,

λ_{max} and λ_{min} are the extreme eigenvalues of the preconditioned operator $M^{-1}A$,

$\|\cdot\|_A$ is the norm induced by the A inner product.

A simplification of this result is the following linear convergence bound:

$$\|\mathbf{x}_* - \mathbf{x}_m\|_A \leq 2\left[\frac{\sqrt{\lambda_{max}/\lambda_{min}} - 1}{\sqrt{\lambda_{max}/\lambda_{min}} + 1}\right]^m \|\mathbf{x}_* - \mathbf{x}_0\|_A. \tag{2.18}$$

Although these estimates are in general not sharp they tell us that, as long as bounds on the spectrum of the preconditioned operator are available, the error at iteration m can be bounded with respect to the original error and these bounds.

Our ambition here is to show that as soon as we consider simulations in heterogeneous media it is easy to build a test case for which the iterative solver becomes rather inefficient.

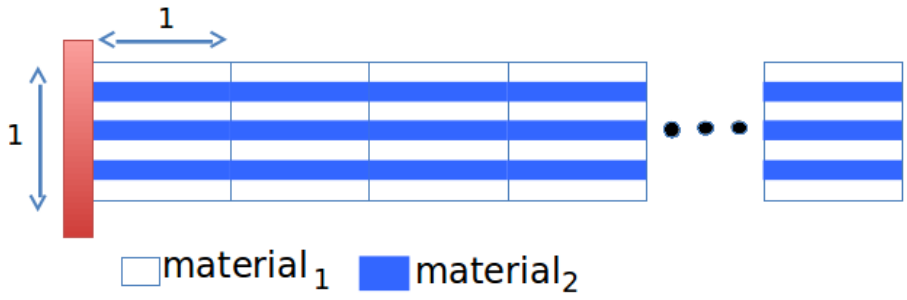


Figure 2.5: Geometry for the robustness test – the domain consists of seven layers of two different materials. Homogeneous Dirichlet boundary conditions are imposed on the left hand side, homogeneous Neumann boundary conditions are imposed on all other boundaries. The number N of subdomains (and thus the length of the domain) varies.

Iteration count:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	18	33	62	120
$\alpha_2 = 10^2$	24	37	64	117
$\alpha_2 = 10^4$	32	63	117	187
$\alpha_2 = 10^6$	21	51	107	208

Estimated Condition number:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$
$\alpha_2 = 10^2$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$
$\alpha_2 = 10^4$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$
$\alpha_2 = 10^6$	321	$1.37 \cdot 10^3$	$5.63 \cdot 10^3$	$2.29 \cdot 10^4$

Table 2.1: Convergence results for the **scalar elliptic problem** (2.19) discretized by $\mathbb{P}_1$ finite elements and preconditioned by **Additive Schwarz** (2.4). The geometry is given in Figure 2.5. Two layers of overlap are added to each subdomain. The coefficient in material 1 is $\alpha_1 = 1$. We make the number of subdomains and the coefficient α_2 in material 2 vary and report the number of iterations needed to reach convergence (top) and the estimate for the condition number of the preconditioned operator based on the Ritz values (bottom).

We consider the Additive Schwarz preconditioner (2.4) applied to a discretization of the scalar elliptic problem (also known as the Darcy equation)

$$\begin{cases} -\nabla \cdot (\alpha \nabla \mathbf{u}) & = f, & \text{in } \Omega, \\ \mathbf{u}(x,y) & = 0, & \text{if } x = 0, \\ \frac{\partial \mathbf{u}}{\partial \mathbf{n}}(x,y) & = 0, & \text{on the remainder of } \partial\Omega, \end{cases} \tag{2.19}$$

where $\Omega = [0, N] \times [0, 1]$. The problem is discretized by standard $\mathbb{P}_1$ finite elements on a $(20N + 1) \times 21$ regular mesh ($N \in \mathbb{N}$) where the diffusivity parameter is a real valued function $\alpha : \Omega \to \mathbb{R}^+$. The domain consists of two different materials (characterized by two values of α: α_1 and α_2) placed in seven alternating layers as illustrated in Figure 2.5. In order to build the partition we first divide Ω into N non overlapping unit squares and then add two layers of overlap to each of these subdomains.

In Table 2.1 we report the results of our convergence test. We give the number of iterations needed to achieve convergence and also the estimate for the condition number

of the preconditioned operator $M^{-1}A$ which is based on the Ritz values of this operator at the final iteration of CG (see [15] for instance). The stopping criterion is based on the relative error:

$$\frac{\|x_* - x_m\|_\infty}{\|x_*\|_\infty} < 10^{-6}.$$

We observe that the number of iterations needed to achieve convergence grows both with the number of subdomains and the magnitude of the jump in the coefficients. An exception is the case where $\alpha_2 = 10^6$ which requires fewer iterations than the case $\alpha_2 = 10^4$. The fact that in the table the estimated condition number depends only on the number of subdomains is not a typo.

The Schwarz preconditioner which we've introduced is far from being the state of the art Schwarz preconditioner. In the next section we will show that there is a well known improvement to recover robustness with respect to the number of subdomains in the constant coefficient case. Even with this improvement, the lack of robustness with respect to heterogeneous materials will remain an issue and is a perfect illustration of the family of problems which we tackle in the subsequent chapters of this manuscript. But first we present some ways of improving robustness by acting on the partition into subdomains.

2.1.4 Acting on the partition to improve robustness

Classical domain decomposition methods are known to be robust for good choices of the partition into subdomains, see e.g. [23, 22, 68]. Although some generalizations of these results exist (cf. [86, 85, 98]...) it is undeniable that acting on the partition into subdomains can really help with robustness.

FETI and BDD As we have already mentioned a great advantage of domain decomposition methods without overlap is that in cases where the simulation domain consists of several materials it is possible to partition the domain in such a way that heterogeneities coincide with the subdomain interfaces. Then by making a smart choice for the weights in the FETI and BDD preconditioners just as good a convergence can be recovered as in the constant coefficient case. For FETI the idea goes back to [93] where weights based on the diagonals of the stiffness matrices are introduced and a mechanical interpretation is given. The subtlety lies in the fact that the weights for the displacement variables and the force variables are not identical but they are connected. In [58] a mathematical formulation of this choice of weights is given which makes it possible to write the preconditioner in a general form and enables the theoretical study. The corresponding results for BDD can also be found there.

More recently, for FETI and the Darcy equation, the authors in [87, 84, 83] prove that even in some particular configurations where the heterogeneity overlaps the boundary it does not affect convergence but this is far from being the general case.

What about additive Schwarz? It is Pierre-Louis Lions who in 1990 derived the first non overlapping version of the Alternating Schwarz method [66]. The trick is to replace the Dirichlet transmission conditions in (2.2) by Robin transmission conditions for a parameter β:

$$u_1^{n+1} + \beta \frac{\partial}{\partial \mathbf{n}_1} u_1^{n+1} = u_2^n + \beta \frac{\partial}{\partial \mathbf{n}_1} u_2^n \text{ on } \partial\Omega_2 \cap \partial\Omega_1,$$

in the first step of the iteration and

$$u_2^{n+1} + \beta \frac{\partial}{\partial \mathbf{n}_2} u_2^{n+1} = u_1^{n+1} + \beta \frac{\partial}{\partial \mathbf{n}_2} u_1^{n+1} \text{ on } \partial\Omega_2 \cap \partial\Omega_1,$$

in the second step of the iteration ($\mathbf{n}_1$ and $\mathbf{n}_2$ are the unit outward normals for subdomains Ω_1 and Ω_2). Lions proves that this algorithm applied to the Poisson problem converges without overlap for any number of subdomains. The idea to change the transmission conditions was further developed into looking for the optimized transmission conditions within the whole range of linear transmission conditions. It appears that the optimal transmission conditions were highly non local with the result that the optimized solver is usually too expensive to use and truncations are necessary. Among the vast literature we refer to [77, 16, 42, 40] and references therein.

AGMG Multigrid methods [48] are very closely related to domain decomposition. From the domain decomposition point of view a multigrid method is a recursive application of domain decomposition: the global domain is divided into subdomains which are themselves divided into subdomains and so on until the partition is the mesh. From the multigrid point of view, a domain decomposition method is a multigrid method where, starting at the mesh level, the coarsening procedure is applied just once and results in the partition into subdomains.

Algebraic multigrid [5, 94] is a variant of the original multigrid algorithms which does not require any information on the geometry of the underlying problem. This is particularly interesting if the matrix stems from a gridless problem or a problem on an unstructured grid but more generally algebraic methods are advantageous because they can be implemented as black box algorithms without any information other than the problem matrix and right hand side.

For problems with heterogeneous coefficients Algebraic Multigrid Techniques may build *aggregates* (the multigrid counterpart for subdomains) which do not resolve the heterogeneities and hence may lead to slow convergence. To overcome this fact, in [75] the authors propose an Algebraic Multigrid method based on aggregation which has a guaranteed convergence rate. This is the first complete convergence analysis of an AMG method with plain aggregation, it is based on their previous work [74]. The result holds for symmetric M-matrices with non negative row sum. The idea is to act on the way that the aggregates are formed: a maximal convergence bound is defined by the user and this is translated into a quality constraint for the aggregates. Then the aggregates are built adaptively in such a way that the quality constraint is always satisfied: just as with the domain decomposition methods, convergence can be highly accelerated by acting on the partition of the domain.

Drawbacks of this strategy In this thesis work, one of the objectives was not to rely on the assumption that a partition can be built which resolves heterogeneities. Indeed the endgame is to solve hard industrial problems and more particularly the problems that Michelin is faced with of which Figure 2.6 is an example. The reasons for which we decided not to rely on a specific partition of the subdomains to ensure good convergence include the following

- The material distribution in figure 2.6 suggests that if the subdomains resolve the heterogeneities then they will have very bad aspect ratios and in some cases they may even be plates or shells. This means that the local problems could potentially be very ill conditioned and solving them would require extra work and fine techniques.

Figure 2.6: An illustration of the type of problems Michelin are faced with

– If the partition into subdomains must accommodate the materials this must be implemented in a part of the code where the material distribution is known. This goes against the objective of having a black box solver which interferes the least possible with existing and future code.
– Perhaps the most decisive argument is that heterogeneous materials is only one of the parameters which slows down convergence and what we are aiming for is a solver that tackles various kinds of difficulties.

To put things into perspective, in an industrial context, the question is not only: 'is there a partition into subdomains which resolves all the heterogeneities and leads to well conditioned local problems?' but rather 'how much work would it require for an engineer to build this partition?'. If it is possible to partition the domain automatically with software such as Metis [54] or Scotch [13] and then let the solver do the hard work this is a very attractive perspective and building that solver is exactly what we will aim for in subsequent chapters.

2.2 Two Level Methods: toward robustness

The lack of robustness which we pointed out in the previous section can be explained by a lack of global communication between subdomains: during one iteration a subdomain only exchanges information with its neighbours or in some cases (preconditioned FETI and BDD) the neighbours of its neighbours. For this reason a remedy is to add a global communication mechanism to the algorithm. This is what is called a second level. Among the first of this type were the BPS preconditioner introduced by Bramble, Paschiak and Schatz [4] and the two-level overlapping Schwarz method introduced by Dryja and Widlund [24]. The idea is to use a direct solver not only in each of the subdomains but also on a subproblem which is shared by all subdomains: the coarse problem. This coarse problem is a rough approximation of A and choosing it will be the subject of a large part of this manuscript. Before diving in we present the abstract Schwarz framework [61, 112]: a theoretical framework for studying domain decomposition methods.

2.2.1 Abstract Schwarz framework

This subsection is adapted from the book by Toselli and Widlund [112](Chapter 2). We refer to there for the bibliographical details of the emergence of the Schwarz theory. We do mention contributions [80] and [117] which are often deemed crucial. Some of the notation which we introduce has already been used in the previous section. This is not a problem since here we generalize the same notions. Given a finite dimensional Hilbert space V, given a symmetric, positive definite bilinear form,

$$a(\cdot,\cdot) : V \times V \to \mathbb{R},$$

and an element $f \in V'$, we consider the problem of finding $u \in V$, such that

$$a(u,v) = f(v),\ v \in V. \tag{2.20}$$

If A is the stiffness matrix relative to the bilinear form $a(\cdot,\cdot)$ and a given basis for V, and $\mathbf{f}$ is the vector relative to f and the same basis then problem (2.20) is equivalent to the linear system

$$A\mathbf{u} = \mathbf{f}, \tag{2.21}$$

with A symmetric, positive definite. We next consider a family of spaces $\{V_i, i = 0,\dots,N\}$ and suppose that there exist interpolation operators

$$R_i^\top : V_i \to V.$$

We assume that V admits the following decomposition (this is not necessarily a direct sum)

$$V = R_0^\top V_0 + \sum_{i=1}^{N} R_i^\top V_i. \tag{2.22}$$

Notice that the subspaces are now numbered between 0 and N. Although this makes no difference for this abstract definition, in many cases V_0 will be a particular space: the coarse space and the N other spaces V_i will be the usual subdomains based on geometry.

We next introduce local symmetric, positive definite, bilinear forms on the subspaces,

$$\tilde{a}_i(\cdot,\cdot) : V_i \times V_i \to \mathbb{R}, \quad i = 0,\dots,N,$$

and the local stiffness matrices associated with them,

$$\tilde{A}_i : V_i \to V_i.$$

Schwarz operators are defined in terms of projection like operators

$$P_i = R_i^\top \tilde{P}_i : V \to R_i^\top V_i \subset V, i = 0,\dots,N,$$

where $\tilde{P}_i : V \to V_i$, is defined by

$$\tilde{a}_i(\tilde{P}_i u, v_i) = a(u, R_i^\top v_i), \quad v_i \in V_i. \tag{2.23}$$

We note that $\tilde{P}_i$ is well defined since the local bilinear forms are coercive.

Remark 2.2. In case we want to use the original bilinear form on a subspace V_i, we choose

$$\tilde{a}_i(u_i, v_i) = a(R_i^\top u_i, R_i^\top v_i), \quad u_i, v_i \in V_i \tag{2.24}$$

and find that

$$\tilde{A}_i = R_i A R_i^\top = A_i. \tag{2.25}$$

In this case we say that we use an *exact solver on* V_i.

We have the following lemma.

Lemma 2.3. The P_i can be written as

$$P_i = R_i^\top \tilde{A}_i^{-1} R_i A, \quad 0 \leq i \leq N. \tag{2.26}$$

In addition the P_i are self adjoint with respect to the inner product induced by $a(\cdot,\cdot)$ and positive semi-definite. If moreover the local bilinear form is given by (2.24), then P_i is a projection, i.e.,

$$P_i^2 = P_i. \tag{2.27}$$

From now on we make the following assumption.

Assumption 2.4. An exact solver is used on the coarse space V_0. (Then, P_0 is an A-orthogonal projection.)

Based on the projection like operators P_i three families of preconditioned Schwarz operators are

1. Additive :
$$P_{ad} := \sum_{i=0}^{N} P_i. \tag{2.28}$$
2. Multiplicative:
$$P_{mu} := I - (I - P_N)(I - P_{N-1}) \dots (I - P_0). \tag{2.29}$$
3. Hybrid:
$$P_{hy} := P_0 + (I - P_0) \sum_{i=1}^{N} P_i (I - P_0). \tag{2.30}$$

The convergence bounds in the Abstract Schwarz Framework rely on Assumption 2.4 and three additional Assumptions.

Assumption 2.5 (Strenghtened Cauchy-Schwarz inequalites)**.** There exist constants $0 \leq \epsilon_{ij} \leq 1$, $1 \leq i, j \leq N$, such that

$$|a(R_i^\top u_i, R_j^\top u_j)| \leq \epsilon_{ij} a(R_i^\top u_i, R_i^\top u_i)^{1/2} a(R_j^\top u_j, R_j^\top u_j)^{1/2},$$

for $u_i \in V_i$ and $u_j \in V_j$. We will denote the spectral radius of $\epsilon = \{\epsilon_{ij}\}$ by $\rho(\epsilon)$.

Assumption 2.6 (Local Stability)**.** There exists $\omega > 0$ such that

$$a(R_i^\top u_i, R_i^\top u_i) \leq \omega \tilde{a}_i(u_i, u_i), \quad u_i \in \text{range}(\tilde{P}_i) \subset V_i,\ 1 \leq i \leq N. \tag{2.31}$$

Assumption 2.7 (Stable splitting)**.** There exists a constant C_0, such that every $u \in V$ admits a decomposition

$$u = \sum_{i=0}^{N} R_i^\top u_i, \quad \{u_i \in V_i, 0 \leq i \leq N\}$$

that satisfies

$$\sum_{i=0}^{N} \tilde{a}_i(u_i, u_i) \leq C_0^2 a(u, u).$$

The following theorem gives some convergence results for the Schwarz Domain Decomposition Methods.

Theorem 2.8. Let Assumptions 2.4, 2.5, 2.6 and 2.7 hold. Then the operators defined by (2.28), (2.29) and (2.30) satisfy, for any $u \in V$,

$$\begin{aligned} C_0^{-2} a(u,u) \leq \quad & a(P_{ad}u,u) \quad \leq \omega(\rho(\epsilon)+1)a(u,u), \\ \max(1,C_0^2)^{-1} a(u,u) \leq \quad & a(P_{hy}u,u) \quad \leq \max(1,\omega\rho(\epsilon))a(u,u). \end{aligned}$$

and, under the Assumption that $\omega < 2$,

$$\|I - P_{mu}\|_A \leq 1 - \frac{2-\omega}{(2\max(1,\omega)^2\rho(\epsilon)^2+1)C_0^2} < 1,$$

In fact, for the hybrid operator it is sufficient that Assumption 2.7 hold on range$(I-P_0)$. The additive and hybrid operators P_{ad} and P_{hy} are symmetric so the natural choice is to solve with PCG. As stated in (2.17) the convergence of PCG is bounded with respect to the extreme eigenvalues of the preconditioned operators and these extreme values are in turn related to the Rayleigh quotients in the theorem so the two first results in Theorem 2.8 are indeed convergence results. The multiplicative variant P_{mu} is not symmetric. Rather than a preconditioned operator to be solved with an iterative method it is the generalized discrete variant of the Alternating Schwarz method (2.2) and would typically be solved by a Richardson iteration. The norm in the theorem is the A-norm of the error propagator which is indeed relevant to the convergence behavior of the algorithm.

Assumption 2.5 is usually proved using the following Lemma and then the constant $\rho(\epsilon)$ in the convergence results depends only on the geometry of the partition (but not on the number of subdomains).

Lemma 2.9. Suppose that the local subspaces V_i, $i = 1, \ldots, N$ have been colored in such a way that two subspaces V_k and V_l which have the same color are orthogonal

$$P_k P_l = P_l P_k = 0$$

and that this required N^C colors. Then Assumption 2.5 holds and $\rho(\epsilon) \leq N^C$.

For the proofs of all of these results and more detail on the Abstract Schwarz framework we, again, refer the reader to [112](Chapter 2).

The Additive Schwarz preconditioner introduced in Subsection 2.1.1 is P_{ad} for an empty coarse space $V_0 = \emptyset$ and exact solvers on each of the local subspaces: *i.e.* $\tilde{a}_i$ is defined according to (2.24). Assuming that we have chosen a non empty coarse space V_0 we define the two level Additive Schwarz preconditioner as follows.

Definition 2.10. The two level Additive Schwarz preconditioner is the Additive preconditioner where exact solvers are used on all subspaces

$$M^{-1} := \sum_{i=0}^{N} R_i^\top A_i^{-1} R_i; \quad A_i := R_i A R_i^\top. \tag{2.32}$$

Remark 2.11. Since the local solvers are all exact solvers Assumption 2.6 is automatically satisfied with $\omega = 1$. Then, by Lemma 2.9 and Theorem 2.8 the condition number of $M^{-1}A$ depends only on the stable splitting property (Assumption 2.7) and the condition number of $M^{-1}A$ is bounded by $C_0^{-2}(N^C + 1)$. To simplify things further, in this expression the number N^C of colors can be replaced by the maximal number of subdomains to which one element of the mesh belongs [25, Section 4].

2.2.2 Coarse spaces based on the zero energy modes

We introduce the most simple family of coarse spaces. They are based on the kernels of the local problems (by this we mean the problem restricted to a subdomain). It is by now common knowledge that a good coarse space should at least contain these zero energy modes. In some cases (FETI or preconditioned BDD) handling the zero energy modes is an absolute requirement.

Two level Additive Schwarz, Nicolaides and Partition of unity coarse spaces Considering the Poisson problem on Ω, Nicolaides proposed as early as 1987 [80] to accelerate convergence by partitioning the domain Ω into non overlapping subdomains Ω_i^*, $i = 1, \ldots, N$ and using the space of functions which are constant on each of these subdomains as a coarse space:

$$V_0^{NICO} = \text{span}(\mathbf{1}_{\Omega_1^*}, \ldots, \mathbf{1}_{\Omega_N^*}),$$

where $\mathbf{1}_{\Omega_i^*}$ is the indicator function of Ω_i^*.

There is however a significant drawback with the Nicolaides coarse space: the basis functions have an energy of the order H/h where H is the subdomain size and h is the mesh size. For this reason it is unrealistic to hope that the convergence of the resulting two level method will not depend on the mesh size. The solution is to make the basis functions decrease as smoothly as possible in the overlap.

In [97] Sarkis introduces and analyzes a *Partition of unity* coarse space (spanned by one basis function per subdomain) for which the two level Schwarz method applied to the Poisson problem converges independently of the mesh size and the number of subdomains. More precisely the condition number depends only linearly on the relative portion of a subdomain that is overlapped by others:

$$\kappa(M^{-1}A) \leq C\left(1 + \frac{H}{\delta}\right), \tag{2.33}$$

where H is the subdomain size, δ is the width of the overlap and C is a constant that depends on the geometry of the partition but not on H, δ or the mesh size h. Some additional assumptions on the shape of the domains are required.

We refer to [97] (or [112](Lemma 3.24)) for the construction of the Partition of unity coarse space in a general framework and the proof of this result. In the case where the mesh is regular the basis function for a given $i = 1 \ldots, N$ takes the value 1 in the part of Ω_i that is not overlapped by other subdomains, 0 outside Ω_i and it decreases linearly from 1 to 0 in the overlap.

The main reason why the functions that are piecewise constant in the interior of each subdomain need to be in the coarse space is what is referred to in [112] as the quotient space argument. A crucial argument in the proof of convergence is the Poincaré inequality: assume that $1 \leq p \leq \infty$ and that $\tilde{\Omega}$ is a bounded connected open subset of $\mathbb{R}^n$ with a Lipschitz boundary. Then there exists a constant C depending only on $\tilde{\Omega}$ and p such that every function u in the Sobolev space $W^{1,p}(\tilde{\Omega})$ satisfies

$$\|u - u_{\tilde{\Omega}}\|_{L^p(\tilde{\Omega})} \leq C\|\nabla u\|_{L^p(\tilde{\Omega})}; \quad u_{\tilde{\Omega}} = \frac{1}{|\tilde{\Omega}|}u(y)dy.$$

Thanks to the Partition of unity coarse space, this inequality (for $p = 2$) can be applied to local functions with zero mean value.

Iteration count:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	18	24	24	23
$\alpha_2 = 10^2$	22	25	25	24
$\alpha_2 = 10^4$	36	62	95	128
$\alpha_2 = 10^6$	31	51	89	154

Estimated Condition number:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	28.0	28.2	28.1	28.1
$\alpha_2 = 10^2$	28.0	28.2	28.1	28.1
$\alpha_2 = 10^4$	415	$1.29 \cdot 10^3$	$2.39 \cdot 10^3$	$1.36 \cdot 10^3$
$\alpha_2 = 10^6$	479	$2.02 \cdot 10^3$	$7.96 \cdot 10^3$	$2.76 \cdot 10^4$

Table 2.2: Convergence results for the scalar elliptic problem (2.19) discretized by $\mathbb{P}_1$ finite elements and preconditioned by two level Additive Schwarz (2.32) where V_0 is the **Partition of Unity** coarse space (piecewise constants in the interior of each subdomain and linear decay to zero in the overlap). The geometry is given in Figure 2.5. Two layers of overlap are added to each subdomain. The coefficient in material 1 is always $\alpha_1 = 1$. We make the number of subdomains and the jump in the coefficient vary through α_2 (coefficient in material 2) and report the number of iterations needed to reach convergence (top) and the estimate for the condition number of the preconditioned operator based on the Ritz values (bottom).

Finally we illustrate the efficiency of the Partition of unity coarse space numerically. We run the same robustness test as in Table 2.1 with the two level preconditioner and display the results in Table 2.2, we observe that in the case where the coefficients are constant ($\alpha_2 = \alpha_1 = 1$) or almost constant ($\alpha_2 = 100$) the convergence no longer deteriorates with the number of subdomains. However the Partition of unity coarse space does not contain enough information to ensure that the method is robust even with large jumps in the coefficients.

BDD, FETI and the Rigid body mode coarse space Recall from (2.11) and (2.12) that the BDD formulation of the elasticity problem is

$$\left(\sum_{i=1}^{N} R_i^\top S_i R_i\right) \hat{\mathbf{u}}^\Gamma = \sum_{i=1}^{N} R_i^\top \tilde{\mathbf{f}}_i \text{ preconditioned by } M^{-1} = \sum_{i=1}^{N} \widetilde{R_i^\top} S_i^{-1} \widetilde{R_i}.$$

We have assumed that the inverse S_i^{-1} is defined. This is usually not the case so the general form of the BDD preconditioner should rather be $\left(\sum_{i=1}^{N} \widetilde{R_i^\top} S_i^\dagger \widetilde{R_i}\right)$ with $S_i^\dagger$ a pseudo-inverse of S_i. Since $S_i^\dagger$ is only defined on range(S_i) it is necessary to introduce projection operators to ensure that the residuals remain in this space. This is done using a hybrid preconditioner (2.30). If the weighted operators $\widetilde{R_i^\top}$ are equal to $R_i^\top D_i^{-1}$ for a diagonal matrix D_i then the general form of the Balancing Domain Decomposition operator [67] is

$$P_{bdd} = P_0 + (I - P_0) \left(\sum_{i=1}^{N} \widetilde{R_i^\top} S_i^\dagger \widetilde{R_i}\right) \underbrace{\left(\sum_{i=1}^{N} R_i^\top S_i R_i\right)}_{:=\hat{S}} (I - P_0), \tag{2.34}$$

where the coarse projector P_0, coarse interpolation operator $R_0^\top : V_0 \to V$ and coarse space V_0 are defined as

$$P_0 := R_0^\top S_0^{-1} R_0 \hat{S},\ S_0 := R_0 \hat{S} R_0^\top,\ V_0 := \text{range}(R_0^\top) = \sum_{i=1}^{N} R_i^\top D_i \left(\text{Ker}(S_i)\right).$$

We call V_0 the rigid body mode coarse space because for elasticity the kernel of S_i is the trace of the rigid body modes of Ω_i on the boundary.

For FETI things are slightly more complicated. Indeed recall from (2.15) and (2.16) that the FETI formulation of the elasticity problem is

$$\left(\sum_{i=1}^{N} B_i S_i^{-1} B_i^\top\right) \boldsymbol{\lambda} = \sum_{i=1}^{N} B_i S_i^{-1} \tilde{\mathbf{f}}_i \text{ preconditioned by } M^{-1} = \sum_{i=1}^{N} \widetilde{B_i} S_i \widetilde{B_i}.$$

As soon as one of the S_i is singular the reformulation of the problem must be adapted. Indeed if $\mathcal{R}_i^\top$ is an interpolator from $\mathbb{R}^{\dim(\text{Ker}(S_i))}$ into the kernel of S_i then $\mathbf{u}_i^\Gamma = S_i^{-1}(\tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda})$ in (2.14) must be replaced by

$$\mathbf{u}_i^\Gamma = S_i^\dagger (\tilde{\mathbf{f}}_i - B_i^\top \boldsymbol{\lambda}) + \mathcal{R}_i^\top \boldsymbol{\alpha_i}, \quad \boldsymbol{\alpha_i} \in \mathbb{R}^{\dim(\text{Ker}(S_i))}.$$

We do not wish to go into the details here, they will be presented in Chapter 5. What is important is that even in the case where S is singular it is possible to rewrite the elasticity problem as a problem on the interfaces forces. The main difference is that the interface forces no longer live in the whole of $U = \text{range}\left(\sum_{i=1}^{N} B_i\right)$ but rather in the subset of admissible constraints

$$\{\boldsymbol{\lambda} \in U; G^\top \boldsymbol{\lambda} = \sum_{i=1}^{N} \mathcal{R}_i^\top \tilde{\mathbf{f}}_i\} \text{ with } G := \sum_{i=1}^{N} B_i \mathcal{R}_i^\top.$$

For this reason, the iterative solver for FETI is the Projected Preconditioned Conjugate Gradient algorithm (PPCG): it is initialized with an initial guess $\boldsymbol{\lambda}_0$ which satisfies the constraint ($G^\top \boldsymbol{\lambda}_0 = \sum_{i=1}^{N} \mathcal{R}_i^\top \tilde{\mathbf{f}}_i$) and then the search directions are all projected into the space $\text{Ker}(G^\top)$ so that all subsequent approximations also satisfy the constraint. If P is a projection operator onto $\text{Ker}(G^\top)$ the preconditioned FETI operator is

$$\underbrace{P \left(\sum_{i=1}^{N} \widetilde{B_i} S_i \widetilde{B_i}\right)}_{M^{-1}} P^\top \underbrace{\left(\sum_{i=1}^{N} B_i S_i^{-1} B_i^\top\right)}_{:=F}.$$

Another big difference with Additive Schwarz and BDD is that we cannot use an exact solver in the definition of projection P in order to get an F-orthogonal projection. Indeed the whole point of P is to project out of the space where the FETI operator F is not defined. Instead we use the next best direction available, namely $(M^{-1})^{-1}$. More precisely the projection operator is

$$P = I - M^{-1} G \left(G^\top M^{-1} G\right)^{-1} G^\top,$$

and it is $(M^{-1})^{-1}$-orthogonal.

The resulting problem is usually solved with the projected preconditioned conjugate gradient algorithm (PPCG) algorithm.

Algorithm 2.2 PPCG: Projected Preconditioned Conjugate Algorithm for $PM^{-1}P^\top F\boldsymbol{\lambda} = PM^{-1}P^\top \mathbf{d}$ (where $\mathbf{d} := BS^\dagger \mathbf{f}$ is the right hand side for FETI)

$\boldsymbol{\lambda}_0 := M^{-1}G\left(G^\top M^{-1}G\right)^{-1}\sum_{i=1}^N \mathcal{R}_i^\top \tilde{\mathbf{f}}_i$
$\mathbf{r}_0 := P^\top(\mathbf{d} - F\boldsymbol{\lambda}_0)$; $\mathbf{z}_0 := M^{-1}\mathbf{r}_0$; $\mathbf{p}_0 := \mathbf{z}_0$
for $j = 0, 1, \ldots$ until convergence **do**
  $\mathbf{p}_j := P\mathbf{p}_j$
  $\alpha_j := \langle \mathbf{r}_j, \mathbf{z}_j\rangle / \langle F\mathbf{p}_j, \mathbf{p}_j\rangle$
  $\boldsymbol{\lambda}_{j+1} := \boldsymbol{\lambda}_j + \alpha_j \mathbf{p}_j$
  $\mathbf{r}_{j+1} := \mathbf{r}_j - \alpha_j P^\top F\mathbf{p}_j$
  $\mathbf{z}_{j+1} := M^{-1}\mathbf{r}_{j+1}$
  $\beta_j := \langle \mathbf{r}_{j+1}, \mathbf{z}_{j+1}\rangle / \langle \mathbf{r}_j, \mathbf{z}_j\rangle$
  $\mathbf{p}_{j+1} := \mathbf{z}_{j+1} + \beta_j \mathbf{p}_j$
end for

The PPCG algorithm can of course be used for any Hybrid Schwarz type operator (2.30). For a detailed study of the variants of CG available for this problem see [110, 56].

Recall that for Additive Schwarz with the Partition of unity coarse space applied to the Poisson problem the convergence bound doesn't depend on the number of subdomains (see (2.33)), instead it depends on the amount of overlap H/δ. For the scalar elliptic problem, FETI is by construction equipped with a second level related to the kernel of S which turns out to be the trace of the constant function on the boundaries of the subdomains. For this reason it is quite natural that, once more, the condition number of the preconditioned operator depends only weakly on the number of subdomains [58] through the number of elements in one subdomain:

$$\kappa(PM^{-1}P^\top F_{|\text{range}(P)}) \leq C\left(1 + \log\left(\frac{H}{h}\right)\right)^2, \tag{2.35}$$

where H is the subdomain size and h is the mesh size. Some regularity assumptions on the subdomains are required.

2.2.3 Analytical coarse spaces

There are a number of problems for which a good, if not the optimal, choice for the coarse space is available in the literature. Here we give a brief review of some of the contributions that propose such an analytical coarse space.

Coarse spaces based on the coefficient distribution For some families of heterogeneous media, coarse spaces have been proposed that are adapted to the heterogeneities. For the scalar elliptic problem in binary media (only two different materials) we refer to [115], [47] and the very many references therein. In [115] the authors consider the same problem as the one in Figure 2.5. Their objective is to apply a Projected Preconditioned Conjugate Gradient method to solve this problem. Instead of a Domain Decomposition preconditioner they use an incomplete Cholesky preconditioner but its behaviour is related to the Additive Schwarz preconditioner because both these preconditioners deal well with high frequency components and encounter problems for the low frequency ones. Their conclusion is that the coarse space must consist of as many vectors as there are high coefficient layers. In [47] a coarse space is proposed with one coarse basis vector per subdomain. It

is proved that this coarse space is sufficient to take care of all heterogeneities as long as they are *small islands* that only intersect the boundary of the subdomain once.

Other Coarse spaces For FETI the first coarse space was introduced in [32] to achieve scalability even for time dependent problems in mechanics. These problems usually have a zero order term and so the local operators are non singular and there is no natural coarse space enticed by the rigid body modes. The authors in [32] propose to use the rigid body modes to define a coarse space anyway.

Another problem for which research of a coarse space was very active is the case of very thin domains. In [35] for FETI a coarse space is proposed for plates and then adapted in [33] to shells. For BDD a coarse space for plate and shell problems is introduced, analyzed theoretically and tested numerically in [63].

Finally we mention that for the linear elasticity equations [17] proposes a coarse space for the two level Schwarz method with guaranteed convergence even in the incompressible limit. In [18] the same authors further reduce the size of the coarse space and also prove a convergence bound that is independent of the material properties. For FETI, a conclusion has been reached by [114] and more recently analyzed in [43]: one coarse vector per subdomain ensures robustness with respect to the almost incompressible behaviour.

Our objective The coarse spaces which we have just mentioned are optimal in the sense that, for a given partition into subdomains and a given problem, we can't hope to achieve robustness with respect to the particular difficulties in the problem by using a smaller coarse space. The drawback is that the spaces cannot be built automatically without prior knowledge of the particular challenges which must be tackled. Our ambition throughout this manuscript is to design coarse spaces which can deal with all kinds of heterogeneities as well as other difficulties and are built automatically without requiring the knowledge of the underlying set of partial differential equations. In cases where the optimal results apply they are good points of comparisons for the new methods. In particular we will test our methods on cases with heterogeneous coefficients and elasticity in the incompressible limit. Generalized eigenvalue problems will be a crucial tool.

2.2.4 Coarse spaces that rely on generalized eigenvalue problems

How generalized eigenvalue problems can help Once a domain decomposition method has been reformulated to fit the Abstract Schwarz framework proving that it converges comes down to ensuring that the local solvers are stable (Assumption 2.6) and that each vector admits a stable splitting (Assumption 2.7). These are inequalities between energy norms and we can ensure robustness by making sure that they hold with constants that do not depend on any of the parameters with respect to which we want the method to be robust. With this in mind we explain why generalized eigenvalues problems will be one of the most crucial tools in the automatic construction of our robust domain decomposition methods. First we define generalized eigenvalue problems.

Definition 2.12 (Generalized eigenvalue problem)**.** Let $\widetilde{A}$ and $\widetilde{B}$ be two symmetric matrices in $\mathbb{R}^{n\times n}$. Then the generalized eigenvalues associated with the 'pencil' $(\widetilde{A}, \widetilde{B})$ are the $\lambda \in \mathbb{R} \cup \{+\infty\}$ which satisfy:

– either $\lambda \in \mathbb{R}$ and there exists $\mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ such that

$$\widetilde{A}\mathbf{x} = \lambda \widetilde{B}\mathbf{x}, \tag{2.36}$$

– or $\lambda = +\infty$ and there exists $\mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ such that

$$\widetilde{B}\mathbf{x} = \mathbf{0}, \text{ and } \widetilde{A}\mathbf{x} \neq \mathbf{0}.$$

In both cases $\mathbf{x}$ is called a generalized eigenvector associated with the generalized eigenvalue λ and $(\lambda, \mathbf{x})$ is a generalized eigenpair of pencil $(\tilde{A}, \tilde{B})$.

The definition above allows for infinite eigenvalues. This can be better understood by realizing that if $(+\infty, \mathbf{x})$ is an eigenpair for the pencil $(\widetilde{A}, \widetilde{B})$ then $(0, \mathbf{x})$ is an eigenpair for the pencil $(\widetilde{B}, \widetilde{A})$ and there is no reason to discriminate between both formulations. In cases where the matrices are symmetric and $\widetilde{B}$ is non singular there are no infinite eigenvalues and crucial properties on the eigenvalues and eigenvectors arise.

Lemma 2.13. Let $\tilde{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix and $\tilde{B} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Then the set of generalized eigenvectors $\{\mathbf{x}^k\}_{k=1,\dots,n}$ associated with pencil $(\tilde{A}, \tilde{B})$ can be chosen so that they form a $\tilde{B}$-orthonormal basis of $\mathbb{R}^n$:

$$\langle \mathbf{x}^k, \tilde{B}\mathbf{x}^k \rangle = 1, \text{ for all } k = 1, \dots, n \text{ and } \langle \mathbf{x}^k, \tilde{B}\mathbf{x}^l \rangle = 0, \text{ for all } k, l = 1, \dots, n;\ k \neq l.$$

Then, for every $k = 1, \dots, n$ the following also holds

$$\langle \mathbf{x}^k, \tilde{A}\mathbf{x}^k \rangle = \lambda^k, \text{ and } \langle \mathbf{x}^k, \tilde{A}\mathbf{x}^l \rangle = 0, \text{ for all } l = 1, \dots, n;\ l \neq k.$$

Proof. Most of this proof is a rewrite of [64]. It is a well known result, for any real symmetric matrix $\tilde{M} \in \mathbb{R}^{n\times n}$, that there exists an orthonormal basis $\{\mathbf{y}^k\}_{k=1,\dots,N}$ of $\mathbb{R}^n$ which consists of eigenvectors $\mathbf{y}^k$ of $\tilde{M}$. This can be written as:

$$\tilde{M}\mathbf{y}^k = \lambda^k \mathbf{x}^k; \quad \langle \mathbf{y}^k, \mathbf{y}^k \rangle = 1; \quad \text{and} \quad \langle \mathbf{y}^k, \mathbf{y}^l \rangle = 0, \text{ if } k \neq l. \tag{2.37}$$

The way to prove the lemma is to reduce generalized eigenvalue problem (2.36) to a classical one. First of all notice that writing the problem as $\tilde{B}^{-1}\tilde{A}\mathbf{x}^k = \lambda^k \mathbf{x}^k$ does not help because there is no reason for the product $\tilde{B}^{-1}\tilde{A}$ to be symmetric. Instead we use the fact that, since $\tilde{B}$ is symmetric positive definite, it admits a Cholesky factorization:

$$\tilde{B} = LL^\top; \text{ where } L \text{ is a lower triangular (invertible) matrix.}$$

Then (2.36) wan be rewritten as $\tilde{M}\mathbf{y}^k = \lambda^k \mathbf{y}^k$ for $\tilde{M} = L^{-1}\tilde{A}L^{\top -1}$, $\mathbf{y}^k = L^\top \mathbf{x}^k$. Assume that the $\mathbf{y}^k$ have been chosen in order for (2.37) to hold, then the set of $\mathbf{x}^k = (L^\top)^{-1}\mathbf{y}^k$ is the basis of $\mathbb{R}^n$ that we are looking for since it satisfies the following conditions:

$$\tilde{A}\mathbf{x}^k = \lambda^k \tilde{B}\mathbf{x}^k; \quad \langle \mathbf{x}^k, \tilde{B}\mathbf{x}^k \rangle = 1; \quad \text{and} \quad \langle \mathbf{x}^k, \tilde{B}\mathbf{x}^l \rangle = 0,\ k \neq l.$$

With this the fact that $\langle \mathbf{x}^k, \tilde{A}\mathbf{x}^k \rangle = \lambda^k$ is obvious. For the last property in the lemma, let $(\lambda^k, \mathbf{x}^k)$ and $(\lambda^l, \mathbf{x}^l)$ be two generalized eigenvectors with $k \neq l$. Assume that $\lambda^l \neq 0$ then

$$\left(\langle \mathbf{x}^k, \tilde{B}\mathbf{x}^l \rangle = 0 \text{ and } \tilde{A}\mathbf{x}^l = \lambda^l \tilde{B}\mathbf{x}^l \right) \Rightarrow \frac{1}{\lambda^l} \langle \mathbf{x}^k, \tilde{A}\mathbf{x}^l \rangle = 0 \Rightarrow \langle \mathbf{x}^k, \tilde{A}\mathbf{x}^l \rangle = 0.$$

If $\lambda^l = 0$ then $\mathbf{x}^l \in \mathrm{Ker}(\tilde{A})$ so $\langle \mathbf{x}^k, \tilde{A}\mathbf{x}^l \rangle = 0$ in this case too. □

A direct consequence is the result in the following Lemma which lets us foresee how solving a generalized eigenvalue problem for pencil $(\tilde{A}, \tilde{B})$ will allow us to find the spaces where Assumptions 2.6 and 2.7 hold.

Lemma 2.14. With the notation introduced in Lemma 2.13 and given a threshold $\tau \in \mathbb{R}^+$ let

$$E_1 = \mathrm{span}\{\mathbf{x}^k; \lambda_k < \tau\} \text{ and } E_2 = \mathrm{span}\{\mathbf{x}^k; \lambda_k \geq \tau\}$$

then

$$\begin{cases} \langle \mathbf{x}, \tilde{A}\mathbf{x} \rangle & < & \tau \langle \mathbf{x}^k, \tilde{B}\mathbf{x} \rangle, & \text{for all } \mathbf{x} \in E_1, \\ \langle \mathbf{x}, \tilde{A}\mathbf{x} \rangle & \geq & \tau \langle \mathbf{x}^k, \tilde{B}\mathbf{x} \rangle, & \text{for all } \mathbf{x} \in E_2. \end{cases}$$

State of the art In practice solving a generalized eigenvalue problem which involves the global matrix is more expensive than solving the original problem. For this reason before resorting to a generalized eigenvalue problem to find the space where an estimate holds the robustness condition must first be rewritten in local form giving rise to one (local) generalized eigenvalue problem per subdomain. These can then be solved in parallel. To the best of our knowledge this strategy goes back to [7] where it was used to build an Algebraic Multigrid with aggregation based on elements method for which any targeted convergence rate can be achieved. The ideas behind spectral AMG [12] are also closely related. Since then this has been quite a prolific direction of research. In particular the methods developed in the remainder of this thesis build on these seminal ideas.

More recently, several contributions propose to build coarse spaces for problems with highly heterogeneous coefficients by solving local eigenproblems. However, compared to the earlier work in the AMG context the recent papers all focus on generalized eigenvalue problems. We may distinguish between three sets of methods that all differ by the choice of the bilinear form on the right hand side of the generalized eigenproblem. In [38, 39], the right hand side is the local mass matrix, or a "homogenised" version obtained by using a multiscale partition of unity. In [78, 79, 21] the right hand side corresponds to an L_2-product on the subdomain boundary, so that the problem can be reduced to a generalized eigenproblem for the Dirichlet-to-Neumann operator on the subdomain boundary. This is the method we present in Chapter 3. It applies to the scalar elliptic problem. The latest set of papers, [99, 26, 105, 106, 70, 103, 109], uses yet another type of bilinear form on the right hand side, inspired by theoretical considerations. The construction in this last set of papers is particularly interesting because it extends also to other equations such as Stokes, Brinkmann, linear elasticity, or the eddy current problem. In Chapter 4 we present our contribution to this family of methods for the Schwarz preconditioner and then for BDD and FETI in Chapter 5.

In the framework of two level additive Schwarz methods, [47, 98, 38] identify the bottleneck for proving a convergence bound which is independent of the jumps in the coefficients to be the so called stable splitting property. Bypassing this bottleneck estimate is the objective behind the choice of many of the aforementioned coarse spaces. All these approaches have their advantages and disadvantages, which depend on many factors, in particular the type of coefficient variations and the size of the overlap. The choice of generalized eigenvalue problem is a delicate compromise between ensuring robustness and a moderate size of the coarse space. In this spirit, for the scalar elliptic equation, [39, 26] use multiscale partition of unity functions to eliminate some of the 'bad' eigenmodes a priori. While very effective in the scalar elliptic case, this may prove tricky in cases where there are several PDEs with several jumping coefficients. With the methods in [79, 21] (see also Chapter 3 of this manuscript) for the scalar elliptic problem the eigenvalue problem is posed on the interface and this alone resolves many of the complications posed by coefficient variations that trigger non necessary coarse basis vectors. Throughout all of our numerical experiments we will keep a close eye on the size of the coarse space that is constructed automatically.

We mention that there have also been some recent multilevel extensions of some of the above approaches (see [28, 27, 116, 100]).

2.3 Contributions of this Thesis

The subject of this thesis was elaborated in cooperation with the tire manufacturer Michelin following the work [78]. For a solver in an industrial context robustness is among the most important properties: it must be guaranteed that when a simulation is run it will converge. For this reason the strategy developed in this thesis takes place at the most algebraic level possible: hardly any assumptions are made on the symmetric positive definite problem at hand. This way we are ready to face a wide range of difficulties (in particular heterogeneous coefficients).

Chapter 3 of this thesis focuses on the scalar problems of the type $-\nabla\cdot(\alpha\nabla u) = f$. The coarse space, for Schwarz, is constructed using the low frequency modes of the Dirichlet-to-Neumann operator defined on the boundary of each subdomain. Chapter 4 concentrates once more on the Schwarz preconditioner. We propose and analyze a coarse space which applies to systems of partial differential equations. This is the first coarse space which we call GenEO for Generalized Eigenproblems in the Overlaps. Then, in Chapter 5 we propose to apply the GenEO strategy to build robust FETI and BDD methods. Although the starting point is the same, its application is quite different. In each case we will prove convergence bounds that don't depend on the specific difficulties of the problem and illustrate these results on numerical examples. Chapter 6 describes the behaviour of the GenEO algorithms on elasticity problems in the almost incompressible limit. Finally in Chapter 7 we propose some leads for future work in particular with the objective to improve the GenEO coarse spaces.

The fundamental idea on which this entire work is based is that within an iterative solver, by using well chosen projections, we can separate the problem into two parts: the first is solved with the iterative solver and a specific treatment is applied to the second (a direct solve). In the remainder of this manuscript the main objective will be to identify the part of the solution space on which the iterative solver does a good job. The complementary of this space is responsible for slow convergence and we make it into the coarse space so that it is taken care of by a direct solver.

We are looking for a (not too large) coarse space which is

– spanned by local vectors (so that the coarse problem matrix is sparse),
– computed automatically,
– such that the corresponding two level method is robust.

The strategy for the choice of the coarse space is the same throughout the manuscript: using the Abstract Schwarz framework the bottleneck estimate in the convergence proof is derived. Based on this a generalized eigenvalue problem is identified which separates the vectors that are problematic and need to be in the coarse space from the others. This way we can guarantee convergence theoretically.

2.3.1 DtN: a coarse space for the scalar elliptic problem

We present here Chapter 3 in which we focus on the scalar elliptic problem. We build on the work in [78] which precedes the beginning of this thesis. The contents of Chapter 3 are a rewrite of articles [79, 21, 52]. More precisely we give the heuristics motivating this particular choice for the coarse space as well as a theoretical analysis that guarantees robustness and some numerical results.

Given a right hand side f, the scalar elliptic problem is: Find u^* such that

$$-\nabla\cdot(\alpha\nabla u^*) = f,$$

where $\alpha : \Omega \to \mathbb{R}^+$ is a coefficient which varies within the domain. Since for now we just want to introduce some ideas we don't give any extra thought to the boundary conditions for the global problem.

Heuristics Consider the case where the domain is divided into slices (all boundaries between subdomains are in one direction only). In Figure 2.7 we present this geometry for the case of three subdomains. Then we apply the alternating Schwarz algorithm (2.2) to this problem. It is easy to prove that the updates of the error $e^n = |u^n - u^*|$ satisfy the same algorithm but for the homogeneous problem. In particular each update of the error e_2 in subdomain Ω_2 satisfies (using notation from Figure 2.7):

$$\begin{aligned} -\nabla \cdot (\alpha \nabla e_2^{n+1}) &= 0, \\ e_2^{n+1}(A, y) &= e_1^{n+1}(A, y), \\ e_2^{n+1}(D, y) &= e_3^n(D, y). \end{aligned} \tag{2.38}$$

Figure 2.8 shows the successive updates of the error in Ω_2 and its neighbours. The general behaviour at each step is known since, by the maximal principal, e_2^{n+1} decreases in the interior of the subdomain with boundary conditions at $x = A$ and $x = D$ which are prescribed by the neighbours. As shown in the figure, convergence can be either fast or slow depending on whether the solutions of the local problems decrease rapidly or slowly in the overlap. (We recall here that since we are considering the error, our objective is to drive it to zero.) It is on this observation that the construction of the DtN coarse space relies: in each subdomain we want to isolate the components of the error which, for a given Dirichlet condition, decrease slowly inside the subdomain and thus pass on to their neighbours a boundary condition that is barely improved.

Under the assumption that the overlap is narrow a reasonable guess is that when the normal derivative of the error at the boundary of the subdomain is large then the boundary condition which is passed on to the neighbours is closer to zero.

Definition of the DtN coarse space The Dirichlet-to-Neumann operator evaluates just the desired quantity. Indeed for a domain Ω_j it is defined as follows.

Definition 2.15. Let $\mathrm{tr}_j\alpha$ be the trace of coefficient α in Ω_j on the boundary $\Gamma := \partial\Omega_j$ of subdomain Ω_j and $\boldsymbol{n}_j$ be the unit outward normal to Ω_j on Γ. For any function $v_\Gamma : \Gamma \to \mathbb{R}$ such that $v_{\Gamma|\partial\Omega} = 0$ if $\Gamma \cap \partial\Omega \neq \emptyset$ we define

$$\mathrm{DtN}_j(v_\Gamma) := \mathrm{tr}_j\alpha \frac{\partial v}{\partial \boldsymbol{n}_j}\bigg|_\Gamma, \text{ where } v \text{ is the solution of } \begin{cases} -\nabla \cdot (\alpha \nabla v) = 0 & \text{in } \Omega_j \\ v = v_\Gamma & \text{on } \Gamma \end{cases}. \tag{2.39}$$

In other words, given a function defined on Γ, the DtN operator extends it harmonically to the interior of the subdomain and returns the normal derivative of the extension at the boundary.

In order to fulfill this definition, the procedure for building the coarse space is the following:

1. Compute (in parallel over the subdomains) the generalized eigenvalues λ and generalized eigenvectors v_Γ of
$$\mathrm{DtN}_j(v_\Gamma) = \lambda\, \mathrm{tr}_j\alpha\, v_\Gamma.$$
2. Select the eigenvectors that correspond to an eigenvalue smaller than $1/\mathrm{diam}(\Omega_j)$ (the inverse of the diameter of the subdomain).

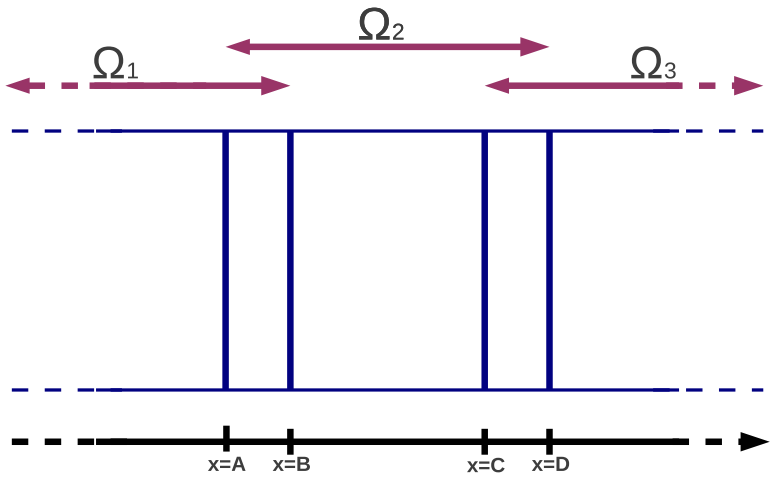


Figure 2.7: Geometry for motivating the choice of the DtN coarse space (see also Figure 2.8).

3. Extend these eigenvectors harmonically to the interior of the subdomain.
4. Multiply by a partition of unity function, reinterpolate into the finite element space and extend by zero to the whole of Ω.

Theoretical Result By building the coarse space this way we can ensure that the resulting two level Schwarz method will converge independently of almost all parameters in the problem.

Theorem 2.16. Under an assumption on α the condition number of A preconditioned by two level Additive Schwarz with the DtN coarse space satisfies

$$\kappa(M_{AS,2}^{-1}A) \lesssim \left(C_P^2 + \max_{j=1}^{N} \frac{\text{diam}(\Omega_j)}{\delta_j}\right).$$

The constant hidden in the symbol $\lesssim$ doesn't depend on the mesh size h, the overlap size δ_j, the size of the subdomain $\text{diam}(\Omega_j)$ or on the choice of α. More detail on the constant C_P is given in the caption of Figure 2.9.

The reason for the assumption on the coefficient distribution is that the proof requires applying weighted Poincaré inequalities [89]. It is not very restrictive and always satisfied in the case of minimal overlap (*i.e.* the size δ_j of the overlap is equal to the mesh size). The case where the inequality does not apply is exactly the case where, because of α, the value of the normal derivative on the boundary of the subdomain is not correlated to the decay of the error in the overlap and thus, even by controlling the normal derivative, it cannot be guaranteed that a good transmission condition will be given to the neighbouring subdomain.

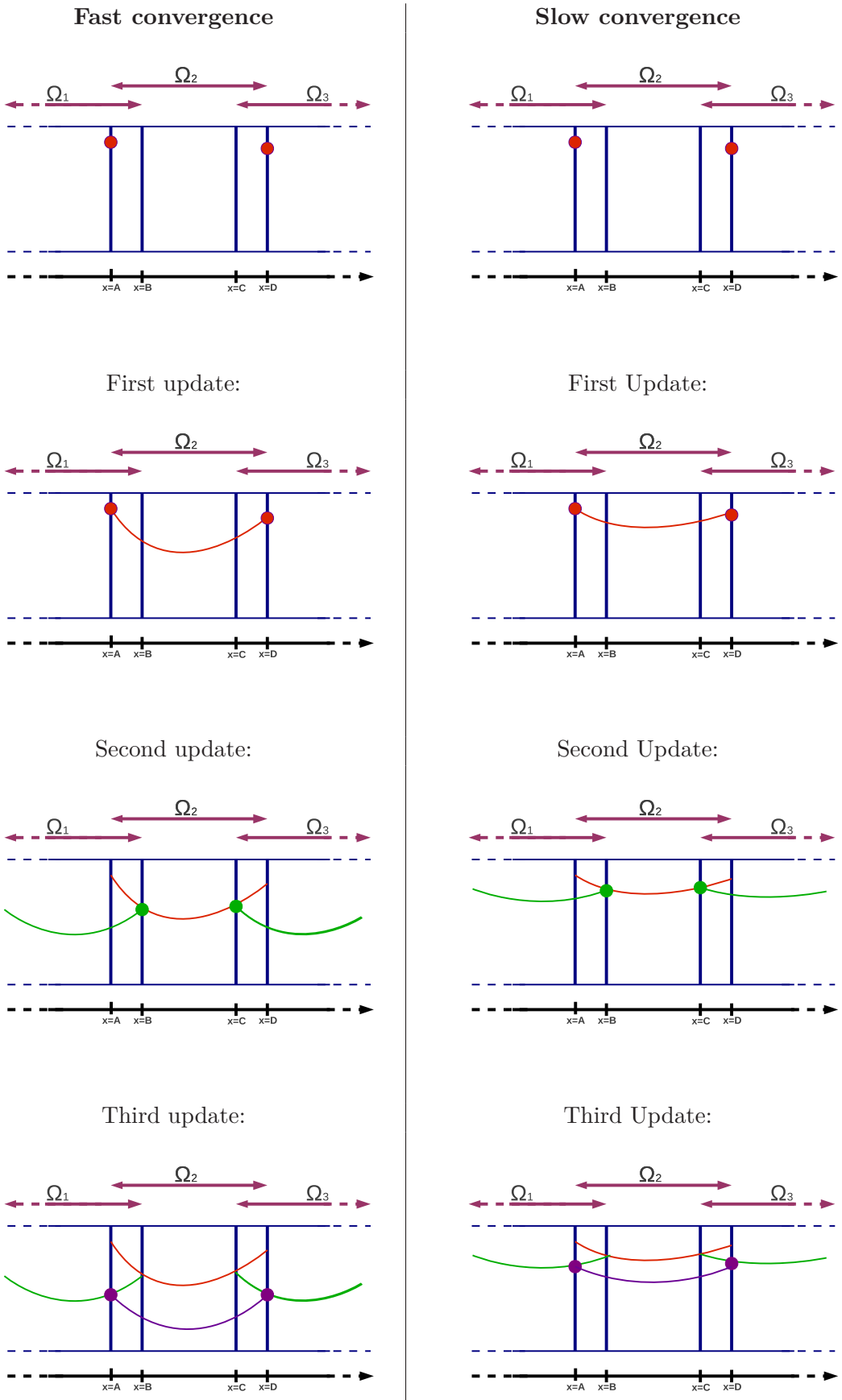


Figure 2.8: Illustration of the fact that the DtN operator can predict the speed of convergence. (Since we are looking at the updates of the error the objective is to drive it to zero as fast as possible.) We notice that if the local component of the error decreases rapidly in the overlap then we give to the neighbouring subdomain a *good* boundary value.

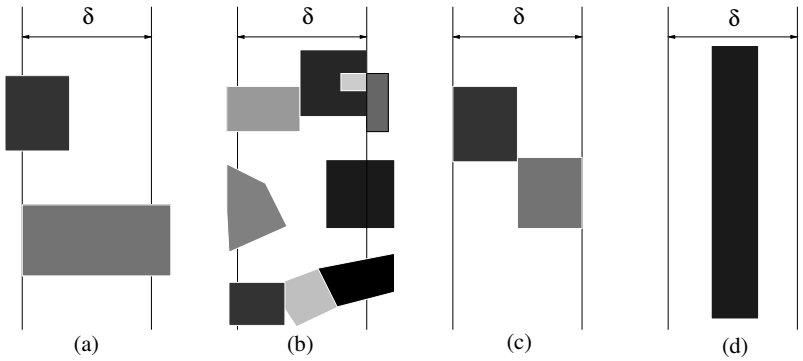


Figure 2.9: Overlap between two subdomains (darker shades of gray correspond to larger values of α). In the two first cases (starting from the left) the assumption holds and $C_P = \mathcal{O}(1)$ in the theorem (a), (b); in the third case the assumption holds and $C_P = \mathcal{O}(\log(\delta_j/h))$ in the theorem (c); in the last case the assumption does not hold (d).

As usual with the Schwarz preconditioner and as we have already stated (see Remark 2.11) most of the proof consists in proving the existence of a stable splitting of any global vector onto the local subspaces and the coarse space (Assumption 2.7).

Numerical Results Now we illustrate the efficiency of the DtN coarse space. Table 2.3 presents the results of the robustness test which we already put the one level Schwarz preconditioner and the two level Schwarz preconditioner with the Partition of unity coarse space through. We observe that this time robustness is achieved. The number of vectors that are selected for the coarse space in each subdomain is equal to the number of layers where α is large (three in our case). This is the optimal choice (see [115, 47] or the discussion in Subsection 1.2.3). In Chapter 3 we present a more elaborate series of tests.

2.3.2 GenEO: a coarse space for the Additive Schwarz method

In Chapter 4 we build a coarse space which allows us to guarantee convergence for a larger range of symmetric positive definite matrices. This work is the object of publications [105, 106].

Ideas We explain here the ideas which led to the construction of the GenEO coarse space and refer to Chapter 4 for a rigorous presentation. The proof of convergence for the DtN coarse space relies on two arguments that cannot be generalized easily to the case of systems of partial differential equations:

- Weighted Poincaré inequalities make it possible to derive estimates between a norm on the boundary of the subdomain (which we know to be bounded as a result of the generalized eigenvalue problem) and a norm in the overlap (which we need for the proof),
- a stability property for the finite element interpolator (with respect to the Euclidean norm and also the operator norm) is required because each time a function is multiplied by the partition of unity, the result $\chi_j u_j$ must be interpolated into the finite element space.

The generalized eigenvalue problem which we have elaborated incorporates both of these obstacles directly into its definition. Indeed one of the bilinear forms in the pencil is defined on the overlapping zone between two subdomains and it is weighted by the

Iteration count:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	18	25	25	25
$\alpha_2 = 10^2$	21	26	27	26
$\alpha_2 = 10^4$	22	28	28	26
$\alpha_2 = 10^6$	17	25	25	25

Estimated Condition number:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	22.4	25.3	26.1	26.2
$\alpha_2 = 10^2$	22.4	25.3	26.1	26.2
$\alpha_2 = 10^4$	22.4	25.3	26.1	26.2
$\alpha_2 = 10^6$	22.4	25.3	26.0	26.2

Table 2.3: Convergence results for the scalar elliptic problem (2.19) discretized by $\mathbb{P}_1$ finite elements and preconditioned by two level Additive Schwarz (2.32) where V_0 is the **DtN** coarse space. The geometry is given in Figure 2.5. Two layers of overlap are added to each subdomain. The coefficient in material 1 is always $\alpha_1 = 1$. We make the number of subdomains and the jump in the coefficient vary through α_2 (coefficient in material 2) and report the number of iterations needed to reach convergence (top) and the estimate for the condition number of the preconditioned operator based on the Ritz values (bottom).

partition of unity operator. This way the estimate that follows from the definition of the eigenproblem and Lemma 2.13 bypasses the two arguments in the proof for the DtN coarse space that we were unable to generalize.

Definition of the GenEO coarse space The definition of this coarse space is also presented in a poster in the Appendix of this thesis. The poster was presented at the conference *Special Semester on Multiscale Simulation and Analysis in Energy and the Environment* at RICAM in Linz (Austria) in November 2011.

We have called the coarse space GenEO for "Generalized Eigenvalues in the Overlaps". These letters also appear in the word HeteroGenEOus which is an amusing coincidence. In order to define the GenEO coarse space we need to introduce some elements of notation which are of course defined rigorously in Chapter 4:

- For every domain $D \subset \Omega$ that is resolved by the mesh, $V_h(D)$ is the set of restrictions to D of all finite element functions,
- for every domain $D \subset \Omega$ that is resolved by the mesh, $V_{h,0}(D)$ is the set of restrictions to D of all finite element functions that are supported in $\bar{D}$,
- for every domain $D \subset \Omega$ that is resolved by the mesh, the bilinear form $a_D : D \times D \to \mathbb{R}^+$ is obtained by assembling all element matrices corresponding to the elements in D,
- Ω_j° is the part of Ω_j that is overlapped by at least one other subdomain,
- for $j = 1, \ldots, N$; $\Xi_j : V_h(\Omega_j) \to V_{h,0}(\Omega_j)$ is a family of functions that form a partition of unity subordinate to the partition into subdomains. Notice that these functions directly return a finite element function so there is no need for interpolation after applying the partition of unity.

Definition 2.17 (GenEO Coarse Space)**.** For each subdomain $j = 1, \ldots, N$, solve the following generalized eigenvalue problem: find (λ, p) such that

$$a_{\Omega_j}(p, v) = \lambda \, a_{\Omega_j^\circ}(\Xi_j(p), \Xi_j(v)), \quad \forall \, v \in V_h(\Omega_j). \tag{2.40}$$

For each $j = 1, \ldots, N$, let $(p_j^k)_{k=1}^{m_j}$ be a set of eigenvectors of (2.40) that correspond to the m_j lowest eigenvalues. The coarse space is defined as

$$V_H := \text{Vect}\{R_j^\top \Xi_j(p_j^k) : k = 1, \ldots, m_j;\ j = 1, \ldots, N\}.$$

Theoretical Result With this coarse space the two level Additive Schwarz method converges independently of the number of subdomains and the parameters in the problem as is guaranteed by the following theorem.

Theorem 2.18. The condition number of matrix A preconditioned by two level Schwarz with the GenEO coarse space is bounded by

$$\kappa(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \ \leq \ (1+k_0)\left[2 + k_0(2k_0+1)\max_{1\leq j\leq N}\left(1+\frac{1}{\lambda_j^{m_j+1}}\right)\right],$$

where the constant k_0 depends only on the geometry of the problem (but not on the number of subdomains) and is defined as in Lemma 2.9.

We refer to Chapter 4 for the exact assumptions under which this result applies. They are not very restrictive. We notice that perhaps the most important quantity in this estimate is $\lambda_j^{m_j+1}$, the smallest eigenvalue not to have been selected for the coarse space. One possibility is to use the test $\lambda_j^k < \delta_j/\text{diam}(\Omega_j)$ to select the vectors for the coarse space. Then the quantity in the estimate is $\text{diam}(\Omega_j)/\delta_j$ as for the DtN coarse space:

$$\kappa(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \ \leq \ (1+k_0)\left[2 + k_0(2k_0+1)\max_{1\leq j\leq N}\left(1+\frac{\text{diam}(\Omega_j)}{\delta_j}\right)\right],$$

.

Numerical results With this choice of the selection process we submit GenEO to the same robustness test as DtN and before that the Partition of unity coarse space and the one level preconditioner. As for DtN as many modes as there are layers where α is large (three) are selected and, thanks to the choice of the coarse space, the two level solver is robust as testified by the results in Table 2.4.

2.3.3 Generalization of GenEO to substructuring methods

We give here a brief summary of the contributions presented in Chapter 5. They are the subject of publication [109] as well as the shorter note [108].

Once more the idea is to rely heavily on the Abstract Schwarz framework to find which estimate in the convergence proof of FETI and BDD requires strong assumptions. The main difference with GenEO for Schwarz is that this time it is the stability of the local solvers (Assumption 2.6) which will determine the choice of the eigenvalue problem while the existence of a stable splitting (Assumption 2.7) is trivial on the entire solution space thanks to the presence of partition of unity operators in the preconditioners.

The major part of the theoretical study consists in reformulating Assumption 2.6 to find the right eigenvalue problem. Once it has been identified, and given a threshold τ, the coarse space is constructed by selecting all the generalized eigenvectors that correspond to an eigenvalue smaller than τ. Thanks to this we are able to prove that the condition numbers of the two level preconditioned BDD and FETI operators are bounded by $\frac{\mathcal{N}}{\tau}$ where $\mathcal{N}$ denotes the maximal number of neighbours of a subdomain.

Iteration count:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	19	24	25	24
$\alpha_2 = 10^2$	23	26	27	26
$\alpha_2 = 10^4$	26	26	27	27
$\alpha_2 = 10^6$	17	21	22	25

Estimated Condition number:

	8 subdomains	16 subdomains	32 subdomains	64 subdomains
$\alpha_2 = 1$	31.8	31.9	31.9	31.9
$\alpha_2 = 10^2$	31.8	31.9	31.9	31.9
$\alpha_2 = 10^4$	31.8	31.9	31.9	31.9
$\alpha_2 = 10^6$	31.8	31.9	31.9	31.9

Table 2.4: Convergence results for the scalar elliptic problem (2.19) discretized by $\mathbb{P}_1$ finite elements and preconditioned by two level Additive Schwarz (2.32) where V_0 is the **GenEO** coarse space. The geometry is given in Figure 2.5. Two layers of overlap are added to each subdomain. The coefficient in material 1 is always $\alpha_1 = 1$. We make the number of subdomains and the jump in the coefficient vary through α_2 (coefficient in material 2) and report the number of iterations needed to reach convergence (top) and the estimate for the condition number of the preconditioned operator based on the Ritz values (bottom).

For BDD the eigenvalue problem stems quite naturally from the formulation of BDD in the abstract Schwarz framework. For FETI the procedure is more complex since it is on the transpose FM^{-1} of the preconditioned operator (which has the same spectrum as $M^{-1}F$) that we have worked. The result can be written not only for the Dirichlet preconditioner that we've already introduced but also for the, cheaper, Lumped preconditioner.

At the end of Chapter 5 numerical results illustrate the behaviour of the method for FETI.

2.3.4 Application to elasticity in the almost incompressible limit

The results obtained with the Schwarz preconditioner and the GenEO coarse space are very satisfying for many problems. Unfortunately there remained one obstacle: the Schwarz preconditioner is so poor for elasticity problems in the almost incompressible limit that our automatic selection procedure leads to a coarse space that spans the whole range of displacements in the overlapping zone. Since our methods must apply to the Michelin simulations we could not ignore this: tires are essentially made of rubber which is the textbook example for an almost incompressible material.

In Chapter 6 we show numerically that contrary to Schwarz-GenEO, FETI-GenEO deals fine with almost incompressible problems. Here we explain how we got the intuition that it was necessary to change solver altogether.

Fourier analysis for the Schwarz Method We consider the case where $\Omega = \mathbb{R}^2$ and the domain is partitioned into two half planes $\Omega_1 = \{(x, y);\ x < \delta\}$ and $\Omega_2 = \{(x, y);\ x > 0\}$. The width of the overlap is $\delta > 0$.

In two dimensions the (expanded) linear elasticity equations with constant coefficients write for the vector unknown $\boldsymbol{u} = (u, v)^{\mathrm{T}}$ and right hand side $\mathbf{f} = (f_1, f_2)^{\top}$ as follows

$$\begin{cases} -\mu\Delta u - (\lambda + \mu)\partial_x \nabla \cdot (\boldsymbol{u}) = f_1, \\ -\mu\Delta v - (\lambda + \mu)\partial_y \nabla \cdot (\boldsymbol{u}) = f_2, \end{cases} \tag{2.41}$$

where λ and μ can be defined with respect to the Lamé parameters

$$\lambda := \frac{E\nu}{(1+\nu)(1-2\nu)},\ \mu := \frac{E}{2(1+\nu)}.$$

Given the particularly simple geometry of the domain we may apply a Fourier transform in the y direction which allows us to write the problem as:

$$\begin{cases} -(2\mu+\lambda)\partial_{xx}\hat{u} + k^2\mu\hat{u} - ik(\lambda+\mu)\partial_x\hat{v} = \hat{f}_1, \\ k^2(2\mu+\lambda)\hat{v} - \mu\partial_{xx}\hat{v} - ik(\lambda+\mu)\partial_x\hat{u} = \hat{f}_2. \end{cases} \tag{2.42}$$

For a given $k > 0$, this is an ordinary differential equation which solution (obtained *via* Maple) is

$$\begin{cases} \hat{u}(x) = & a_1\,\mathrm{e}^{-kx} + a_2\,\mathrm{e}^{-kx}x + a_3\,\mathrm{e}^{kx} + a_4\,\mathrm{e}^{kx}x, \\ \hat{v}(x) = & \frac{-i\left(\mu\,a_1\,k\mathrm{e}^{-kx} + \mu\,a_2\,k\mathrm{e}^{-kx}x - 3\,\mu\,a_2\,\mathrm{e}^{-kx} - \mu\,a_3\,k\mathrm{e}^{kx} - \mu\,a_4\,k\mathrm{e}^{kx}x - 3\,\mu\,a_4\,\mathrm{e}^{kx}\right)}{k(\mu+\lambda)} \\ + & \frac{-i\left(k\lambda\,a_1\,\mathrm{e}^{-kx} + k\lambda\,a_2\,\mathrm{e}^{-kx}x - \lambda\,a_2\,\mathrm{e}^{-kx} - k\lambda\,a_3\,\mathrm{e}^{kx} - k\lambda\,a_4\,\mathrm{e}^{kx}x - \lambda\,\mathrm{e}^{kx}a_4\right)}{k(\mu+\lambda)}, \end{cases} \tag{2.43}$$

with a_1, a_2, a_3 and a_4 (complex) integration constants.

Denote by $(\hat{u}_1, \hat{v}_1)$ the solution of this problem restricted to subdomain Ω_1 ($x < \delta$) and by $(\hat{u}_2, \hat{v}_2)$ the solution of this problem restricted to subdomain Ω_2 ($x > 0$). By a classical argument we know that $\hat{u}_1$ and $\hat{v}_1$ must be bounded at $-\infty$ and that $\hat{u}_2$ and $\hat{v}_2$ must be bounded at $+\infty$. With this we can eliminate half of the terms in (2.43):

$$\begin{cases} \hat{u}_1 = (a_1 + b_1 x)\mathrm{e}^{kx}, \\ \hat{v}_1 = \frac{i\,(a_1\,\mu k + b_1\,\mu k x + 3b_1\,\mu + a_1\,\lambda k + b_1\,\lambda k x + b_1\,\lambda)\mathrm{e}^{kx}}{k(\mu+\lambda)}, \\ \hat{u}_2 = (a_2 + b_2 x)\mathrm{e}^{kx}, \\ \hat{v}_2 = \frac{i\,(a_2\,\mu k + b_2\,\mu k x + 3b_2\,\mu + a_2\,\lambda k + b_2\,\lambda k x + b_2\,\lambda)\mathrm{e}^{kx}}{k(\mu+\lambda)}. \end{cases} \tag{2.44}$$

Comparison between four types of transmission conditions In order to find the solution of the global problem we use the alternating Schwarz algorithm (2.2): the problem is solved alternately in each of the subdomains using values of the local solution that was just computed by the neighbour as boundary conditions. We generalize the classical algorithm (2.2) by considering four different types of transmission conditions: (A) continuity of the displacements (normal and tangential), (B) continuity of the constraints (normal and tangential), (C) continuity of the normal constraint and tangential displacement, (D) continuity of the tangential constraint and normal displacement.

$$\text{(A)}\ \begin{cases} u_1^{n+1}(\delta, y) = u_2^n(\delta, y), \\ v_1^{n+1}(\delta, y) = v_2^n(\delta, y), \\ u_2^{n+1}(0, y) = u_1^n(0, y), \\ v_2^{n+1}(0, y) = v_1^n(0, y). \end{cases} \qquad \text{(B)}\ \begin{cases} {\sigma_1}_n^{n+1}(\delta, y) = {\sigma_2}_n^n(\delta, y), \\ {\sigma_1}_t^{n+1}(\delta, y) = {\sigma_2}_t^n(\delta, y), \\ {\sigma_2}_n^{n+1}(0, y) = {\sigma_1}_n^n(0, y), \\ {\sigma_2}_t^{n+1}(0, y) = {\sigma_1}_t^n(0, y). \end{cases}$$

$$\text{(C)}\ \begin{cases} v_1^{n+1}(\delta, y) = v_2^n(\delta, y), \\ {\sigma_1}_n^{n+1}(\delta, y) = {\sigma_2}_n^n(\delta, y), \\ v_2^{n+1}(0, y) = v_1^n(0, y), \\ {\sigma_2}_n^{n+1}(0, y) = {\sigma_1}_n^n(0, y). \end{cases} \qquad \text{(D)}\ \begin{cases} u_1^{n+1}(\delta, y) = u_2^n(\delta, y), \\ {\sigma_1}_t^{n+1}(\delta, y) = {\sigma_2}_t^n(\delta, y), \\ u_2^{n+1}(0, y) = u_1^n(0, y), \\ {\sigma_2}_t^{n+1}(0, y) = {\sigma_1}_t^n(0, y). \end{cases}$$

where:

$$\sigma_n = (2\mu+\lambda)\frac{\partial u}{\partial x} + \lambda\frac{\partial v}{\partial y},\ \text{and}\ \sigma_t = \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right),$$

are respectively the normal and tangential components of the interface force.

Following a Fourier transform in the y direction the interface forces write

$$\hat{\sigma_n} = (2\mu + \lambda)\frac{\partial \hat{u}}{\partial x} + ik\lambda\hat{v}, \text{ and } \hat{\sigma_t} = \mu\left(\frac{\partial \hat{v}}{\partial x} + ik\hat{u}\right).$$

By the same argument as for the heuristic analysis of DtN, the updates of the error satisfy the same equations as the iterative scheme but for the homogeneous problem. For this reason in what follows a *good* solving scheme is one which converges to zero rapidly.

With well chosen changes of variables and the use of Maple we can describe the values of the constants a_1 and b_1 through an iteration matrix that, when applied to the values of a_1 and b_1 at iteration $n-1$, returns the values of a_1 and b_1 at iteration $n+1$:

$$\begin{pmatrix} a_1^{n+1} \\ b_1^{n+1} \end{pmatrix} = M \begin{pmatrix} a_1^{n-1} \\ b_1^{n-1} \end{pmatrix}. \tag{2.45}$$

Since Ω_1 and Ω_2 are interchangeable the coefficients for Ω_2 also satisfy this equation. Of course M depends on the choice of the transmission conditions. In the case where they are mixed ((C) or (D)) the iteration matrix takes a simple form:

$$M_{C \text{ or } D} = \begin{pmatrix} \mathrm{e}^{-2\,k\delta} & -2\,\delta\,\mathrm{e}^{-2\,k\delta} \\ 0 & \mathrm{e}^{-2\,k\delta} \end{pmatrix}$$

In the two other cases the matrices are a lot more complicated. For this reason, from now on we focus on their two eigenvalues eig_1 and eig_2 indexed by A, B, C or D keeping in mind that convergence is good when these eigenvalues are close to zero and it is bad when they are close to 1. We get

$$\begin{cases} eig_{A1} &= \left[1 + 2\frac{(\delta k)^2}{(3-4\nu)^2} + 2\sqrt{\frac{(\delta k)^2}{(3-4\nu)^2} + \frac{(\delta k)^4}{(3-4\nu)^4}}\right]\mathrm{e}^{-2\,k\delta}, \\ eig_{A2} &= \left[1 + 2\frac{(\delta k)^2}{(3-4\nu)^2} - 2\sqrt{\frac{(\delta k)^2}{(3-4\nu)^2} + \frac{(\delta k)^4}{(3-4\nu)^4}}\right]\mathrm{e}^{-2\,k\delta}. \end{cases} \tag{2.46}$$

$$\begin{cases} eig_{B1} = \left(1 + 2\,\delta^2k^2 + 2\,\sqrt{\delta^2k^2 + \delta^4k^4}\right)\mathrm{e}^{-2\,k\delta}, \\ eig_{B2} = \left(1 + 2\,\delta^2k^2 - 2\,\sqrt{\delta^2k^2 + \delta^4k^4}\right)\mathrm{e}^{-2\,k\delta}. \end{cases} \tag{2.47}$$

$$\left\{ eig_{C1} = eig_{C2} = eig_{D1} = eig_{D2} = \mathrm{e}^{-2\,k\delta} := eig_C := eig_D. \right. \tag{2.48}$$

The following remarks are straightforward:

–

$$eig_{A1} > eig_C = eig_D > eig_{A2}, \tag{2.49}$$

and

$$eig_{B1} > eig_C = eig_D > eig_{B2}. \tag{2.50}$$

– $eig_{B1}, eig_{B2}, eig_C = eig_D$ do not depend on the physical parameters λ and μ. They do however depend on the size of the overlap δ and the frequency k.
– In all four cases if there is no overlap then the algorithm will not converge and conversely if there is overlap then the eigenvalues are all < 1 and convergence is guaranteed.
– The convergence is at its worse (eigenvalues are close to 1) for low frequencies. This is what is expected with a primal domain decomposition method.

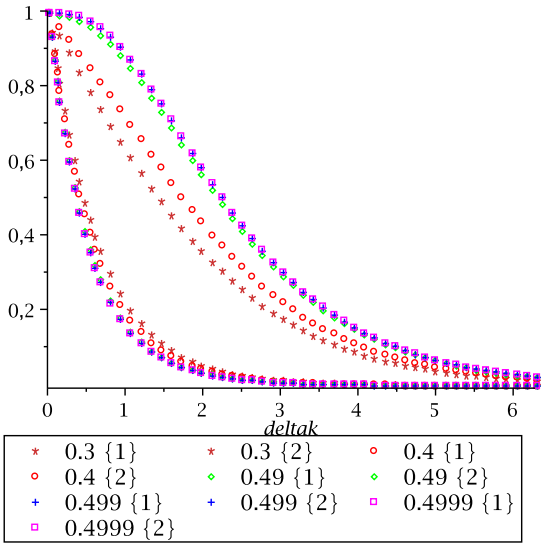


Figure 2.10: Case (A): eigenvalues with respect to δk for different values of Poisson's ratio ν.

Study of the incompressible limit The material parameters only have an impact in case (A) when the transmission conditions are pure displacement. In this case the almost incompressible limit is

$$\begin{cases} \lim\limits_{\nu\to 0.5} eig_{A1} &= \left[1+2(\delta k)^2+2\sqrt{(\delta k)^2+(\delta k)^4}\right] \mathrm{e}^{-2\,k\delta}, \\ \lim\limits_{\nu\to 0.5} eig_{A2} &= \left[1+2(\delta k)^2-2\sqrt{(\delta k)^2+(\delta k)^4}\right] \mathrm{e}^{-2\,k\delta}. \end{cases} \tag{2.51}$$

We observe that $\lim\limits_{\nu\to 0.5} eig_{A1} = eig_{B1}$ and $\lim\limits_{\nu\to 0.5} eig_{A2} = eig_{B2}$. For this reason the fact that with choice (B) the eigenvalues do not depend on the Lamé parameters is actually not an advantage: with these pure constraint transmission conditions convergence is always worse than with pure displacement. In Figure 2.10 we plot the two eigenvalues in case (A) with respect to δk for different values of Poisson's ratio. We indeed observe a convergence behaviour when ν approaches 0.5. Finally, in Figure 2.11 we plot the largest of the eigenvalues (also the worst) with respect to Poisson's ratio ν for different values of δk. We focus on values $\delta k \leq 1$ because the problem that we solve is discretized and so the functions cannot 'oscillate' faster than the mesh: the field of frequencies is restrained to $k \leq \frac{1}{h}$. If moreover the overlap is minimal ($\delta = h$) then we indeed find $\delta k \leq 1$.

The conclusion that we draw from this study is that Dirichlet boundary conditions such as the ones implemented in the Schwarz algorithm are not well adapted to the elasticity problem in the almost incompressible limit. The Fourier analysis suggests that using mixed boundary conditions (one displacement component and one constraint displacement) would lead to better performances in the almost incompressible limit. This has already been observed [44, 45, 82, 20]. We have not continued down this path since it seemed difficult to take advantage of mixed boundary conditions while acting at the algebraic level. For this reason we focused on the substructuring (without overlap) methods that are BDD and FETI.

We find in [76] an additional argument in favor of the substructuring formulation. The results in [14] show that for the geometry that we considered ($\mathbb{R}^2$ divided into two half planes), if the substructuring formulation is used with a Richardson iteration then we get

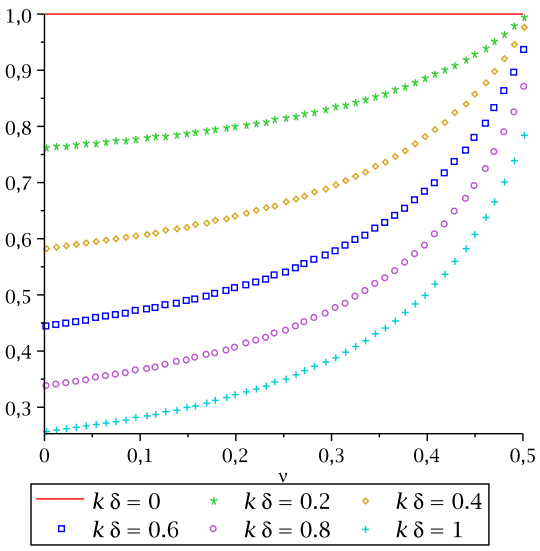


Figure 2.11: Case (A): largest eigenvalue with respect to Poisson's ratio ν for different values of δk.

an exact solver for the Poisson problem [14]: we find the solution in just one iteration. In the case of the Stokes equation (that is very strongly connected to the mixed formulation of nearly incompressible elasticity) the authors in [76] propose an exact solver and show that the classical BDD formulation leads to an algorithm that converges independently of the Lamé parameters.

In Chapter 6 we mention some problems related to the discretization of the almost incompressible elasticity equations but mostly we illustrate the behaviour of our generalized eigenvalue problems on them. In particular we show that with FETI-GenEO one coarse vector per subdomain is sufficient to deal with the almost incompressible behaviour and so our objective is achieved.

2.3.5 Perspectives

To end this manuscript but prepare for the future we present in Chapter 7 some ideas to improve our algorithm or explore their behaviour some more. There are three main directions of research that we mention. First we present how, thanks to the rather abstract formulation that we use, it is possible to extend the GenEO coarse spaces to multilevel methods. This is an important feature in case the GenEO coarse space becomes excessively large. Then we present a different way to build the coarse space where the coarse vectors are no longer selected *a priori* with an eigenproblem but rather selected on the fly within the conjugate gradient iterations. We call this method Frugal FETI because the idea is to use computational resources frugally. Finally, we present a first result for Frugal FETI on a Michelin test case. This result was obtained with the implementation of Frugal FETI at Michelin that is also a part of this thesis work.

Chapter 3

DtN: a coarse space for the scalar elliptic problem

The contents of this chapter were presented in Section 2.3.1 of the introduction. We have merged the following published work into this chapter:

- [79] in collaboration with Frédéric Nataf, Hua Xiang and Victorita Dolean published in *SIAM Journal on Scientific Computing*.
- [21] in collaboration with Victorita Dolean, Frédéric Nataf and Robert Scheichl published in *Computational Methods in Applied Mathematics*.
- [52] in collaboration with Pierre Jolivet, Victorita Dolean, Frédéric Hecht, Frédéric Nataf, and Christophe Prud'Homme published in *Journal of Numerical Mathematics*.

The idea which we build upon was introduced in [78] (Compte rendu à l'Académie des Sciences) by Frédéric Nataf, Hua Xiang and Victorita Dolean.

Contents

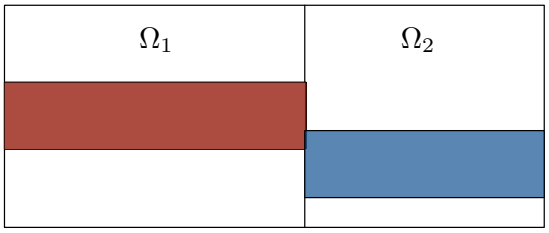


Figure 3.1: Coefficient α varying along and across the interface.

3.1 Introduction

We consider the variational formulation of a second order elliptic boundary value problem with Dirichlet boundary conditions: Find $u^* \in H_0^1(\Omega)$, for a given polygonal (polyhedral) domain $\Omega \subset \mathbb{R}^d (d = 2 \text{ or } 3)$ and a source term $f \in L_2(\Omega)$, such that

$$\underbrace{\int_\Omega \alpha(\boldsymbol{x})\, \nabla u^*(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) d\boldsymbol{x}}_{:= a(u^*,\, v)} = \underbrace{\int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) d\boldsymbol{x}}_{:= \langle f,\, v\rangle}, \quad \text{for all} \quad v \in H_0^1(\Omega). \tag{3.1}$$

In the following we often omit the arguments of the functions we integrate and $d\boldsymbol{x}$. We are interested in the case where the diffusion coefficient $\alpha = \alpha(\boldsymbol{x})$ is a positive piecewise constant function that may have large variations within Ω. In particular we aim to solve the case where the discontinuities in α are along subdomain interfaces as illustrated in Figure 3.1. In this case classical results break down.

In [78], Nataf, Xiang and Dolean proposed the construction of a coarse space with the ambition that the two-level method be robust with respect to heterogeneous coefficients for fairly arbitrary partitions into subdomains, *e.g.* provided by an automatic graph partitioner such as Metis or Scotch [54, 13]. The construction is based on the low-frequency modes associated with the eigenvalue problem for the Dirichlet-to-Neumann (DtN) operator on the boundary of each subdomain. It is the harmonic extensions of these low-frequency eigenvectors to the whole subdomain that span the coarse space. With this method, even for discontinuities along (rather than across) the subdomain interfaces, the solver is robust with respect to arbitrarily large jumps in the coefficients leading to a very efficient, automatic, preconditioning method for heterogeneous problems. As usual with domain decomposition methods, it is also well suited for parallel implementation.

The DtN coarse space is, by construction, ideally designed to deal with coefficient variations that are strictly interior to the subdomain. In this chapter, we recall the definition of the construction of the DtN coarse space in [78] and prove the robustness of the two-level Additive Schwarz preconditioner with this coarse space. The proof uses weighted Poincaré inequalities to derive some crucial estimates. These were introduced and successfully applied to different domain decomposition methods in [89, 88, 87, 38, 100]. Our analysis is inspired by the approach in [38], as well as by the abstract framework developed in [99]. The result that we obtain, generalizes the classical estimates of overlapping Schwarz methods to the case where the coarse space is richer than just the kernel of the local operators (which is the set of constant functions) [80], or other classical coarse spaces (cf. [112]). It is particularly well suited in the small overlap case.

The rest of the chapter is organized as follows. In Section 3.2 we present the model problem and some domain decomposition methods with an emphasis on coarse spaces.

In Section 3.3 the DtN coarse space is introduced as well as the heuristics on which it is based. In section 3.4 we give the convergence theorem (Theorem 3.10) and the proof of this result. Finally, in Section 3.5 we present some numerical results.

3.2 Preliminaries and notation

3.2.1 Model problem and discretization

We consider a discretization of the variational problem (3.1) with continuous, piecewise linear finite element functions. To define the finite element spaces and the approximate solution, we assume that we have a quasi-uniform, simplicial triangulation $\mathcal{T}_h$ of Ω:

$$\overline{\Omega} = \bigcup_{\tau \in \mathcal{T}_h} \tau.$$

The standard space of continuous and piecewise linear (with respect to $\mathcal{T}_h$) functions is then denoted by V_h, and the subspace of functions from V_h that vanish on the boundary of Ω by $V_{h,0}$.

The Galerkin approximation of (3.1) is: Find $u_h \in V_{h,0}$ such that

$$a(u_h, v_h) = \langle f, v_h \rangle, \quad \text{for all} \quad v_h \in V_{h,0}. \tag{3.2}$$

Let $\{\phi_k\}_{k=1}^n$ be the usual basis for $V_{h,0}$ consisting of nodal 'hat' functions with $n := \dim(V_{h,0})$. Then (3.2) is equivalent to the linear system

$$A\boldsymbol{u} \;=\; \boldsymbol{f}, \tag{3.3}$$

where $A_{k,l} := a(\phi_k, \phi_l)$, $f_k = \langle f, \phi_k \rangle$, $k, l = 1, \ldots, n$, and $\boldsymbol{u}$ is the vector of coefficients corresponding to the unknown finite element function u_h in (3.2). We use boldface for vectors and roman for finite element functions. We will frequently switch between bilinear forms and the corresponding matrices (*e.g.* $a(\cdot, \cdot)$ and A), as well as finite element functions in $V_{h,0}$ and the vectors of their coefficients in $\mathbb{R}^n$ (*e.g.* u_h and $\boldsymbol{u}$).

3.2.2 Two level Additive Schwarz preconditioner

In order to automatically construct robust two-level Schwarz type methods for (3.3) we first partition our domain Ω into a set of non overlapping subdomains $\{\Omega_j^*\}_{j=1}^N$ using for example a graph partitioner such as METIS or SCOTCH [54, 13]. Each subdomain Ω_j^* is then extended to a domain Ω_j by adding a number of adjacent fine grid elements, thus creating an overlapping decomposition $\{\Omega_j\}_{j=1}^N$ of Ω. This is illustrated in Figure 3.2.

Next we define the local functional spaces. Let $D \subset \Omega$ be any subset of Ω that is resolved by $\mathcal{T}_h$. The space of restrictions to D of the functions in V_h is denoted by $V_h(D)$

$$V_h(D) = \{v_{|D}; v \in V_h\}. \tag{3.4}$$

Similarly, the space of restrictions of functions from V_h, which are supported in $\bar{D}$ is denoted by $V_{h,0}(D)$

$$V_{h,0}(D) = \{v_{|D}; v \in V_h, \text{ supp}(v) \subset \bar{D}\}. \tag{3.5}$$

In particular we have that $V_h(D) \subset H^1(D)$ and $V_{h,0}(D) \subset H_0^1(D)$.

The one level Additive Schwarz preconditioner for (3.3) is introduced by defining restriction operators R_j from functions in $V_{h,0}$ to functions in $V_{h,0}(\Omega_j)$ or from vectors in $\mathbb{R}^n$ to vectors in $\mathbb{R}^{n_j}$, where $n_j := \dim(V_{h,0}(\Omega_j))$. As usual we use simple injection, i.e. for

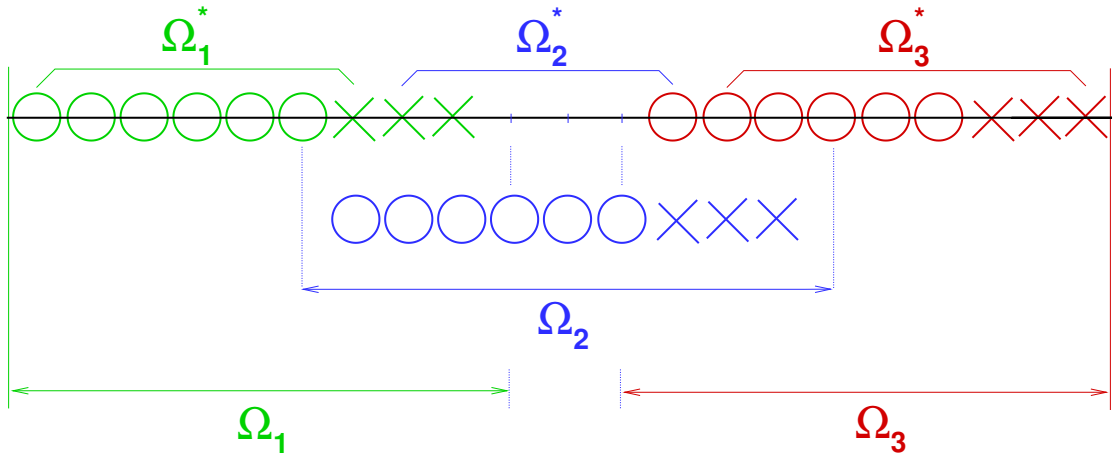


Figure 3.2: Illustration of the Domain Decomposition on a one dimensional three subdomain case. There are two partitions of Ω: an overlapping partition into Ω_1, Ω_2, Ω_3 and a non overlapping partition into Ω_1^*, Ω_2^*, Ω_3^*.

any $u \in V_{h,0}$ we set $(R_j u)(x_i) = u(x_i)$ at every grid node $x_i \in \Omega_j$. With this the one-level Additive Schwarz preconditioner is

$$M_{AS,1}^{-1} = \sum_{j=1}^{N} R_j^\top A_j^{-1} R_j \quad \text{where} \quad A_j := R_j A R_j^\top . \tag{3.6}$$

In terms of implementation this preconditioner is particularly well suited for preconditioning parallel iterative solvers, such as the conjugate gradient algorithm for (3.3) because all subdomain solves can be carried out independently of each other so $M_{AS,1}^{-1}$ is significantly less expensive to compute than A^{-1}. In terms of performance, a better and more flexible preconditioner is the following. Lets assume that we have a coarse space $V_H \subset V_{h,0}$ and a restriction operator R_H from $V_{h,0}$ to V_H, the two-level Additive Schwarz preconditioner can be defined just by adding a coarse solve to $M_{AS,1}^{-1}$:

$$M_{AS,2}^{-1} = M_{AS,1}^{-1} + R_H^\top A_H^{-1} R_H \quad \text{where} \quad A_H := R_H A R_H^\top . \tag{3.7}$$

For the scalar elliptic problem the coarse space V_H classically consists of finite element functions on a coarser triangulation $\mathcal{T}_H$ of Ω and R_H is the canonical restriction from $V_{h,0}$ to V_H. In [80], Nicolaides defines the interpolator $R_H^\top$ as follows (assuming that the degrees of freedom have been renumbered)

$$(R_H^\top)_{ij} = \begin{cases} 1, \text{ if } i \in \Omega_j^*, \\ 0, \text{ otherwise,} \end{cases} \Leftrightarrow R_H^\top = \begin{bmatrix} \mathbf{1}_{\Omega_1^*} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\Omega_2^*} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{1}_{\Omega_N^*} \end{bmatrix} \tag{3.8}$$

where, again, the subdomains Ω_j^* constitute the non overlapping partition of Ω and $\mathbf{1}_{\Omega_j^*}$ is a vector of all ones whose length is the number of unknowns in Ω_j^*.

Let D_j, $j = 1, \dots, N$, be diagonal matrices which correspond to a partition of unity subordinate to the partition of Ω into subdomains

$$\sum_{j=1}^{N} R_j^\top D_j R_j = I .$$

then the Nicolaides coarse space is adapted to the overlapping partition as follows

$$R_H^\top = \left[R_1^\top \mathbf{diag(D_1)} \quad | \quad R_2^\top \mathbf{diag(D_2)} \quad | \quad \dots \quad | \quad R_N^\top \mathbf{diag(D_N)} \right], \tag{3.9}$$

where $\mathbf{diag(D_j)}$ is the vector of the diagonal entries in D_j. The length of this vector is the number of unknowns in the overlapping subdomain Ω_j. We call this the Partition of unity coarse space and refer to [97] for a particular choice of D_j and the corresponding proof of convergence.

3.2.3 Two variants

The two level Additive Schwarz preconditioner (3.7) is only one choice among a family of one and two level domain decomposition methods. Here we present an alternate way to include the coarse correction into the solver and a non symmetric variant for the one level preconditioner. These will be studied numerically in Section 4.5.

The balancing preconditioner As it is nicely presented in [110] (see also references therein) two-level domain decomposition methods as well as multigrid methods and methods based on deflation are all defined by two ingredients: a full rank matrix $R_H^\top \in \mathbb{R}^{n\times m}$ whose columns span the m dimensional coarse space and an algebraic formulation of the coarse correction. These techniques imply solving a reduced size problem of order $m \times m$ called the coarse problem. The space spanned by the columns of $R_H^\top$ should contain the vectors responsible for the stagnation of the iterative method since they will be taken care of by the coarse solve (usually a direct solve).

In the next section we come back to the choice of $R_H^\top$. For now lets focus on the way to include the coarse problem into the solver. Lets assume that we are given a problem matrix A, a preconditioner M^{-1} and a coarse interpolator $R_H^\top : V_H \to V_h$. The first way to include the coarse space is to proceed additively as we did to build $M_{AS,2}^{-1}$ from $M_{AS,1}^{-1}$:

$$M_{ad}^{-1} = M^{-1} + R_H^\top A_H^{-1} R_H, \text{ where } A_H = R_H A R_H^\top. \tag{3.10}$$

The main advantage is that then the coarse solve can be performed in parallel at the same time as the application of M^{-1}.

Another choice is to apply the coarse correction in a multiplicative way. This is the balancing preconditioner and it was proposed by Mandel [67]. The abstract balancing preconditioner [67] for symmetric systems reads

$$P_{BNN} = (I - R_H^\top A_H^{-1} R_H A) M^{-1} (I - A R_H^\top A_H^{-1} R_H) + R_H^\top A_H^{-1} R_H, \text{ where } A_H = R_H A R_H^\top. \tag{3.11}$$

This is closely related to the hybrid Schwarz preconditioner in the abstract Schwarz framework (see (2.30) in section 2.2.1 of the introduction). For the preconditioner P_{BNN}, if we choose the initial approximation $\mathbf{x}_0 = R_H^\top A_H^{-1} R_H \mathbf{f}$, then the action of $(I - A R_H^\top A_H^{-1} R_H)$ is not required in practice, see [112, p.48]: the authors in [110] define

$$P_{ADEF2} = (I - R_H^\top A_H^{-1} R_H A) M^{-1} + R_H^\top A_H^{-1} R_H. \tag{3.12}$$

and prove that with the right initial guess P_{ADEF2} and P_{BNN} are equivalent. In particular P_{ADEF2} is as robust as P_{BNN} but requires one less coarse solve at each iteration. We mention also that for non symmetric problems the abstract balancing preconditioner is presented in [30].

The restricted additive Schwarz preconditioner In [9] a variant of the Additive Schwarz preconditioner (3.6) that requires fewer communications between subdomains was introduced. This is the Restricted Additive Schwarz (RAS) preconditioner:

$$M_{RAS}^{-1} := \sum_{j=1}^{N} \tilde{R}_j^\top A_j^{-1} R_j, \text{ where once more } A_j := R_j A R j^\top, \tag{3.13}$$

the interpolation operators $R_j^\top$ are unchanged and the $\tilde{R}_j^\top$ are their counterparts for the non overlapping partition (assuming the unknowns have been renumbered)

$$\tilde{R}_j^\top = \begin{bmatrix} \vdots \\ \mathbf{0} \\ I_{\Omega_j^*} \\ \mathbf{0} \\ \vdots \end{bmatrix} \text{ and } R_j = \begin{bmatrix} \dots & \mathbf{0} & I_{\Omega_j} & \mathbf{0} & \dots \end{bmatrix},$$

where we have denoted with I_{Ω_j} (respectively $I_{\Omega_j^*}$) the identity matrix whose dimension is the number of unknowns in Ω_j (respectively Ω_j^*). A simple way to generalize the RAS preconditioner is to replace the $\tilde{R}_j^\top$ corresponding to the non overlapping partition by $\tilde{R}_j^\top := D_j R_j$ where once more the D_j are partition of unity operators ($\sum_{j=1}^{N} R_j^\top D_j R_j = I$).

The reason why in the numerical section we also study the behavior of the DtN coarse space with RAS is that it is expected to converge faster than the symmetric preconditioner. The reason why is explained in [29]: the Additive Schwarz preconditioner overcorrects the solution in the overlap by adding contributions from multiple subdomains.

Preconditioner M_{AS}^{-1} is symmetric so the Krylov method of choice is the conjugate gradient (CG algorithm). Preconditioner M_{RAS}^{-1} is not symmetric so we use GMRES [96] as a solver.

Using M_{AS}^{-1} or M_{RAS}^{-1}, takes care of the very large eigenvalues of the coefficient matrix. The small eigenvalues on the other hand still exist and may hamper convergence. Next we design a coarse space which efficiently deals with them.

3.3 DtN coarse space

An ideal choice for the coarse basis would be to use exactly the eigenvectors of $M^{-1}A$ corresponding to small eigenvalues. However the cost of computing the lower part of the spectrum of a matrix is larger than the cost of solving the linear system so this is not an great plan.

We will look for a coarse interpolator $R_H^\top$ that consists of local contributions:

$$R_H^\top = \begin{bmatrix} W^1 & 0 & \cdots & 0 \\ \vdots & W^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & W^N \end{bmatrix}, \tag{3.14}$$

where the W^j are rectangular matrices whose columns are vectors in $V_h(\Omega_j)$. This way the computation of the basis vectors for the coarse space can be done locally and implemented

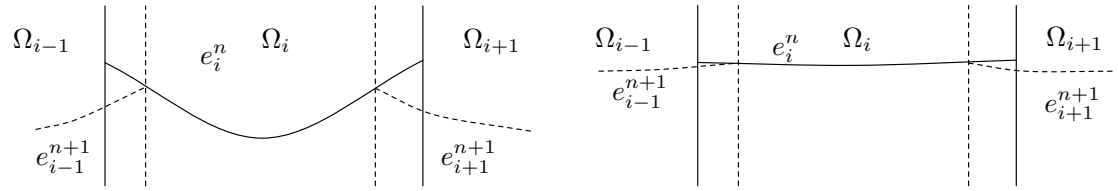


Figure 3.3: Fast or slow convergence of the Schwarz algorithm.

in parallel. Also, the coarse operator $A_H = R_H A R_H^\top$ will be sparse as a result of the sparsity of $R_H^\top$, the non zero components of A_H corresponding to adjacent subdomains. In the next subsections we give the heuristic motivating the particular choice of W_i which we call the DtN coarse space and its rigorous definition.

3.3.1 Definition of the DtN coarse space

Lets motivate the choice for the DtN coarse space. Because the vectors in the coarse space receive a special treatment (the coarse solve), it is most efficient to span the coarse space with vectors that slow down the convergence of the one level method. The question is: how do we identify these problematic vectors?

The heuristics for the choice of DtN coarse space have already been presented in the introduction (Section 2.3.1). We sum up the main ideas here based on the one dimensional example in Figure 3.3. We consider applying the original alternating Schwarz algorithm (2.2) to this problem. At each step of the algorithm the error e_j^n in subdomain Ω_j is computed by extending harmonically to the interior of the subdomain the boundary conditions that are transmitted by the neighbouring subdomains. By the maximum principle we are ensured that the error is a convex function: in non pathological cases the error decreases in the overlap and the transmission condition that is given to the neighbours is an improved condition. How much this condition is improved by depends on how much the error decreases in the overlap. This is what is illustrated in Figure 3.3: on the left hand side the error decreases faster that on the right leading to a faster convergence.

A good guess is that a fast decay of the error in the overlap is the direct consequence of a large normal derivative of the error on the boundary. This is why the Dirichlet to Neumann map plays such an important role. We introduce it next. To this end, let $j = 1, \ldots, N$ and let

$$\mathrm{tr}_j \alpha(\boldsymbol{x}) := \limsup_{\Omega_j \ni \boldsymbol{y} \to \boldsymbol{x}} \alpha(\boldsymbol{y}), \text{ for almost every } \boldsymbol{x} \in \partial\Omega_j,$$

denote the trace of α on $\partial\Omega_j$.

Definition 3.1. Let $\Gamma := \partial\Omega_j$ and $\boldsymbol{n}_j$ be the unit outward normal to Ω_j on Γ. Then for any $v_\Gamma : \Gamma \to \mathbb{R}$ such that $v_\Gamma|_{\partial\Omega} = 0$ if $\Gamma \cap \partial\Omega \neq \emptyset$ define

$$\mathrm{DtN}_j(v_\Gamma) := \mathrm{tr}_j \alpha \frac{\partial v}{\partial \boldsymbol{n}_j}\bigg|_\Gamma, \text{ where } v \text{ is the solution of } \begin{cases} -\nabla\cdot(\alpha\nabla v) = 0 & \text{in } \Omega_j \\ v = v_\Gamma & \text{on } \Gamma \end{cases}. \tag{3.15}$$

With words, the DtN operator takes boundary data on Γ, computes its harmonic extension v to the whole of Ω_j and returns the normal derivative of v on Γ. This is indeed a map between Dirichlet and Neumann data.

Our strategy is to span the coarse space with the low frequency modes of the Dirichlet to Neumann operator DtN_j with respect to the weighted l_2 norm on Γ. More precisely the coarse space W^j is spanned by the harmonic extensions of the generalized eigenvectors corresponding to the smallest generalized eigenvalues of the following generalized eigenproblem:

$$\mathrm{DtN}_j(v_\Gamma) \;=\; \lambda\, \mathrm{tr}_j \alpha\, v_\Gamma. \tag{3.16}$$

For simplicity let us consider an interior subdomain Ω_j that does not touch the (Dirichlet) boundary of the global domain Ω. The other case carries through in a similar way. Instead of looking for an eigenpair of equation (3.16) and then computing v, the harmonic extension of v_Γ, we directly search for the pair (λ, v). It is straightforward to check that it satisfies:

$$\left\{ \begin{array}{rcll} -\nabla\cdot(\alpha\nabla v) & = & 0 & \text{in } \Omega_j\ , \\ \mathrm{tr}_j \alpha \dfrac{\partial v}{\partial \boldsymbol{n}_j} & = & \lambda\, \mathrm{tr}_j \alpha\, v & \text{on } \Gamma\ . \end{array} \right. \tag{3.17}$$

The step by step construction of the coarse interpolator (at the continuous level) is described in Algorithm 3.1 with the use of partition of unity functions $\{\chi_j\}_{j=1}^N$ defined on Ω, subordinate to the overlapping decomposition $\{\Omega_j\}_{j=1}^N$. One particular choice for χ_j (for which we prove the convergence theorem) is given later on in (3.25).

Algorithm 3.1 DtN coarse space construction

– In parallel for all subdomains $1 \leq j \leq N$ do:

1. Compute the generalized eigenpairs (v_j^1, λ_j^1), $(v_j^2, \lambda_j^2), \ldots, (v_j^{m_j}, \lambda_j^{m_j}) \in V_h(\Omega_j) \times \mathbb{R}$ of (3.17) such that

$$\lambda_j^1 \leq \ldots \leq \lambda_j^{m_j} < 1/diam(\Omega_j) \leq \lambda_j^{m_j+1} \leq \ldots.$$

2. Let the rectangular matrix W^j with m_j columns be

$$W^j = [I_h(\chi_j\, v_j^1) | \ldots | I_h(\chi_j\, v_j^{m_j})]\,,$$

where I_h is the standard nodal interpolator onto the finite element space and χ_j is the partition of unity function.

– The global coarse interpolator is

$$R_H^\top = \begin{bmatrix} W^1 & 0 & \cdots & 0 \\ \vdots & W^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & W^N \end{bmatrix}, \tag{3.18}$$

The strategy we advocate is to select the generalized eigenvectors such that

$$\lambda < 1/diam(\Omega_j) \tag{3.19}$$

where $diam(\Omega_j)$ is the diameter of subdomain Ω_j. The reason why is that we want to do as well as the classical (Partition of unity) coarse space in the constant coefficient case. In that case, for a shape regular subdomain, the first non zero eigenvalue is of order $1/diam(\Omega_j)$ (see [31]) the corresponding eigenvector is not in the coarse space and convergence is good.

3.3.2 DtN coarse space at the discrete level

The variational formulation of (3.17) is: Find $(\lambda, v) \in \mathbb{R} \times V_h(\Omega_j)$ such that

$$\int_{\Omega_j} \alpha \nabla v \cdot \nabla w = \lambda \int_\Gamma \mathrm{tr}_j \alpha \, v \, w \,, \quad \forall w \in V_h(\Omega_j). \tag{3.20}$$

To obtain the discrete form of the generalized eigenvalue problem (3.20), we first introduce bilinear forms $a_j : V_h(\Omega_j) \times V_h(\Omega_j) \to \mathbb{R}$ and $b_j : V_h(\Omega_j) \times V_h(\Omega_j) \to \mathbb{R}$ as

$$a_j(v, w) := \int_{\Omega_j} \alpha \nabla v \cdot \nabla w \text{ and } b_j(v, w) := \int_\Gamma \mathrm{tr}_j \alpha \, v \, w \,, \quad \forall v,\, w \in V_h(\Omega_j) \,. \tag{3.21}$$

Then, for the finite element basis $\{\phi_k\}_{k=1,\dots,n}$, we introduce the matrices corresponding to these bilinear forms. Let $A^{(j)}$ be the coefficient matrix associated with the variational form a_j

$$(A^{(j)})_{kl} = \int_{\Omega_j} \alpha \nabla \phi_k \cdot \nabla \phi_l,$$

and let $M^{(j)}$ be the weighted mass matrix on Γ associated with the variational form b_j

$$(M^{(j)})_{kl} := \int_\Gamma \mathrm{tr}_j \alpha \phi_k \, \phi_l \,.$$

Then we can write the finite element approximation of generalized eigenproblem (3.20): Find $(\lambda, \mathbf{V})$ such that

$$A^{(j)} \mathbf{V} = \lambda M^{(j)} \mathbf{V}. \tag{3.22}$$

Let I (resp. Γ) be the set of indices corresponding to the interior (resp. boundary) degrees of freedom and $n_\Gamma := \#\Gamma$ be the number of interface degrees of freedom. Then with block notations we get

$$A^{(j)} = \begin{bmatrix} A_{\mathrm{II}}^{(j)} & A_{\mathrm{I\Gamma}}^{(j)} \\ A_{\mathrm{\Gamma I}}^{(j)} & A_{\mathrm{\Gamma\Gamma}}^{(j)} \end{bmatrix}, \text{ and } M^{(j)} = \begin{bmatrix} 0 & 0 \\ 0 & M_\Gamma^{(j)} \end{bmatrix}$$

Remark 3.2. Notice that, the matrix $A^{(j)}$ in the generalized eigenvalue problem is not $A_j = R_j A R_j^\top$ from the definition of the Schwarz preconditioners and that it cannot be extracted from the original global matrix. Indeed, for the global domain Ω, the coefficient matrix is given by $A_{kl} = \int_\Omega \alpha \nabla \phi_k \cdot \nabla \phi_l$ and that may differ from $\int_{\Omega_j} \alpha \nabla \phi_k \cdot \nabla \phi_l$. More precisely, three of the four blocks are identical $A_{\mathrm{II}}^{(j)} = A_{\mathrm{II}}$, $A_{\mathrm{\Gamma I}}^{(j)} = A_{\mathrm{\Gamma I}}$ and $A_{\mathrm{I\Gamma}}^{(j)} = A_{\mathrm{I\Gamma}}$. However $A_{\mathrm{\Gamma\Gamma}}^{(j)} \neq A_{\mathrm{\Gamma\Gamma}}$, since $A_{\mathrm{\Gamma\Gamma}}^{(j)}$ refers to the matrix prior to assembly with the neighbouring subdomains.

With block notation, generalized eigenproblem (3.22) can be rewritten as

$$A^{(j)} \mathbf{V} = \lambda \begin{bmatrix} 0 & 0 \\ 0 & M_\Gamma^{(j)} \end{bmatrix} \mathbf{V} = \lambda M^{(j)} \, \mathbf{V}, \tag{3.23}$$

or, as a system,

$$\begin{cases} A_{\mathrm{II}} \mathbf{V}_\mathrm{I} + A_{\mathrm{I\Gamma}} \mathbf{V}_\Gamma &= \mathbf{0}, \\ A_{\mathrm{\Gamma\Gamma}}^{(j)} \mathbf{V}_\Gamma + A_{\mathrm{\Gamma I}} \mathbf{V}_\mathrm{I} &= \lambda M_\Gamma^{(j)} \mathbf{V}_\Gamma. \end{cases}$$

We use the first equation to eliminate the interior unknowns $\mathbf{V}_\mathrm{I}$ and introduce the Schur complement $S_\Gamma^{(j)} := A_{\Gamma\Gamma}^{(j)} - A_{\Gamma\mathrm{I}}A_{\mathrm{II}}^{-1}A_{\mathrm{I}\Gamma}$, then we get a generalized eigenvalue problem

$$S_\Gamma^{(j)}\mathbf{V}_\Gamma \;=\; \lambda\, M_\Gamma^{(j)}\mathbf{V}_\Gamma\,, \tag{3.24}$$

which is exactly the discrete form of the original generalized eigenproblem (3.16). Indeed, it is well known that the discrete counterpart of the DtN map is the Schur complement $S_\Gamma^{(j)}$ (this results from the divergence theorem).

Let $(\lambda_j^k, \mathbf{V}_j^k)_{k=1}^{n_\Gamma}$ be the n_Γ eigenpairs of (3.24) numbered in increasing order of λ_j^k. Since matrices $S_\Gamma^{(j)}$ and $M_\Gamma^{(j)}$ are symmetric, the eigenvectors $\mathbf{V}_{\Gamma,j}^k,\ k = 1, \ldots, n_\Gamma$, satisfy

$$\langle \mathbf{V}_{\Gamma,j}^k, S_\Gamma^{(j)}\mathbf{V}_{\Gamma,j}^l\rangle = 0 \quad \text{and} \quad \langle \mathbf{V}_{\Gamma,j}^k, M_\Gamma^{(j)}\mathbf{V}_{\Gamma,j}^l\rangle = 0, \quad \text{if} \quad k \neq l.$$

Matrix $M_\Gamma^{(j)}$ is positive definite so we may say that this is an orthogonality property with respect to the norm induced by $M_\Gamma^{(j)}$. Matrix $S_\Gamma^{(j)}$ is symmetric positive semi-definite and in the case of an interior subdomain, there is exactly one eigenvalue that is 0 corresponding to the constant eigenvector. The harmonic extension $\mathbf{V}_j^k = \begin{bmatrix} \mathbf{V}_{\Gamma,j}^k \\ -A_{\mathrm{II}}^{-1}A_{\mathrm{I}\Gamma}\mathbf{V}_{\Gamma,j}^k \end{bmatrix}$ of $\mathbf{V}_{\Gamma,j}^k$ is an eigenvector corresponding to eigenvalue λ_j^k of generalized eigenproblem (3.22). These harmonic extensions also satisfy an orthogonality type property with respect to $A^{(j)}$ since for any $k \neq l$

$$\langle \mathbf{V}_j^k, A^{(j)}\mathbf{V}_j^l\rangle \;=\; \begin{bmatrix} -A_{\mathrm{II}}^{-1}A_{\mathrm{I}\Gamma}\mathbf{V}_{\Gamma,j}^k \\ \mathbf{V}_{\Gamma,j}^k \end{bmatrix}^\top \begin{bmatrix} \mathbf{0} \\ S_\Gamma^{(j)}\mathbf{V}_{\Gamma,ij}^l \end{bmatrix} \;=\; \langle \mathbf{V}_{\Gamma,i_j}^k, S_\Gamma^{(j)}\mathbf{V}_{\Gamma,j}^l\rangle \;=\; 0,$$

and an orthogonality type property with respect to $M_\Gamma^{(j)}$ since for any $k \neq l$

$$\langle \mathbf{V}_j^k, M^{(j)}\mathbf{V}_j^l\rangle \;=\; \begin{bmatrix} -A_{\mathrm{II}}^{-1}A_{\mathrm{I}\Gamma}\mathbf{V}_{\Gamma,j}^k \\ \mathbf{V}_{\Gamma,j}^k \end{bmatrix}^\top \begin{bmatrix} 0 & 0 \\ 0 & M_\Gamma^{(j)}\mathbf{V}_{\Gamma,j}^l \end{bmatrix} \;=\; \langle \mathbf{V}_{\Gamma,j}^k, M_\Gamma^{(j)}\mathbf{V}_{\Gamma,j}^l\rangle \;=\; 0.$$

Since the kernel of $\begin{bmatrix} 0 & 0 \\ 0 & M_\Gamma^{(j)} \end{bmatrix}$ consists of the vectors whose components in Γ are zero, all the remaining eigenvalues of (3.23) are ∞, and so the smallest eigenvalues of (3.24) are also the smallest eigenvalues of (3.23). This means that these two discrete generalized eigenproblems are suitable alternatives for implementing the DtN coarse space.

3.3.3 A remark

The construction we propose is, to some extent, inspired by two observations already made elsewhere, namely

- that robust coarse basis functions can in many cases be obtained on standard simplicial meshes by harmonically extending suitable boundary data to the interior of coarse mesh elements [47] (i.e. multiscale finite element coarse spaces),
- that local spectral information about the underlying differential operator can be used to obtain fully robust coarse spaces [38, 99].

By combining both ideas the goal is to identify all the vectors which need to be in the coarse space and at the same time not too many vectors. Inclusions that are inside a subdomain will not "trigger" (unnecessary) additional coarse basis functions. This can

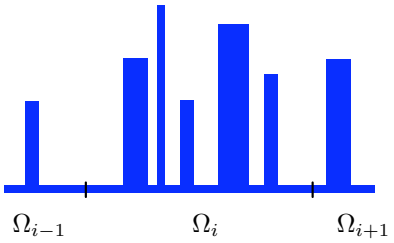


Figure 3.4: 1D example with many high coefficient inclusions per subdomain.

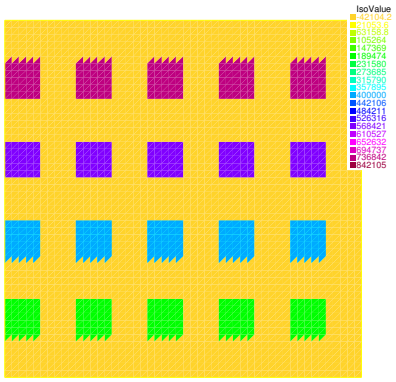


Figure 3.5: Coefficient distribution on a subdomain Ω_j for a two-dimensional model problem with high-permeability inclusions.

happen in the case of generalized eigenvalue problems that are defined on the entire subdomain, such as those in [38], unless, as described in [39], they are combined with a partition of unity constructed *via* the multiscale finite element techniques in [47].

Let us illustrate one clear benefit of working with the DtN map on the one-dimensional example in Figure 3.4. Independently of the coefficient variations in the interior of Ω_j the DtN coarse space consists of at most two basis functions per subdomain since the DtN map is a two by two matrix which has exactly two eigenmodes.

For one more illustration, consider the two-dimensional permeability field α on the subdomain Ω_j shown in Figure 3.5. We see in Figure 3.6 (left) a typical DtN eigenvector associated with one of the boundary inclusions. Since it is harmonic in the interior of Ω_j, it has much lower energy than a typical eigenvector of the corresponding eigenproblem $-\nabla \cdot (\alpha \nabla v) = \lambda \alpha v$, posed on the entire subdomain Ω_j and shown in Figure 3.6 (right). This is achieved without the use of a coefficient-adapted partition of unity (such as in [39]).

3.4 Theoretical analysis

For the theoretical analysis we focus on two level Additive Schwarz preconditioner (3.7). With the Balanced preconditioner (3.11) convergence is always better than with the Additive preconditioner so the theory presented here goes through also in this case.

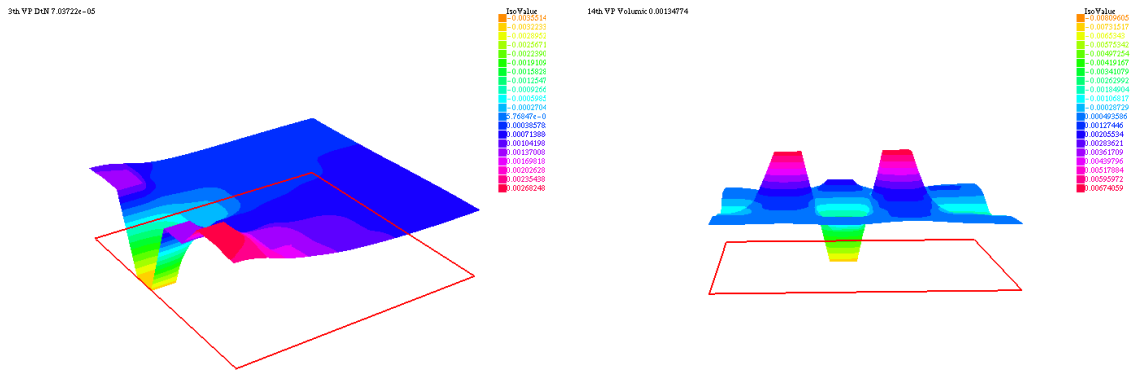


Figure 3.6: Typical eigenvector of the DtN map for the model problem in Figure 3.5 (left plot) and typical eigenvector of the full subdomain eigenproblem $-\nabla\cdot(\alpha\nabla v) = \lambda\,\alpha v$ (right plot).

3.4.1 A few theoretical assumptions and tools

Throughout the remainder of this chapter, the notation $E \lesssim F$ (for two quantities E, F) means that E/F is bounded from above independently, not only of the mesh size h and the method specific parameters (such as the diameter $\mathrm{diam}\,(\Omega_j)$ of the subdomain and the size of the overlap δ_j defined below), but also of the values taken by the coefficient α. Moreover $E \eqsim F$ means that $E \lesssim F$ and $E \lesssim F$.

Particular choice for the partition of unity Assume that the coarse space is built following the procedure in Algorithm 3.1 with the partition of unity functions defined in the following way: for each fine grid node $\boldsymbol{x}_k \in \Omega$, let the index set $\mathcal{N}(\boldsymbol{x}_k)$ contain the indices of all the domains Ω_i such that $\boldsymbol{x}_k \in \Omega_i$, then, for each subdomain Ω_j define $\chi_j \in V_{h,0}(\Omega_j)$ by setting

$$\chi_j(\boldsymbol{x}_k) := \frac{\mathrm{dist}(\boldsymbol{x}_k, \partial\Omega_j)}{\sum_{i\in\mathcal{N}(\boldsymbol{x}_k)} \mathrm{dist}(\boldsymbol{x}_k, \partial\Omega_i)}, \quad \text{at all nodes } \boldsymbol{x}_k \in \Omega_j. \tag{3.25}$$

Clearly these functions form a partition of unity on Ω and satisfy $0 \le \chi_j \le 1$. Moreover, if

$$\Omega_j^\circ := \{\boldsymbol{x} \in \Omega_j : \chi_j(\boldsymbol{x}) < 1\}$$

denotes the part of Ω_j that overlaps neighbouring domains, then it is also easy to verify that

$$|\nabla\chi_j| \lesssim \delta_j^{-1} \tag{3.26}$$

where δ_j denotes the width of Ω_j° at the narrowest place. Since $\mathcal{T}_h$ was assumed to be quasi-uniform and the overlapping decomposition $\{\Omega_j\}_{j=1}^N$ was obtained by adding layers of fine grid elements, we have $\delta_j \eqsim \delta_{j'}$, for any two neighbouring subdomains Ω_j and $\Omega_{j'}$.

Assumptions on the coefficient distribution We have assumed that the coefficient α is piecewise constant. To be more precise, we assume that the domain is a union of polygonal (polyhedral) subdomains $\mathcal{Y}_l$, such that:

$$\overline{\Omega} = \bigcup_{l=1}^m \overline{\mathcal{Y}}_l \quad \text{and} \quad \alpha(\boldsymbol{x}) = \alpha_l, \text{ for all } \boldsymbol{x} \in \mathcal{Y}_l \text{ and } l = 1, \ldots, m.$$

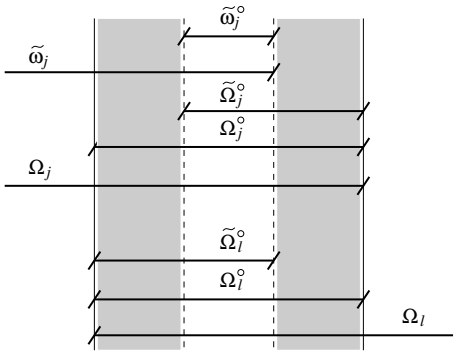


Figure 3.7: Overlap region between two subdomains Ω_j and Ω_l with the various subsets used in the analysis.

We also assume that the triangulation $\mathcal{T}_h$ resolves $\mathcal{Y}_l$, namely, for $l = 1, \ldots, m$, we have:

$$\overline{\mathcal{Y}}_l = \bigcup_{\tau \in \mathcal{T}_{h,l}} \tau, \tag{3.27}$$

where $\mathcal{T}_{h,l} \subset \mathcal{T}_h$, for $l = 1, \ldots, m$.

By definition of the parameter δ_j, the overlap Ω_j° contains all points which are at most at a distance δ_j from the boundary $\partial\Omega_j$. For each $j \in \{1, \ldots, N\}$, we define the following subset of the overlap:

$$\widetilde{\Omega}_j^\circ := \text{interior}\Big(\bigcup_{k=1,\ldots,K_j} \overline{D}_{jk} \Big) \subset \Omega_j^\circ ,$$

where the regions $D_{jk} \subset \Omega_j^\circ$, $k = 1, \ldots, K_j$, are assumed to form a *shape-regular* (overlapping or non overlapping) partition of $\widetilde{\Omega}_j^\circ$, such that for each $k = 1, \ldots, K_j$, $\text{diam}\,(D_{jk}) \eqsim \delta_j$, $|D_{jk}| \eqsim \delta_j^d$ and $\overline{D}_{jk} \cap \partial\Omega_j \neq \emptyset$. We call $\widetilde{\Omega}_j^\circ$ the *boundary layer* of Ω_j. An example is shown in Figure 3.7. Without loss of generality, we assume that the triangulation $\mathcal{T}_h$ resolves each of the regions D_{jk}.

We now make two technical assumptions on the distribution $\alpha(\boldsymbol{x})$ and on its interplay with the partition into subdomains. These assumptions are needed in our analysis. They are always satisfied in the case of minimal overlap (i.e. the case of one or two layers). Some examples are given in Figure 3.8.

Assumption 3.3. *We assume that there exists a (second) partition of unity* $\{\widetilde{\chi}_j\}_{j=1}^N \subset V_{h,0}$ *associated with* $\{\Omega_j\}_{j=1}^N$, *such that* $0 \leq \widetilde{\chi}_j \leq 1$, $\text{supp}(\nabla\widetilde{\chi}_j) \subset \overline{\widetilde{\Omega}_j^\circ}$ *and* $|\nabla\widetilde{\chi}_j| \lesssim \delta_j^{-1}$, *in other words we assume that the overlap of the boundary layer* $\widetilde{\Omega}_j^\circ$ *with the boundary layers of the neighbouring domains is at least of width* $\eqsim \delta_j$ *everywhere.* [1]

Assumption 3.4. *Again, let* $\text{tr}_j\alpha(\boldsymbol{x}) := \limsup_{\Omega_j \ni \boldsymbol{y} \to \boldsymbol{x}} \alpha(\boldsymbol{y})$, *for almost every* $\boldsymbol{x} \in \partial\Omega_j$, *denote the trace of* α *on* $\partial\Omega_j$. *For each* $j \in \{1, \ldots, N\}$ *and for each* $k = 1, \ldots, K_j$, *we assume that*

(i) there exists a $(d-1)$*–dimensional manifold* $X_k \subset \overline{D}_{jk} \cap \partial\Omega_j$ *with* $|X_k| \eqsim \delta_j^{d-1}$, *such that* $\text{ess sup}_{\boldsymbol{x},\boldsymbol{y} \in X_k} \text{tr}_j\alpha(\boldsymbol{x})/\text{tr}_j\alpha(\boldsymbol{y}) = \mathcal{O}(1)$,

1. We do not need to construct this second partition of unity in practice. It is only needed for the analysis.

(ii) there exists a path $P_{\boldsymbol{y}}$ from each point $\boldsymbol{y} \in D_{jk}$ to X_k, such that $\alpha(\boldsymbol{x})$ is an increasing function along $P_{\boldsymbol{y}}$ (from $\boldsymbol{y}$ to X_k, except possibly on a subset of $P_{\boldsymbol{y}}$ of measure zero).

When (ii) holds, $\alpha(\boldsymbol{x})$ is called *quasi–monotone* on D_{jk} with respect to X_k and $P_{\boldsymbol{y}}$ is called a *quasi-monotone path.*

Definition of weighted norms and semi-norms For any domain $D \subset \Omega$ we need the usual norms, with the standard notations $\|\cdot\|_{L_2(D)}$, $|\cdot|_{H^1(D)}$ and $\|\cdot\|_{H^1(D)}$, as well as the L_2 inner product $(v,w)_{L_2(D)}$. In addition to this, we need to define some related weighted quantities, which will prove very useful in the following:

– the weighted H^1 (or energy) norm

$$|v|^2_{a,D} = \int_D \alpha |\nabla v|^2. \tag{3.28}$$

Note that $|\cdot|_{a,D}$ is indeed a norm on $H^1_0(D)$; on all of $H^1(D)$ it is only a seminorm.

– the weighted L_2 norm and the weighted L_2 inner product

$$\|v\|^2_{0,\alpha,D} = \int_D \alpha v^2 \quad \text{and} \quad (v,w)_{0,\alpha,D} = \int_D \alpha v w. \tag{3.29}$$

When $D = \Omega$ we omit the domain from the subscript and write $\|\cdot\|_a$ and $\|\cdot\|_{0,\alpha}$ instead of $\|\cdot\|_{a,\Omega}$ and $\|\cdot\|_{0,\alpha,\Omega}$, respectively.

Finally, we will also need averages and norms defined on $(d-1)$–dimensional manifolds $X \subset \mathbb{R}^d$, namely for any $v \in L_2(X)$ and for any $\beta \in L_\infty(X)$ we define

$$\overline{v}^X := \frac{1}{|X|}\int_X v \quad \text{and} \quad \|v\|^2_{0,\beta,X} := \int_X \beta v^2.$$

3.4.2 Intermediary estimates

The following lemma, based on Assumption 3.4, is from [89].

Lemma 3.5. Let Assumption 3.4 be satisfied. There exists a uniform constant $C_P > 0$ independent of the coefficient values $\{\alpha_l\}_{l=1}^m$, such that the following *weighted Poincaré/ Friedrichs type inequalities* hold for all $j = 1, \ldots, N$ and $k = 1, \ldots, K_j$:

$$\|v - \overline{v}^{X_k}\|_{0,\alpha,D_{jk}} \leq C_P\, \delta_j\, |v|_{a,D_{jk}}, \qquad \text{for all } v \in V_h(D_{jk}), \text{ and} \tag{3.30}$$

$$\|v\|_{0,\alpha,D_{jk}} \leq C_P\, \delta_j\, |v|_{a,D_{jk}}, \qquad \text{for all } v \in V_h(D_{jk}) \text{ with } v|_{X_k} = 0. \tag{3.31}$$

The constant C_P may depend on δ_j/h (see Remark 3.6 for details).

Proof. Theorems 2.2, 2.7 and 3.3 in [89]. □

Remark 3.6. *(a) Assumptions 3.3 and 3.4 are technical and so, in Figure 3.8, we give some typical examples where the assumptions are either verified or not verified. Essentially the only situation where they can not be verified is when a region $\mathcal{Y}_l$ where the coefficient is large separates the remainder of the overlap Ω_j° into two parts, but does not touch any of the boundaries of Ω_j°, neither inner nor outer (see Figure 3.8 (d)). The assumptions are always satisfied in the case of minimal overlap (i.e. in the case of one or two layers).*

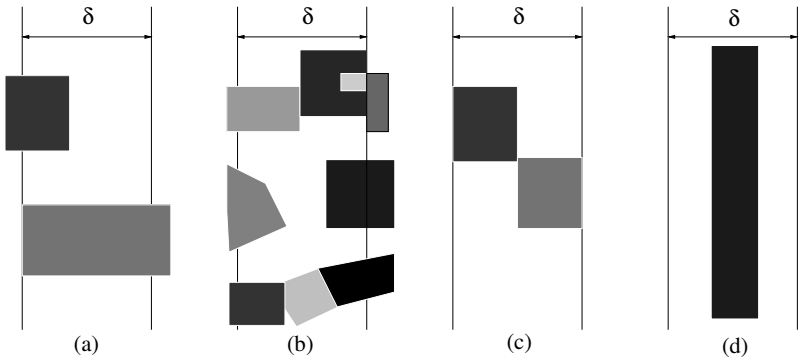


Figure 3.8: Overlap region between two subdomains with high-permeability inclusions (darker color represents higher permeability). We distinguish three cases: Assumption 3.4 is verified and Lemma 3.5 holds with $C_P = \mathcal{O}(1)$ (a & b), with $C_P = \mathcal{O}(\log(\delta_j/h))$ (c) and not verified (d).

(b) Provided Assumptions 3.3 and 3.4 are satisfied, the constant C_P in Lemma 3.5 will always be independent of the coefficient values, and thus of any jumps. It will also be independent of $\operatorname{diam}(\Omega_j)$*, but it may depend on the mesh size h and on the size of the overlap δ_j through the ratio δ_j/h. This was analyzed extensively in [89, 88]. The constant C_P is independent of δ_j/h if the regions $\mathcal{Y}_l$, where the coefficient is constant, intersect always in $(d-1)$–dimensional manifolds of measure $\gtrsim \delta_j^{d-1}$ (e.g. Figure 3.8(a & b)). If any regions $\mathcal{Y}_l$ intersect only in $(d-2)$–dimensional manifolds (i.e. a point in 2D or an edge in 3D) or if some of the intersections are only of measure $\eqsim h^{d-1}$, then C_P will in general be $\mathcal{O}(\log(\delta_j/h))$. In 3D, if some of the regions $\mathcal{Y}_l$ touch each other only in a point, then C_P may be $\mathcal{O}(\delta_j/h)$. Since we are mainly interested in the small overlap case (i.e. $\delta_j \leq ch$ for some small constant $c = 2, 4, 6$, etc.), we will not discuss this further.*

(c) Extensions similar to those in [87] to cases where some of the regions D_{jk} only touch $\partial\Omega_j$ in a point (or in an edge in 3D), or where the regions D_{jk} may become long and thin would also be possible. These may also add dependencies of C_P on $\operatorname{diam}(\Omega_j)/\delta_j$*.*

(d) Due to (i) *and* (ii)*, the manifold X_k has to lie in the closure of the region $\mathcal{Y}_l$ where α takes its maximum on D_{jk}.*

The following result, which is essentially a corollary of Lemma 3.5 will be the key tool in the analysis below.

Lemma 3.7. Let Assumption 3.4 be satisfied. Then

$$\|u\|^2_{0,\alpha,\widetilde{\Omega}_j^\circ} \lesssim C_P^2\delta_j^2|u|^2_{a,\widetilde{\Omega}_j^\circ} + \delta_j\|u\|^2_{0,\mathrm{tr}_j\alpha,\partial\Omega_j\setminus\partial\Omega}, \quad \text{for all } u \in V_h(\widetilde{\Omega_j^\circ}).$$

Proof. Let $\{D_{jk}\}_{k=1}^{K_j}$ be as defined above and let X_k be the $(d-1)$–dimensional manifold associated with D_{jk} in Assumption 3.4. Let $\|\alpha\|_{\infty,D_{jk}} := \operatorname{esssup}\{\alpha(\boldsymbol{x}) : \boldsymbol{x} \in D_{jk}\}$. Then it follows from Lemma 3.5, as well as the triangle and the Cauchy-Schwarz inequalities,

that

$$\begin{aligned}\tfrac{1}{2}\|u\|^2_{0,\alpha,D_{jk}} &\leq \|u-\overline{u}^{X_k}\|^2_{0,\alpha,D_{jk}} + \|\overline{u}^{X_k}\|^2_{0,\alpha,D_{jk}} \\ &\leq C_P^2\delta_j^2|u|^2_{a,D_{jk}} + \frac{|D_{jk}|}{|X_k|^2}\|\alpha\|_{\infty,D_{jk}}\left(\textstyle\int_{X_k} u\right)^2 \\ &\leq C_P^2\delta_j^2|u|^2_{a,D_{jk}} + \frac{|D_{jk}|}{|X_k|}\|\alpha\|_{\infty,D_{jk}}\textstyle\int_{X_k} u^2 \\ &\lesssim C_P^2\delta_j^2|u|^2_{a,D_{jk}} + \delta_j\,\|u\|^2_{0,\alpha,X_k}\,.\end{aligned} \tag{3.32}$$

In the last step, we have used Assumption 3.4(i) and the fact that $\alpha(x)$ reaches its maximum on $\overline{D}_{jk}$ in a set containing X_k (cf. Remark 3.6(d)). If $\partial D_{jk}\cap\partial\Omega\neq\emptyset$, we use (3.31) instead of (3.30). In this case, only the first term on the right hand side of inequality (3.32) appears. The final result follows by summing over $k=1,\ldots,K_j$. □

Note that this lemma is an extension of the small overlap trick in [25] to the variable coefficient case (see also [112, Lemma 3.10]).

Let us assume further, that the eigenvectors are normalised in the $M^{(j)}$ semi-norm:

$$\langle \mathbf{V}_j^k, M^{(j)}\mathbf{V}_j^k\rangle = 1 \quad\Leftrightarrow\quad b_j(v_j^k, v_j^k) = 1,$$

where b_j is the bilinear form associated with $M^{(k)}$ as in (3.21). For any $u\in V_h(\Omega_j)$, we can define the projection onto $\mathrm{span}\{v_j^k\}_{k=1}^{m_j}$ by

$$\Pi_j u \;:=\; \sum_{k=1}^{m_j} b_j\left(v_j^k, u\right)\, v_j^k. \tag{3.33}$$

The projection is stable and satisfies a weak approximation property, as the following theorem shows.

Theorem 3.8. Let Assumption 3.4 hold. Then, for any $u\in V_h(\Omega_j)$,

$$|\Pi_j u|_{a,\Omega_j} \;\leq\; |u|_{a,\Omega_j} \quad\text{and} \tag{3.34}$$

$$\|u-\Pi_j u\|_{0,\alpha,\widetilde{\Omega}_j^\circ} \;\lesssim\; \sqrt{c_j(m_j)}\,\delta_j\,|u|_{a,\Omega_j}\,, \tag{3.35}$$

where $c_j(m_j) := C_P^2 + \left(\delta_j\lambda^{(j)}_{m_j+1}\right)^{-1}$.

Proof. The stability estimate (3.34) follows immediately from the fact that Π_j is a projection satisfying $a_j(\Pi_j u, u-\Pi_j u)=0$. To prove (3.35) let us first apply Lemma 3.7, i.e.

$$\|u-\Pi_j u\|^2_{0,\alpha,\widetilde{\Omega}_j^\circ} \;\lesssim\; C_P\delta_j^2|u-\Pi_j u|^2_{a,\widetilde{\Omega}_j^\circ} \;+\; \delta_j\|u-\Pi_j u\|^2_{0,\mathrm{tr}_j\alpha,\Gamma}\,, \tag{3.36}$$

It follows from the triangle inequality and (3.34) that

$$|u-\Pi_j u|^2_{a,\widetilde{\Omega}_j^\circ} \;\leq\; |u-\Pi_j u|^2_{a,\Omega_j} \;\lesssim\; |u|^2_{a,\Omega_j} \tag{3.37}$$

and so it only remains to bound $\|u-\Pi_j u\|^2_{0,\mathrm{tr}_j\alpha,\Gamma}$ with respect to $|u-\Pi_j u|_{a,\Omega_j}\leq|u|_{a,\Omega_j}$. This is where the particular choice of the DtN coarse space comes in.

The restriction of the functions $\{v_j^k\}_{k=1}^{n_\Gamma}$ to the boundary Γ forms a complete basis of $V_h(\Gamma)$. This implies that $\|u-\Pi_j u\|^2_{0,\mathrm{tr}_j\alpha,\Gamma} = \|\sum_{k=m_j+1}^{n_\Gamma} b_j\Big(v_j^k,u\Big)v_j^k\|^2_{0,\mathrm{tr}_j\alpha,\Gamma}$. It follows

from the fact that the functions $\{v_j^k\}_{k=1}^{n_\Gamma}$ are orthogonal in the $(\cdot,\cdot)_{0,\mathrm{tr}_j\alpha,\Gamma}$ inner product and that $\|v_j^k\|^2_{0,\mathrm{tr}_j\alpha,\Gamma}=1$ that

$$\begin{aligned}
\|u-\Pi_j u\|^2_{0,\mathrm{tr}_j\alpha,\Gamma} &= \sum_{k=m_j+1}^{n_\Gamma} b\Big(v_j^k,u\Big)^2\,\|v_j^k\|^2_{0,\mathrm{tr}_j\alpha,\Gamma}\\
&= \sum_{k=m_j+1}^{n_\Gamma} b\left(v_j^k,u\right)^2\\
&= \sum_{k=m_j+1}^{n_\Gamma} \frac{1}{\lambda_j^k} a\left(v_j^k,u\right) b\left(v_j^k,u\right)\\
&\le \frac{1}{\lambda_{(j)}^{m_j+1}} \sum_{k=m_j+1}^{n_\Gamma} a\left(v_j^k,u\right) b\left(v_j^k,u\right)\\
&= \frac{1}{\lambda_{(j)}^{m_j+1}} |u-\Pi_j u|^2_{a,\Omega_j}\\
&\le \frac{1}{\lambda_{(j)}^{m_j+1}} |u|^2_{a,\Omega_j} \qquad (3.38)
\end{aligned}$$

and the result follows from (3.36), (3.37) and (3.38). □

For any Ω_j, $1\le j\le N$, let Π_j be the projection onto the first m_j local DtN eigenvectors defined in (3.33) and let χ_j be the partition of unity function associated with Ω_j defined in (3.25). For a given function $u\in V_{h,0}$, we introduce the coarse interpolation of u as

$$u_0 := I_h\left(\sum_{j=1}^N \chi_j \Pi_j u|_{\Omega_j}\right)\in V_H\,. \qquad (3.39)$$

In the following, to ease the presentation when there is no confusion and it is clear from the context, we will simply denote the restriction $u|_{\Omega_j}$ of u onto Ω_j by u, and write, e.g., $\Pi_j u$ instead of $\Pi_j u|_{\Omega_j}$.

3.4.3 Stable splitting – Final convergence result

The next theorem is the main result needed to prove the robustness of the DtN coarse space construction. It states that Assumption 2.7 in the Abstract Schwarz framework is satisfied which, according to Remark 2.11, is the only challenge left in proving convergence for the two level additive Schwarz preconditioner.

Theorem 3.9. Let Assumptions 3.3 and 3.4 hold. Let $u\in V_{h,0}$ be given and let $u_0\in V_H$ be the coarse interpolation of u, defined in (3.39). Then there exists a stable splitting

$$u \;=\; \sum_{j=0}^N u_j \quad\text{and}\quad \sum_{j=0}^N \|u_j\|_a^2 \;\lesssim\; \max_{j=1}^N\{c_j(m_j)\}\,\|u\|_a^2\,,$$

with $u_j\in V_{h,0}(\Omega_j)$, $j=1,\dots,N$. The constants $c_j(m_j)$ are defined as in Theorem 3.8 by $c_j(m_j):=C_P^2+\left(\delta_j\lambda_{m_j+1}^{(j)}\right)^{-1}$.

Proof. For any $j \in \{1, \dots, N\}$, we choose

$$u_j := I_h(\widetilde{\chi}_j(u - u_0)).$$

Then, since by definition $\sum_{j=1}^{N} \widetilde{\chi}_j \equiv 1$ and $I_h(u - u_0) = u - u_0$ it is clear that

$$\sum_{j=1}^{N} u_j \;=\; I_h\left(\sum_{j=1}^{N} \widetilde{\chi}_j(u - u_0)\right) \;=\; u - u_0\,.$$

Each point belongs to k_0 subdomains at most so

$$\|u_0\|_a^2 \;\lesssim\; \|u\|_a^2 + \sum_{j=1}^{N} \|u_j\|_a^2$$

and so it suffices to bound the sum of the local energies.

Since the interpolator I_h is stable with respect to the a-norm (cf. [99, Lemma 2.3]),

$$\begin{aligned} \|u_j\|_a^2 \;\lesssim\; \left\|\widetilde{\chi}_j(u - u_0)\right\|_a^2 \;&\lesssim\; \|\widetilde{\chi}_j\|_\infty^2 |u - u_0|_{a,\widetilde{\omega}_j}^2 + \|\nabla\widetilde{\chi}_j\|_\infty^2 \|u - u_0\|_{0,\alpha,\widetilde{\omega}_j^\circ}^2 \\ &\lesssim\; |u - u_0|_{a,\widetilde{\omega}_j}^2 + \delta_j^{-2} \|u - u_0\|_{0,\alpha,\widetilde{\omega}_j^\circ}^2 \end{aligned} \tag{3.40}$$

where we denote $\widetilde{\omega}_j :=$ interior(supp($\widetilde{\chi}_j$)) and $\widetilde{\omega}_j^\circ :=$ interior(supp($\nabla\widetilde{\chi}_j$)) (see Figure 3.7 for a sketch).

Now, since I_h is also stable with respect to the weighted L_2-norm (cf. [99, Lemma 2.3]), using the definition of u_0 and the fact that $\text{supp}(\chi_i) \subset \overline{\Omega}_i$ and $\widetilde{\omega}_j^\circ \cap \Omega_i \subset \widetilde{\Omega}_i^\circ$ we get from Theorem 3.8 that

$$\begin{aligned} \|u - u_0\|_{0,\alpha,\widetilde{\omega}_j^\circ}^2 \;&\lesssim\; \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} \|\chi_i(u - \Pi_i u)\|_{0,\alpha,\widetilde{\omega}_j^\circ}^2 \\ &\lesssim\; \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} \|\chi_i\|_\infty^2 \|u - \Pi_i u\|_{0,\alpha,\widetilde{\omega}_j^\circ \cap \Omega_i}^2 \;\lesssim\; \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} \delta_i^2 c_i(m_i) |u|_{a,\Omega_i}^2 \end{aligned} \tag{3.41}$$

Similarly,

$$\begin{aligned} |u - u_0|_{a,\widetilde{\omega}_j}^2 \;&\lesssim\; \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} |\chi_i(u - \Pi_i u)|_{a,\widetilde{\omega}_j}^2 \\ &\lesssim\; \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} \|\chi_i\|_\infty^2 |u - \Pi_i u|_{a,\widetilde{\omega}_j \cap \Omega_i}^2 + \|\nabla\chi_i\|_{\infty,\,\widetilde{\omega}_j \cap \Omega_i^\circ}^2 \|u - \Pi_i u\|_{0,\alpha,\widetilde{\omega}_j \cap \Omega_i^\circ}^2 \end{aligned}$$

and since $(\widetilde{\omega}_j \cap \Omega_i^\circ) \subset \widetilde{\Omega}_i^\circ$ we have, again using Theorem 3.8,

$$|u - u_0|_{a,\widetilde{\omega}_j}^2 \lesssim \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} |u|_{a,\widetilde{\omega}_j \cap \Omega_i}^2 + |\Pi_i u|_{a,\widetilde{\omega}_j \cap \Omega_i}^2 + \delta_i^{-2} \|u - \Pi_i u\|_{0,\alpha,\widetilde{\Omega}_i^\circ}^2 \lesssim \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} c_i(m_i) |u|_{a,\Omega_i}^2. \tag{3.42}$$

Substituting (3.41) and (3.42) into (3.40) and using the facts that $\delta_j \eqsim \delta_l$ for two neighbouring domains and that each point is contained in at most k_0 subdomains, we finally get

$$\sum_{j=1}^{N} \|u_j\|_a^2 \;\lesssim\; \sum_{j=1}^{N} \sum_{i:\Omega_i \cap \Omega_j \neq \emptyset} c_i(m_i) |u|_{a,\Omega_i}^2 \;\lesssim\; \max_{i=1}^{N}\{c_i(m_i)\} \|u\|_a^2\,,$$

which completes the proof. □

As usual, the existence of a stable splitting established in Theorem 3.9 is sufficient to deduce the following bound on the condition number of $M_{AS,2}^{-1}A$ from the abstract Schwarz theory (see e.g. [112]).

Theorem 3.10. Let Assumptions 3.3 and 3.4 hold. Then the condition number of the two-level Schwarz algorithm with a coarse space based on the spectra of local DtN maps can be bounded by

$$\kappa(M_{AS,2}^{-1}A) \lesssim \max_{j=1}^{N}\{c_j(m_j)\} \lesssim \left(C_P^2 + \max_{j=1}^{N} \frac{1}{\delta_j \lambda_{m_j+1}^{(j)}}\right).$$

The hidden constant is independent of h, δ_j, and $\mathrm{diam}(\Omega_j)$, as well as of any jumps in α.

We have only analyzed the additive preconditioner, but we note that other symmetric versions (in particular the balanced preconditioner in (3.11)) can be analyzed in the same way (cf. [112]). The restricted additive Schwarz (RAS) [9] variant is different since it leads to a non symmetric iteration. However, it behaves in a similar way and even gives slightly better results than the classical additive version above. Numerical tests with the DtN coarse space in the next section will confirm this.

Remark 3.11. By choosing the number m_j of modes per subdomain in the coarse space as prescribed in Algorithm 3.1 we ensure that $\lambda_{m_j+1}^{(j)} \geq \mathrm{diam}(\Omega_j)^{-1}$ so

$$\kappa(M_{AS,2}^{-1}A) \lesssim \left(C_P^2 + \max_{j=1}^{N} \frac{\mathrm{diam}(\Omega_j)}{\delta_j}\right).$$

Provided the weighted Poincaré constant C_P in Assumption 3.4 is uniformly bounded, independently of any jumps in the coefficients, we retrieve the classical estimate for the Additive Schwarz Method. An interesting observation is that the bound depends only in an additive way on the constant C_P and on the ratio of subdomain diameter to overlap. Note also that due to the small overlap "trick" in Lemma 3.7 (and contrary to the results in [38, 39]) the bound in Theorem 3.10 only depends on δ_j^{-1} and not on δ_j^{-2}.

3.5 Numerical results

For a first illustration of the performance of the DtN coarse space we refer the reader to the introduction (Section 2.3.1). There, in the case of a long domain with alternating layers we observed that the DtN coarse space makes it possible to achieve both scalability and robustness with respect to the jumps in the coefficient. This required adding as many vectors per subdomain to the coarse space as the number of high coefficient layers: applying the automatic selection of coarse modes (Algorithm 3.1) led to the optimal coarse space.

From now on we solve the model problem

$$\begin{cases} -\nabla \cdot (\alpha \nabla u_*) &= 1 \quad \text{in } \Omega, \\ u_* &= 0 \quad \text{on } \partial\Omega_D, \\ \frac{\partial u_*}{\partial \mathbf{n}} &= 0 \quad \text{on } \partial\Omega_N, \end{cases} \tag{3.43}$$

where $\Omega = [0,1]^2$ is the unit square and its boundary is divided into $\partial\Omega_D$ (D for Dirichlet) and $\partial\Omega_N$ (N for Neumann). The coefficient α varies within Ω and we will give its definition when we describe each test case.

A regular simplicial mesh of Ω with $n_{nodes} \times n_{nodes}$ nodes is given and in all but the last case (3.43) is approximated by standard linear ($\mathbb{P}_1$) finite elements. We use partitions of Ω into $N \times N$ overlapping subdomains which are obtained by adding n_{layers} of elements to a non overlapping partition. We distinguish two cases: a partition into $N \times N$ regular subdomains or a partition into $N \times N$ non regular subdomains obtained using the automatic graph partitioner Metis [54].

We make three things vary in the solver:

- The one level preconditioner is either the Additive Schwarz preconditioner (AS) introduced in (3.6) or the Restricted Additive Schwarz preconditioner (RAS) defined in (3.13). In the first case the iterative solver is conjugate gradient and in the second it is GMRES. To make the comparison fair the stopping criterion will always be based on the error in the infinity norm

$$\frac{\|u_* - u_m\|_\infty}{\|u_*\|_\infty} < 10^{-6}.$$

- The coarse space correction is based on (3.10) (two level additive preconditioner) or (3.11) (balanced preconditioner).
- The coarse space is either
  - empty, in which case we are considering a one level method,
  - the Partition of unity coarse space [97] (one vector per subdomain) referred to as POU,
  - the DtN coarse space (introduced in this chapter).

The corresponding discretizations and data structures are obtained using the software FreeFem++ [50] in connection with the Metis graph partitioner [54]. Every time a condition number for the preconditioned operator is given it is in fact the estimate given by the extreme Ritz values at the final iteration (see *e.g.* [15]).

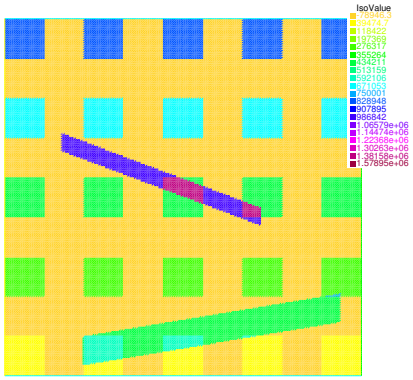
3.5.1 Influence of the partition

The boundary conditions are zero-Dirichlet on the entire boundary ($\partial\Omega_D = \partial\Omega$ and $\partial\Omega_N = \emptyset$). The mesh consists of 257×257 nodes and each subdomain is obtained by adding 2 layers of elements to each subdomain. Figure 3.9 shows both the problem setting and the convergence result for the test case. The coefficient α takes values between 1 and approximately 1.5×10^6 and the distribution contains both inclusions and channels. We consider 2×2, 4×4 and 8×8 subdomain partitions of Ω both regular and obtained with Metis. We compare the three types of coarse spaces and the AS and RAS preconditioners. The coarse correction is computed via the balanced preconditioner. As an illustration, we have chosen to present more extensively the 4×4 subdomain cases. In all cases the automatic selection process picks up no more than four vectors per subdomains. We observe that convergence with our new coarse space requires significantly fewer iterations and that it is robust (iteration counts vary between 22 and 31 for two level AS and 14 and 23 for two level RAS). As expected, even though we cannot prove it theoretically, convergence is best with RAS.

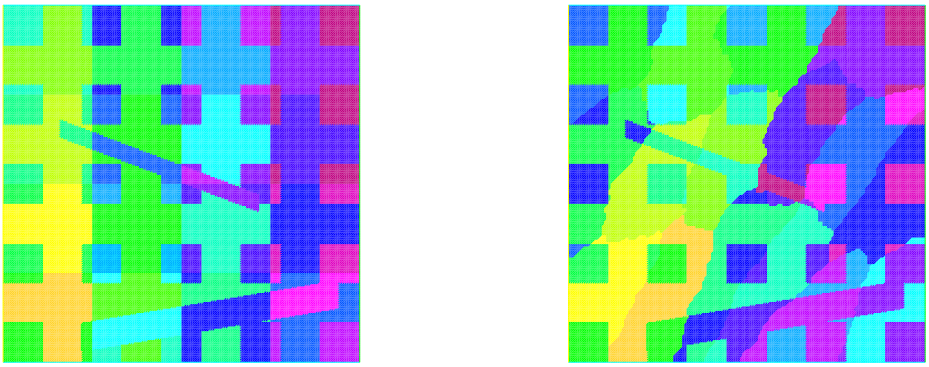
3.5.2 Channels

This time the boundary condition is $u = 0$ on the left hand side boundary $\partial\Omega_D$ and $\frac{\partial u}{\partial \mathbf{n}} = 0$ on the remainder $\partial\Omega_N$. The mesh consists of 161×161 nodes and each subdomain is obtained by adding 1 layer of elements to a non overlapping partition.

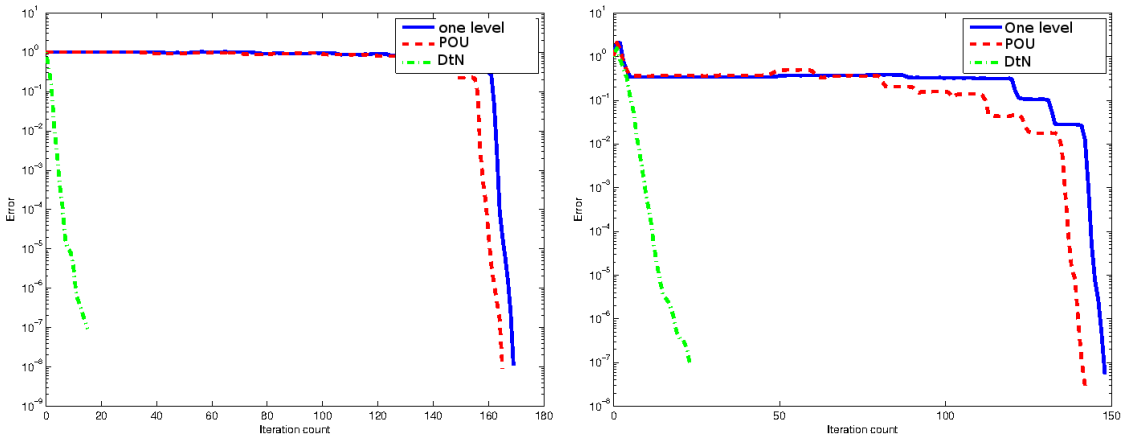
Coefficient distribution:



Partitions into 4×4 uniform (left) and Metis (right) subdomains - This shows both the subdomain boundaries and the coefficient distribution:



RAS convergence for all three choices of the coarse space and the balanced preconditioner, 4×4 regular subdomains (left) and Metis subdomains (right):



Number of iterations required to achieve convergence for various partitions:

	AS	AS+POU	AS+DtN	RAS	RAS+POU	RAS+DtN
2×2	103	110	22	70	70	14
2×2 with Metis	76	76	22	57	57	18
4×4	603	722	26	169	165	15
4×4 with Metis	483	425	36	148	142	23
8×8	461	141	34	205	95	21
8×8 with Metis	600	542	31	240	196	19

Figure 3.9: Test case in Subsection 3.5.1: Geometry and convergence results

The geometry for the test problem is given if Figure 3.10. The partition into subdomains is the 4×4 Metis partition. The diffusion coefficient α contains both high-permeability inclusions and channels. When there are no channels, α varies between 1 and 10^6. With all three channels present, α varies between 1 and 2.8×10^6. We analyze the performance of the method by increasing the number of channels.

All results are reported in Figure 3.11. Our algorithm performs significantly better than the classical methods. The Partition of unity coarse space has virtually no effect on the performance of either AS or RAS, leading to iteration numbers that differ little from the results without any coarse grid in all four cases. Our new coarse space, on the other hand, is fully robust with respect to the coefficient variations and it leads to a gain of at least a factor 6 compared to the one-level method in all cases. The condition number is bounded independently of the coefficient variation. The situation is even more pronounced, if we use the balanced two level preconditioners: the gain is more than a factor 10 in all cases.

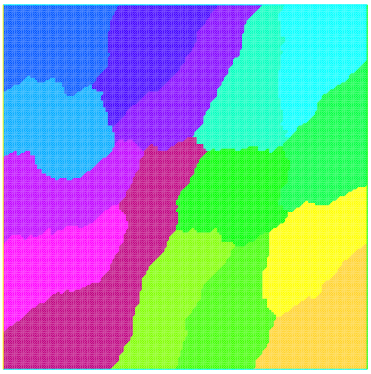
As well as the convergence results we give some information on the size of the coarse space that we build using our automatic selection strategy: for each number of channels we give $\min_j m_j$ and $\max_j m_j$, as well as the global coarse space size $n_H = \sum_j m_j$ and the average number of modes included per subdomain n_H/N. For comparison, we also include information on the total number $n_{\Gamma j}$ of eigenmodes of the discrete DtN operator on each subdomain. We note that adding a small number of channels does not seem to have any significant influence on the size of the coarse space: the difference is less than 10% between the case of three channels and no channel. To sum up Test Problem 1: by using about 3 eigenvectors on average per subdomain we have reduced the condition numbers from $O(10^7)$ to $O(10 - 100)$ in all four test cases.

The last series of tests in Figure 3.11 aims to prove that the number m_j of eigenvectors per subdomain chosen by our automatic algorithm is indeed optimal in some sense. For the test problem with one channel, we first reduce the number of coarse basis functions per subdomain by one, this has a huge influence on the iteration count and the condition number: it goes from $7.7 \cdot 10^1$ to $4 \cdot 10^6$ so by removing one coarse vector per subdomain we've essentially ruined the robustness. Then we add one basis function per subdomain and notice that this has much less effect, with the condition number decreasing only from $7.7 \cdot 10^1$ to $4.0 \cdot 10^1$. This suggests that the selection process we have designed is indeed in some sense an optimal compromise between robustness and size of the coarse space.

3.5.3 Large Inclusions

Now, using the same domain and the same partition we successively add inclusions without any channels present, as shown in Figure 3.12. The results are also presented in Figure 3.12. Again, the Partition of unity coarse space is ineffective for this test problem. The DtN coarse space, on the other hand, is robust to an increase in the number of inclusions and once more requires significantly fewer iterations than the one-level method in all cases. Note that the subdomain partition is not aligned with the inclusions at all. We see that for this test problem also, the coarse space size grows only slowly with the number of inclusions (i.e. roughly by a factor 2 when the number of inclusions has grown by a factor 9), and even in the hardest test case with 36 inclusions, n_H is only 44 (compared to the global dimension n of $V_{h,0}$, and thus of A, which is 25600). As in Test Problem 1, by using on average less than three eigenvectors per subdomain, we have reduced the condition numbers from $O(10^7)$ to $O(10 - 100)$ in all cases.

Partition into 16 subdomains using Metis:



Coefficient distribution – we add channels:

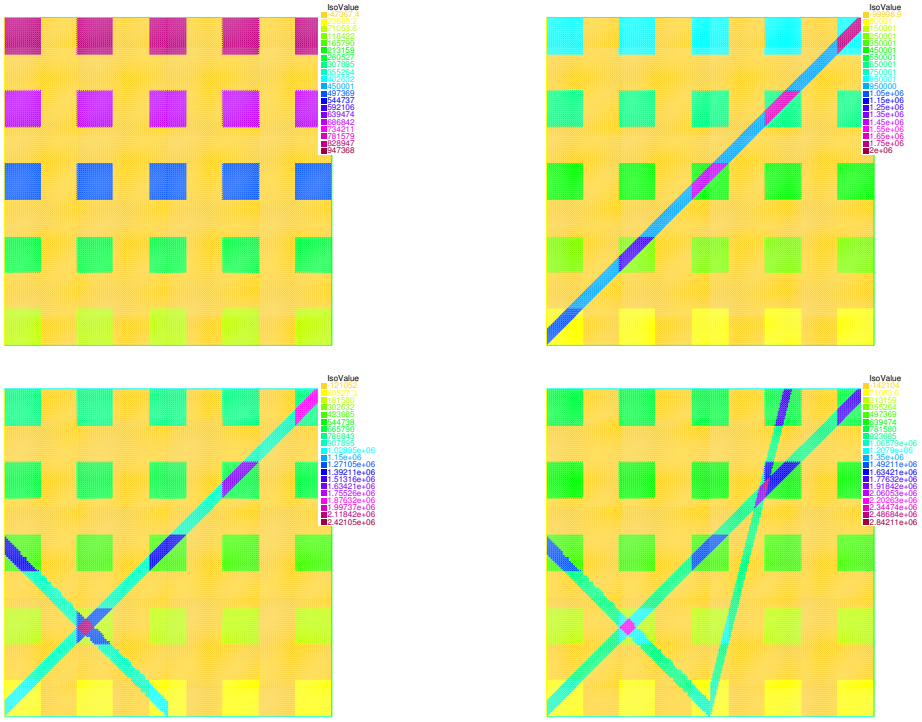


Figure 3.10: Test Problem from subsection 3.5.2: Geometry (see Figure 3.11 for the results)

Number of iterations and condition number (in brackets)
with the **additive** coarse grid correction:

	AS – iteration count (condition number)			RAS – iteration count		
	1−level	POU	DtN	1−level	POU	DtN
no channel	385 $(6.0 \cdot 10^7)$	393 $(6.0 \cdot 10^7)$	42 $(6.1 \cdot 10^1)$	264	255	41
1 channel	430 $(1.2 \cdot 10^7)$	454 $(8.2 \cdot 10^6)$	44 $(7.7 \cdot 10^1)$	243	246	36
2 channels	479 $(1.2 \cdot 10^7)$	499 $(8.6 \cdot 10^6)$	43 $(8.1 \cdot 10^1)$	237	240	42
3 channels	460 $(1.1 \cdot 10^7)$	470 $(8.4 \cdot 10^6)$	46 $(7.1 \cdot 10^1)$	232	234	38

with the **balanced** coarse grid correction:

	AS – iteration count (condition number)			RAS – iteration count		
	1−level	POU	DtN	1−level	POU	DtN
no channel	385 $(6.0 \cdot 10^7)$	349 $(7.2 \cdot 10^6)$	25 $(1.9 \cdot 10^1)$	264	237	20
1 channel	430 $(1.2 \cdot 10^7)$	419 $(5.8 \cdot 10^6)$	29 $(3.6 \cdot 10^1)$	243	232	21
2 channels	479 $(1.2 \cdot 10^7)$	423 $(5.9 \cdot 10^6)$	29 $(3.7 \cdot 10^1)$	237	227	21
3 channels	460 $(1.1 \cdot 10^7)$	433 $(5.8 \cdot 10^6)$	29 $(3.3 \cdot 10^1)$	232	220	20

Size of the coarse space :

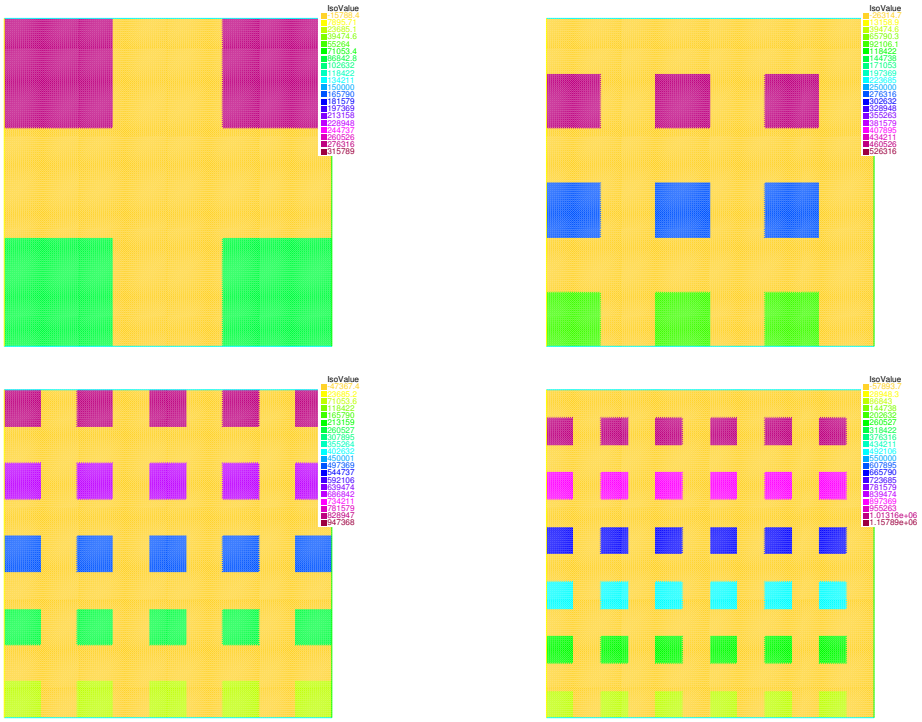
	Total number n_{Γ_j} of eigenvalues on Γ_j	Number m_j of functions included in V_H from Ω_j			
		no channel	1 channel	2 channels	3 channels
Minimum	71	1	1	1	1
Maximum	207	4	4	4	4
Average	143	2.75	2.75	2.88	3
Sum	2280	44	44	46	48

"Optimality" of the automatic selection strategy for m_j, the number of coarse basis functions per subdomain:

AS with additive coarse space correction	iterations	condition number
DtN space with $\max\{m_j - 1, 1\}$ functions from Ω_j	409	$4.6 \cdot 10^6$
DtN space with m_j functions from Ω_j	**44**	$\mathbf{7.7 \cdot 10^1}$
DtN space with $m_j + 1$ functions from Ω_j	35	$4.0 \cdot 10^1$

Figure 3.11: Test Problem from subsection 3.5.2: Results (see Figure 3.10 for the geometry)

We successively add inclusions to the coefficient distribution:



Number of iterations and Condition number (in brackets) (with the **additive** coarse grid correction):

	AS – iteration count (condition number)			RAS – iteration count		
	1−level	POU	DtN	1−level	POU	DtN
2×2 incl.	107 $(2.9 \cdot 10^7)$	82 $(2.2 \cdot 10^6)$	43 $(8.8 \cdot 10^1)$	109	87	41
3×3 incl.	184 $(2.8 \cdot 10^7)$	185 $(3.8 \cdot 10^6)$	47$(8.4 \cdot 10^1)$	164	150	45
5×5 incl.	385 $(6.0 \cdot 10^7)$	393 $(1.4 \cdot 10^7)$	42 $(6.1 \cdot 10^1)$	264	255	41
6×6 incl.	425 $(5.1 \cdot 10^7)$	475 $(8.5 \cdot 10^6)$	46 $(1.0 \cdot 10^2)$	262	248	44

Size of the coarse space:

	Total number n_{Γ_j} of eigenvalues on Γ_j	Number m_j of functions included in V_H from Ω_j			
		2×2 inc.	3×3 inc.	5×5 inc.	6×6 inc.
Minimum	71	1	1	1	1
Maximum	207	3	4	4	5
Average	143	1.43	1.87	2.75	2.75
Sum	2280	23	30	44	44

Figure 3.12: Test problem from Subsection 3.5.3: Geometry and Results

3.5.4 Scalability test on a parallel architecture

The implementation for these test cases is the work of Pierre Jolivet. The detailed techniques are presented in [52].

Two dimensional test case The model problem is solved on $\Omega = [0;1]^2$ with mixed Dirichlet and Neumann boundary conditions using $\mathbb{P}_2$ finite elements. The diffusivity is a highly heterogeneous function of $\Omega \to \mathbb{R}$, *c.f.* Figure 3.13, defined as:

$$\alpha(x,y) = \begin{cases} 10^5(\lfloor 9x \rfloor + 1) & \text{if } \lfloor 9x \rfloor \equiv 0 \bmod 2 \text{ and } \lfloor 9y \rfloor \equiv 0 \bmod 2 \\ 1 & \text{otherwise} \end{cases}$$

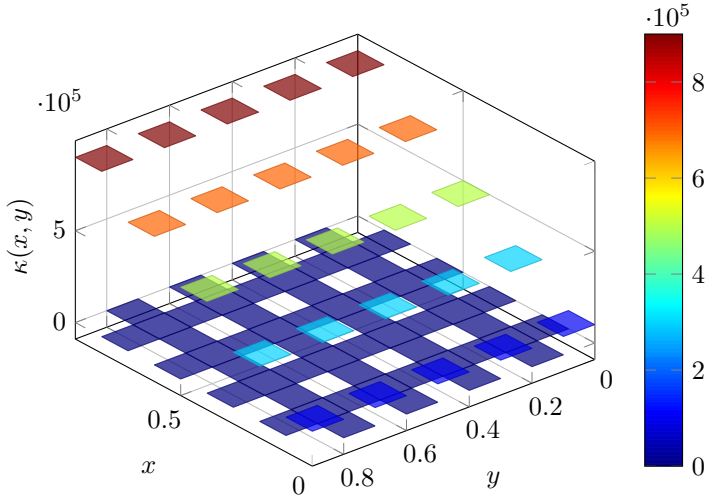


Figure 3.13: Two dimensional diffusivity α in the scalability test cases

Three dimensional test case The same model problem as for the two dimensional test case is solved once again with mixed Dirichlet and Neumann boundary conditions on $\Omega = [0;1]^3$ using $\mathbb{P}_2$ finite elements. The diffusivity is defined as:

$$\alpha(x,y,z) = \begin{cases} 10^5(\lfloor 9x \rfloor + 1)\alpha_\uparrow(z) & \text{if } \lfloor 9x \rfloor \equiv 0 \bmod 2 \text{ and } \lfloor 9y \rfloor \equiv 0 \bmod 2 \\ 1 & \text{otherwise} \end{cases}$$

where

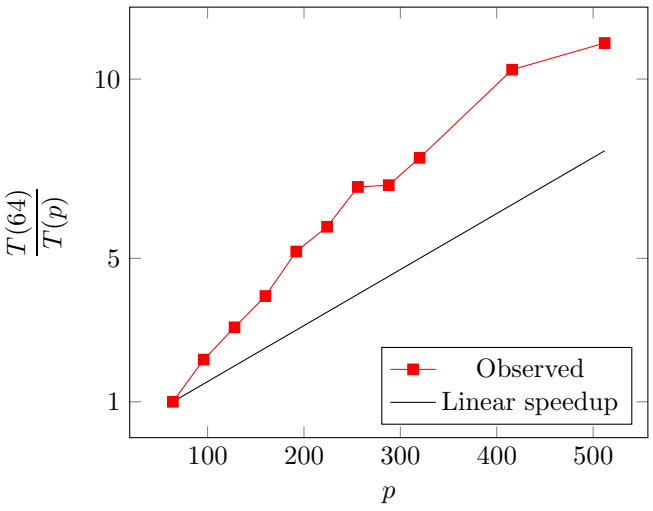
$$\alpha_\uparrow(z) = \begin{cases} \lfloor 9z \rfloor & \text{if } \lfloor 9z \rfloor \not\equiv 0 \bmod 3 \\ 1 & \text{otherwise} \end{cases}$$

Results In order to assess the performance of the implementation of the parallel solver and the scalability of the two level solver with DtN, a speedup test is performed by solving the same problem with different numbers of processors (and hence of subdomains). The results are reported in Figures 3.14 (speedup normalized to 64 subdomains) and 3.15 (speedup normalized to 96 subdomains). A superlinear speedup is observed both in the two and three dimensional cases. These particular tests were performed on `titane`, a 40960-core computer hosted at CEA [2].

2. Commissariat à l'Energie Atomique et aux Energies Alternatives, Bruyères-le-Châtel, France

In the tables, all the timings are obtained using the routine `MPI_Wtime()`. Only the GMRES is timed, meaning that the construction of the meshes, the partitioning of unity and the construction of the coarse operator A_H are not considered. The stopping criterion is chosen so that the relative residual of the GMRES is inferior to a certain tolerance ε at convergence. The first way to assess the performance of the implementation of our parallel solver was done by checking its speedup when increasing the number of processes.

More tests were performed on `babel`, a 121912-core computer hosted at IDRIS [3] and even when greatly increasing the number of subdomains, the Krylov method still converges quite quickly in terms of number of iterations, i.e. in less than 25 iterations for partitions into as many as 4096 subdomains.



p	T	$\#V_H$	$\#V_h$
64	65.7 s	1890	36.5×10^6
128	21.4 s	3810	36.7×10^6
192	12.7 s	5730	36.9×10^6
256	9.4s	7650	37.1×10^6
320	8.4 s	9570	37.2×10^6
512	6.0 s	15330	37.5×10^6

Figure 3.14: Strong scalability observed when solving a two dimensional test case on a fixed size problem using $\mathbb{P}_2$ finite elements and a tolerance $\varepsilon = 10^{-9}$. In the table, p is the number of processors, T is the computation time, $\#V_H$ is the number of DtN coarse vectors and $\#V_h$ is the total number of unknowns counting the ones that are in the overlap multiple times.

Acknowledgments This work has been granted access to the HPC resources of CCRT and IDRIS under the allocation 2011-6654 made by GENCI through ANR project PETALh (ANR-10-COSI-013). The assistance of Philippe Wautelet from IDRIS played an important role in the completion of the scalability tests.

3. Institut du Développement et des Ressources en Informatique Scientifique, Orsay, France

p	T	$\#V_H$	$\#V_h$
96	26.5 s	1920	6.9×10^6
128	18.6 s	2560	7.1×10^6
160	13.0 s	3200	7.4×10^6
192	10.4s	3840	7.6×10^6
224	8.1 s	4480	7.7×10^6
288	6.9 s	5760	8.1×10^6

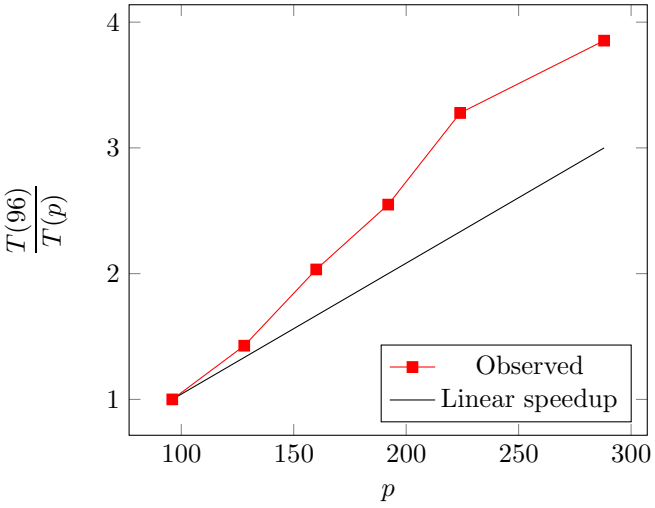


Figure 3.15: Strong scalability observed when solving a three dimensional test case on a fixed size problem using $\mathbb{P}_2$ finite elements and a tolerance $\varepsilon = 10^{-12}$. In the table p is the number of processors, T is the computation time, $\#V_H$ is the number of DtN coarse vectors and $\#V_h$ is the total number of unknowns counting the ones that are in the overlap multiple times.

Chapter 4

GenEO: a coarse space for the Additive Schwarz method

The content of this chapter was published in Numerische Mathematik [106] in collaboration with Victorita Dolean, Patrice Hauret, Frédéric Nataf, Clemens Pechstein and Robert Scheichl. We first presented the method and convergence result in the note [105]. The numerical results in the proceedings paper [107] have also been incorporated into the last section of this chapter. Finally in section 4.3.3 we present results for slightly modified preconditioners, these were first studied in the proceedings of the LSSC conference [104].

Contents

4.1 Introduction

Once more we work in the already extensively studied framework of the overlapping additive Schwarz preconditioner [102, 112], and focus on the definition of a suitable coarse

space with the aim to achieve robustness with respect to heterogeneities in any of the coefficients in the PDEs and the number of subdomains. In the previous chapter we proposed and studied the DtN coarse space for scalar elliptic problems. The proof for DtN relies on uniform (in the coefficients) weighted Poincaré inequalities [89]. While this allows for full robustness in the small overlap case (cf. [21]), in a completely general setting it has two drawbacks: (i) for larger overlap some assumptions are needed on the coefficient distribution in the overlaps and (ii) the arguments cannot be generalized easily to the case of systems of PDEs. This second point was the motivation to look for a new coarse space.

In this Chapter, we propose a coarse space construction based on Generalized Eigenproblems in the Overlap (which we will refer to as the *GenEO coarse space*). We define the coarse space, prove a convergence result and illustrate it with some numerical results. The coarse space construction applies to systems of PDEs discretized by finite elements with only a few extra assumptions. The implementation only relies on having access to element stiffness matrices and the connectivity graph between elements. The subdomain partition is carried out using Metis. Overlap is added based on the connectivity graph and the coarse space is constructed automatically solving a generalized eigenproblem on each subdomain. In our analysis, we identify the fact that the abstract Schwarz framework makes it possible to reduce the proof of convergence to an energy bound in the overlap, and for this reason, the second matrix in the pencil of our generalized eigenvalue problem is a matrix that has zero blocks corresponding to the interior of the subdomain.

The generalized eigenvalue problems which we solve here are closely related, but different to the ones proposed in [26]. The major theoretical advance with respect to [26] is that there, in order for the proof to go through for classical finite element spaces, a stable interpolation operator with a constant independent of the coefficients is needed. In many cases (elasticity for instance), such a stable interpolator does not yet exist to our knowledge. We overcame this problem by introducing partition of unity operators that work directly on the degrees of freedom bypassing the need for a stable interpolation operator. From a practical point of view, thanks to these partition of unity operators, the right hand side of the generalized eigenproblems can be constructed fully automatically from element stiffness matrices and diagonal weighting matrices. We only require access to some topological information (to build a suitable partition of unity), and to the element stiffness matrices (as in AMGe methods, cf. [12]). This is reasonable in standard finite element packages such as FreeFEM++ [50].

The rest of this Chapter is organized as follows. In Section 4.2 we define the problem that we solve and introduce the two-level additive Schwarz framework along with some elements of generalized eigenvalue problem theory. In Section 4.3 we define the abstract procedure to construct our coarse space and give the main convergence result (Theorem 4.33). Section 4.4 gives detailed guidelines on how to implement the two-level Schwarz preconditioner with the GenEO coarse space in a finite element code. Finally in Section 4.5 we test our method for Darcy and linear elasticity and make sure that it indeed converges robustly even for highly varying coefficients in two and three dimensions.

4.2 Preliminaries and notations

4.2.1 Problem Description

Given a Hilbert space V, a symmetric and coercive bilinear form $a : V \times V \to \mathbb{R}$ and an element f in the dual space V', we consider the abstract variational problem: Find

$v \in V$ such that

$$a(v,\, w) = \langle f, w\rangle, \quad \text{for all} \quad w \in V, \tag{4.1}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing. This variational problem is associated with an elliptic boundary value problem (BVP) on a given domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) with suitable boundary conditions posed in a suitable space V of functions on Ω.

We consider a discretization of the variational problem (4.1) with finite elements based on a mesh $\mathcal{T}_h$ of Ω:

$$\overline{\Omega} = \bigcup_{\tau\in\mathcal{T}_h} \tau.$$

Let $V_h \subset V$ denote the chosen conforming space of finite element functions. In the case where $a(\cdot,\cdot)$ is a bilinear form derived from a system of PDEs, V_h is a space of vector functions. The discretization of (4.1) then reads: Find $v_h \in V_h$ such that

$$a(v_h, w_h) = \langle f, w_h\rangle, \quad \text{for all} \quad w_h \in V_h. \tag{4.2}$$

Let $\{\phi_k\}_{k=1}^n$ be a basis for V_h with $n := \dim(V_h)$, then from (4.2) we can derive a linear system

$$\boldsymbol{A}\boldsymbol{v} \,=\, \boldsymbol{f}, \tag{4.3}$$

where the coefficients of the stiffness matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and the load vector $\boldsymbol{f} \in \mathbb{R}^n$ are given by $A_{k,l} = a(\phi_l, \phi_k)$ and $f_k = \langle f, \phi_k\rangle$ ($k, l = 1, \ldots, n$) and $\boldsymbol{v}$ is the vector of coefficients corresponding to the unknown finite element function v_h in (4.2).

The basis $\{\phi_k\}_{k=1}^n$ can be quite arbitrary but it should fulfil a unisolvence property: the basis functions supported on each element $\tau \in \mathcal{T}_h$ are linearly independent when restricted to τ. This is the case for standard finite element bases.

The only significant assumption we make on the problem is that the stiffness matrix $\boldsymbol{A}$ is assembled from positive semi-definite element stiffness matrices.

Assumption 4.1. Let $V_h(\tau) = \{v|_\tau : v \in V_h\}$. We assume that there exist positive semi-definite bilinear forms $a_\tau : V_h(\tau) \times V_h(\tau) \to \mathbb{R}$, for all $\tau \in \mathcal{T}_h$, such that

$$a(v, w) = \sum_{\tau\in\mathcal{T}_h} a_\tau(v|_\tau, w|_\tau), \quad \text{for all} \quad v, w \in V_h.$$

Remark 4.2. If the variational problem is obtained from integrating local forms on the domain then this is not a problem at all. For instance in the case of the Darcy equation we can write for all $v, w \in H_0^1(\Omega)$:

$$a(v, w) = \int_\Omega \kappa\boldsymbol{\nabla} v \cdot \boldsymbol{\nabla} w = \sum_{\tau\in\mathcal{T}_h} \int_\tau \kappa\boldsymbol{\nabla} v \cdot \boldsymbol{\nabla} w = \sum_{\tau\in\mathcal{T}_h} a_\tau(v|_\tau, w|_\tau).$$

4.2.2 Additive Schwarz setting

In order to automatically construct a robust two-level Schwarz preconditioner for (4.3), we first partition our domain Ω into a set of non-overlapping subdomains $\{\Omega_j'\}_{j=1}^N$ resolved by $\mathcal{T}_h$ using for example a graph partitioner such as METIS [54] or SCOTCH [13]. Each subdomain Ω_j' is then extended to a domain Ω_j by adding one or several layers of mesh elements in the sense of Definition 4.3, thus creating an overlapping decomposition $\{\Omega_j\}_{j=1}^N$ of Ω.

Definition 4.3. Given a subdomain $D' \subset \Omega$ which is resolved by $\mathcal{T}_h$, the extension of D' by one layer of elements is

$$D = \mathrm{Int}\Big(\bigcup_{\{k:\mathrm{supp}(\phi_k)\cap D' \neq \emptyset\}} \mathrm{supp}(\phi_k) \Big)$$

and $\mathrm{Int}(\cdot)$ denotes the interior of a domain. Extensions by more than one layer can then be defined recursively.

The proof of the following lemma is a direct consequence of Definition 4.3.

Lemma 4.4. For every degree of freedom k, with $1 \leq k \leq n$, there is a subdomain Ω_j, with $1 \leq j \leq N$, such that $\mathrm{supp}(\phi_k) \subset \overline{\Omega}_j$.

Now, for each $j = 1, \ldots, N$, let

$$V_h(\Omega_j) := \{v|_{\Omega_j} : v \in V_h\}$$

denote the space of restrictions of functions in V_h to Ω_j. Furthermore, let

$$V_{h,0}(\Omega_j) := \{v|_{\Omega_j} : v \in V_h,\ \mathrm{supp}\,(v) \subset \Omega_j\}$$

denote the space of finite element functions supported in Ω_j. By definition, the extension by zero of a function $v \in V_{h,0}(\Omega_j)$ to Ω lies again in V_h. We denote the corresponding extension operator by

$$R_j^\top : V_{h,0}(\Omega_j) \to V_h\,. \tag{4.4}$$

Lemma 4.4 guarantees that $V_h = \sum_{j=1}^N R_j^\top V_{h,0}(\Omega_j)$. The adjoint of $R_j^\top$

$$R_j : V_h' \to V_{h,0}(\Omega_j)'\,,$$

called the restriction operator, is defined by $\langle R_j g, v\rangle = \langle g, R_j^\top v\rangle$, for $v \in V_{h,0}(\Omega_j)$, $g \in V_h'$. However, for the sake of simplicity, we will often leave out the action of $R_j^\top$ and view $V_{h,0}(\Omega_j)$ as a subspace of V_h.

The final ingredient is a coarse space $V_H \subset V_h$ which will be defined later. Let $R_H^\top : V_H \to V_h$ denote the natural embedding and R_H its adjoint. Then the two-level additive Schwarz preconditioner (in matrix form) reads

$$\boldsymbol{M}_{AS,2}^{-1} = \boldsymbol{R}_H^T \boldsymbol{A}_H^{-1} \boldsymbol{R}_H + \sum_{j=1}^N \boldsymbol{R}_j^T \boldsymbol{A}_j^{-1} \boldsymbol{R}_j, \quad \boldsymbol{A}_H := \boldsymbol{R}_H \boldsymbol{A} \boldsymbol{R}_H^T \quad \text{and} \quad \boldsymbol{A}_j := \boldsymbol{R}_j \boldsymbol{A} \boldsymbol{R}_j^T, \tag{4.5}$$

where $\boldsymbol{R}_j$, $\boldsymbol{R}_H$ are the matrix representations of R_j and R_H with respect to the basis $\{\phi_k\}_{k=1}^n$ and the chosen basis of the coarse space V_H. As usual for standard elliptic BVPs, $\boldsymbol{A}_j$ corresponds to the original (global) system matrix restricted to subdomain Ω_j with Dirichlet conditions on the artificial boundary $\partial\Omega_j \setminus \partial\Omega$.

To simplify the notation, if D is the union of elements of $\mathcal{T}_h$ and

$$V_h(D) := \{v|_D : v \in V_h\},$$

we write, for any $v, w \in V_h(D)$,

$$a_D(v,w) := \sum_{\tau \in D} a_\tau(v|_\tau, w|_\tau) \quad \text{and} \quad |v|_{a,D} = \sqrt{a_D(v,v)},$$

where the latter is the energy seminorm. The definition of $a_D(\cdot,\cdot)$ extends naturally to $v, w \in V_h(D')$, for any $D \subset D' \subset \Omega$ which simplifies notations. On each of the local spaces $V_{h,0}(\Omega_j)$, the bilinear form $a_{\Omega_j}(\cdot,\cdot)$ is positive definite since

$$a_{\Omega_j}(v,w) = a(R_j^\top v, R_j^\top w), \quad \text{for all } v, w \in V_{h,0}(\Omega_j),$$

and because $a(\cdot,\cdot)$ is coercive on V. For the same reason, the matrix $\boldsymbol{A}_j$ in (4.5) is invertible. Hence, $|\cdot|_{a,\Omega_j}$ becomes a norm on $V_{h,0}(\Omega_j)$ and so we write

$$\|v\|_{a,\Omega_j} = \sqrt{a_{\Omega_j}(v,v)}, \quad \text{for all } v \in V_{h,0}(\Omega_j).$$

If $D = \Omega$, we omit the domain from the subscript and write $\|\cdot\|_a$ instead of $\|\cdot\|_{a,\Omega}$.

We use here the abstract framework for additive Schwarz (see [112, Chapter 2] or Section 2.2.1 of this manuscript). In the following we summarize the most important ingredients.

Definition 4.5. We define $k_0 = \max_{\tau\in\mathcal{T}_h} \Big(\#\{\Omega_j : 1 \le j \le N, \tau \subset \Omega_j\}\Big)$.

This means that each point in Ω belongs to at most k_0 of the subdomains Ω_j.

Lemma 4.6. With k_0 as in Definition 4.5, the largest eigenvalue of $\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}$ satisfies

$$\lambda_{max}(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \;\le\; k_0 + 1.$$

Proof. See, e.g., [25, Section 4]. □

Definition 4.7 (Stable decomposition)**.** Given a coarse space $V_H \subset V_h$, local subspaces $\{V_{h,0}(\Omega_j)\}_{1\le j\le N}$ and a constant C_0, a C_0-stable decomposition of $v \in V_h$ is a family of functions $\{z_j\}_{0\le j\le N}$ that satisfies

$$v = \sum_{j=0}^{N} R_j^\top z_j\,, \qquad \text{with} \quad z_0 \in V_H\,, \quad z_j \in V_{h,0}(\Omega_j), \text{ for } j \ge 1, \tag{4.6}$$

and

$$\|z_0\|_a^2 + \sum_{j=1}^{N} \|z_j\|_{a,\Omega_j}^2 \le C_0^2\,\|v\|_a^2\,. \tag{4.7}$$

Theorem 4.8. If every $v \in V_h$ admits a C_0-stable decomposition (with uniform C_0), then the smallest eigenvalue of $\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}$ satisfies

$$\lambda_{min}(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \;\ge\; C_0^{-2}.$$

Therefore, the condition number of the two-level Schwarz preconditioner (4.5) can be bounded by

$$\kappa(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \;\le\; C_0^2(k_0+1).$$

Proof. The statement is a direct consequence of [112, Lemma 2.5] and Lemma 4.6. □

In the following, we will construct a C_0-stable decomposition in a specific framework, but prior to that we will provide in an abstract setting, a sufficient and simplified condition of stability.

Lemma 4.9. Using the notations introduced in Definition 4.7, if there exists a constant C_1 such that

$$\|z_j\|^2_{a,\Omega_j} \leq C_1 |v|^2_{a,\Omega_j}, \quad \text{for all} \quad j = 1, \ldots, N, \tag{4.8}$$

then the decomposition (4.6) is C_0-stable with $C_0^2 = 2 + C_1 k_0 (2k_0 + 1)$ where k_0 is given in Definition 4.5.

Proof. From (4.8) and Definition 4.5 we get successively

$$\sum_{j=1}^{N} \|z_j\|^2_{a,\Omega_j} \leq C_1 \sum_{j=1}^{N} |v|^2_{a,\Omega_j} \leq C_1 k_0 \|v\|^2_a . \tag{4.9}$$

We also have:

$$\|z_0\|^2_a = \left\| v - \sum_{j=1}^{N} z_j \right\|^2_a \leq 2\|v\|^2_a + 2 \left\| \sum_{j=1}^{N} z_j \right\|^2_a , \tag{4.10}$$

and from Definition 4.5 and (4.9) we get

$$\left\| \sum_{j=1}^{N} z_j \right\|^2_a \leq k_0 \sum_{j=1}^{N} \|z_j\|^2_{a,\Omega_j} \leq C_1 k_0^2 \|v\|^2_a . \tag{4.11}$$

Using (4.11) in (4.10) yields

$$\|z_0\|^2_a \leq 2(1 + C_1 k_0^2) \|v\|^2_a . \tag{4.12}$$

By adding (4.9) and (4.12) we get (4.7) with $C_0^2 = 2 + C_1 k_0 (2k_0 + 1)$.

□

When $\|z_0\|^2_a$ can be bounded directly in terms of $\|v\|^2_a$ (independently of the coefficient variation), this lemma is superfluous and leads to a suboptimal quadratic dependence on k_0. In general, however, it is not possible to provide such a uniform bound on $\|z_0\|^2_a$, which is why Lemma 4.9 is in fact absolutely crucial for our analysis.

4.2.3 Abstract generalized eigenproblems

In order to construct the coarse space we will use generalized eigenvalue problems in each subdomain. Since several variations of generalized eigenvalue problems exist in the literature (particularly concerning the interpretation of the 'infinite eigenvalue'), we state the definition that we use. It is in agreement with the matrix counterpart in Definition 2.12 of this manuscript.

Definition 4.10 (Generalized eigenvalue problem)**.** Let $\widetilde{V}$ be a finite-dimensional Hilbert space, let $\widetilde{a} : \widetilde{V} \times \widetilde{V} \to \mathbb{R}$ and $\widetilde{b} : \widetilde{V} \times \widetilde{V} \to \mathbb{R}$ be two symmetric bilinear forms. Then the generalized eigenvalues associated with the so called 'pencil' $(\widetilde{a}, \widetilde{b})$ are the following values $\lambda \in \mathbb{R} \cup \{+\infty\}$: either $\lambda \in \mathbb{R}$ and there exists $p \in \widetilde{V} \backslash \{0\}$ such that

$$\widetilde{a}(p, v) = \lambda \widetilde{b}(p, v), \quad \text{for all } v \in \widetilde{V}, \tag{4.13}$$

or $\lambda = +\infty$ and there exists $p \in \widetilde{V} \backslash \{0\}$ such that

$$\widetilde{b}(p, v) = 0, \quad \text{for all } v \in \widetilde{V}, \quad \text{and} \quad \widetilde{a}(p, v) \neq 0, \quad \text{for a certain} \quad v \in \widetilde{V}.$$

In both cases p is called a generalized eigenvector associated with the eigenvalue λ.

The definition above allows for infinite eigenvalues. This results from the fact that if $(+\infty, p)$ is an eigenpair for the pencil $(\widetilde{a}, \widetilde{b})$ then $(0, p)$ is an eigenpair for the pencil $(\widetilde{b}, \widetilde{a})$ and there is no reason to discriminate between both formulations. In cases where the bilinear form $\widetilde{b}$ is positive definite, the problem can be simplified and crucial properties on the eigenvalues and eigenvectors arise. In particular, it leads quite naturally to optimal projectors onto subspaces of the functional space, as the next lemma shows in an abstract setting.

Lemma 4.11. Let $\tilde{a}$ be positive semi-definite and $\tilde{b}$ positive definite, and let the eigenpairs $\{(p^k, \lambda_k)\}_{k=1}^{\dim(\widetilde{V})}$ of the generalized eigenvalue problem (4.13) be ordered such that

$$0 \leq \lambda_1 \leq \ldots \leq \lambda_{\dim(\widetilde{V})} \quad \text{and} \quad \widetilde{b}(p^k, p^l) = \delta_{kl}\,, \quad \text{for any} \quad 1 \leq k, l \leq \dim(\widetilde{V}),$$

where δ_{kl} denotes the Kronecker delta. Then, for any integer $1 \leq m < \dim(\widetilde{V})$, the projection

$$\widetilde{\Pi}_m v := \sum_{k=1}^{m} \widetilde{b}(v, p^k) p^k$$

satisfies

$$|\widetilde{\Pi}_m v|_{\widetilde{a}} \leq |v|_{\widetilde{a}} \quad \text{and} \quad |v - \widetilde{\Pi}_m v|_{\widetilde{a}} \leq |v|_{\widetilde{a}}, \quad \text{for all } v \in \widetilde{V}. \tag{4.14}$$

Additionally, if m is such that $\lambda_{m+1} > 0$, we have the stability estimate

$$\|v - \widetilde{\Pi}_m v\|_{\widetilde{b}}^2 \;\leq\; \frac{1}{\lambda_{m+1}} |v - \widetilde{\Pi}_m v|_{\widetilde{a}}^2, \quad \text{for all } v \in \widetilde{V}.$$

Proof. Due to the additional assumptions on $\widetilde{a}$ and $\widetilde{b}$, the generalized eigenvalue problem can be simplified to a standard eigenvalue problem, for which the existence of eigenvectors $\{p^k\}_{k=1}^{\dim(\widetilde{V})}$ with associated non-negative real eigenvalues $\{\lambda_k\}_{k=1}^{\dim(\widetilde{V})}$ is guaranteed by standard spectral theory. Moreover, $\{p^k\}_{k=1}^{\dim(\widetilde{V})}$ can be chosen such that it is a basis of $\widetilde{V}$ fulfilling the conditions:

$$\widetilde{a}(p^k, p^l) = \widetilde{b}(p^k, p^l) = 0 \quad \forall\, k \neq l, \quad |p^k|_{\widetilde{b}}^2 = 1 \quad \text{and} \quad |p^k|_{\widetilde{a}}^2 = \lambda_k.$$

The proof of this result, in matrix formulation is given in the proof of Lemma 2.13. Now let $v \in \widetilde{V}$ be fixed. From the $\widetilde{b}$-orthonormality of the basis we get

$$v = \sum_{k=1}^{\dim(\widetilde{V})} \widetilde{b}(v, p^k) p^k.$$

For any index set $I \subset \{1, ..., \dim(\widetilde{V})\}$ the fact that $\widetilde{a}(p^k, p^l) = 0 \quad \forall\, k \neq l$ implies

$$\left| \sum_{k \in I} \widetilde{b}(v, p^k) p^k \right|_{\widetilde{a}}^2 = \sum_{k \in I} \widetilde{b}(v, p^k)^2 |p^k|_{\widetilde{a}}^2,$$

and thus

$$|v|_{\widetilde{a}}^2 = |\widetilde{\Pi}_m v|_{\widetilde{a}}^2 + |v - \widetilde{\Pi}_m v|_{\widetilde{a}}^2.$$

and (4.14) follows directly. Finally,

$$\begin{aligned}
\|v - \widetilde{\Pi}_m v\|_{\widetilde{b}}^2 &= \Big\| \sum_{k=m+1}^{\dim(\widetilde{V})} \widetilde{b}(v, p^k)\, p^k \Big\|_{\widetilde{b}}^2 \\
&= \sum_{k=m+1}^{\dim(\widetilde{V})} \widetilde{b}(v, p^k)^2 && \text{(by the } \widetilde{b}\text{-orthonormality of } p^k) \\
&= \sum_{k=m+1}^{\dim(\widetilde{V})} \widetilde{b}(v, p^k)^2 \, \frac{1}{\lambda_k} \, |p^k|_{\widetilde{a}}^2 && \text{(since } \lambda_k = |p^k|_{\widetilde{a}}^2) \\
&\leq \frac{1}{\lambda_{m+1}} \sum_{k=m+1}^{\dim(\widetilde{V})} \widetilde{b}(v, p^k)^2 \, |p^k|_{\widetilde{a}}^2 && \text{(since } \lambda_1 \leq \ldots \leq \lambda_{\dim(\widetilde{V})}) \\
&= \frac{1}{\lambda_{m+1}} |v - \widetilde{\Pi}_m v|_{\widetilde{a}}^2 && \text{(by } \widetilde{a}(p^k, p^l) = 0 \, \forall \, k \neq l).
\end{aligned}$$

□

This lemma will be one of the core arguments to prove the existence of a stable decomposition onto the new GenEO (Generalized Eigenproblems in the Overlap) coarse space and the local subspaces. It is in fact the central part in all the approaches that rely on solving eigenvalue problems, cf. Lemma 3.2 in the pioneering work [7] where b is the l_2 (euclidean) inner product or equation (2.8) in [26] where b is a particular bilinear form defined there. The choice of b is one of the defining elements that characterizes each of these methods and for GenEO it will be introduced in the next section.

4.3 Algebraic construction of a robust coarse space and its analysis

In this section we introduce the coarse space and give a bound on the condition number of the two-level additive Schwarz method with it along with a rigorous proof of this result. The proof will consist in proving the existence of a stable splitting for any function in V_h in the sense of Definition 4.7.

4.3.1 The coarse space

The GenEO coarse space is constructed as follows. In each subdomain we pose a suitable generalized eigenproblem and select a number of low frequency eigenfunctions. These local functions are converted into global coarse basis functions using a partition of unity operator. As mentioned before, the eigenproblems are restricted to the overlapping zone, which is introduced in the next definition. Following this definition, we will then define the partition of unity operator, which will appear both in the eigenproblems themselves and in the construction of the coarse basis functions.

Definition 4.12 (Overlapping zone)**.** For each subdomain Ω_j $(1 \leq j \leq N)$, the overlapping zone is given by

$$\Omega_j^\circ = \{x \in \Omega_j : \exists \, j' \neq j \quad \text{such that} \quad x \in \Omega_{j'}\}.$$

We will also require the set of degrees of freedom associated with $V_h(\Omega_j)$, as well as those associated with $V_{h,0}(\Omega_j)$, for $1 \leq j \leq N$.

Definition 4.13. Given a subdomain D that is a union of elements from $\mathcal{T}_h$, let

$$\overline{\text{dof}}(D) := \{k = 1, \ldots, n : \text{supp}\,(\phi_k) \cap D \neq \emptyset\}$$

denote the set of degrees of freedom that are 'active' in D, including those associated with the boundary. Similarly, we denote by

$$\text{dof}(D) := \{k : 1 \leq k \leq n \text{ and } \text{supp}(\phi_k) \subset \overline{D}\}$$

the set of internal degrees of freedom in D.

Remark 4.14. Since the basis functions ϕ_k of V_h fulfil a unisolvence property on each element they also fulfil a unisolvence property on each subdomain Ω_j, in other words the functions $\{\phi_k|_{\Omega_j}\}_{k\in\overline{\text{dof}}(\Omega_j)}$ (resp. $\{\phi_k|_{\Omega_j}\}_{k\in\text{dof}(\Omega_j)}$) are linearly independent. A direct consequence is that these functions form a basis of $V_h(\Omega_j)$ (resp. $V_{h,0}(\Omega_j)$).

Now we can introduce the partition of unity operators. Recall that, for any $v \in V_h$, we write $v = \sum_{k=1}^{n} v_k\, \phi_k$.

Definition 4.15 (Partition of unity)**.** For each degree of freedom $k \in \text{dof}(\Omega) := \{1, \ldots, n\}$, let $\{\mu_{j,k} : k \in \text{dof}(\Omega_j),\ 1 \leq j \leq N\}$ be a family of weights such that

$$\mu_{j,k} \geq 1 \quad \text{and} \quad \sum_{\{j;k\in\text{dof}(\Omega_j)\}} \frac{1}{\mu_{j,k}} = 1.$$

Then, for $1 \leq j \leq N$, the local partition of unity operator $\Xi_j : V_h(\Omega_j) \to V_{h,0}(\Omega_j)$ is defined by

$$\Xi_j(v) := \sum_{k\in\text{dof}(\Omega_j)} \frac{1}{\mu_{j,k}}\, v_k\, \phi_k|_{\Omega_j}, \quad \text{for all } v \in V_h(\Omega_j).$$

Remark 4.16. A possible choice for the weights in Definition 4.15 is to use the multiplicity of each degree of freedom: for any degree of freedom $k \in \text{dof}(\Omega)$, let μ_k denote the number of subdomains for which k is an internal degree of freedom, i.e.

$$\mu_k := \#\{j : 1 \leq j \leq N \text{ and } k \in \text{dof}(\Omega_j)\}$$

and then use the equal weights $\mu_{j,k} := \mu_k$, for any $j = 1, \ldots, N$ with $k \in \text{dof}(\Omega_j)$.

Lemma 4.17. The operators Ξ_j from Definition 4.15 form a partition of unity in the following sense:

$$\sum_{j=1}^{N} R_j^\top \Xi_j(v|_{\Omega_j}) \;=\; v, \quad \text{for all } v \in V_h. \tag{4.15}$$

Moreover,

$$\Xi_j(v)|_{\Omega_j\setminus\Omega_j^\circ} = v|_{\Omega_j\setminus\Omega_j^\circ}, \quad \text{for all } v \in V_h(\Omega_j) \text{ and } 1 \leq j \leq N. \tag{4.16}$$

Proof. Property (4.15) follows directly from the definition. To show (4.16), let $v \in V_h(\Omega_j)$ and recall that by definition

$$\Xi_j(v)|_{\Omega_j\setminus\Omega_j^\circ} \;=\; \sum_{k\in\text{dof}(\Omega_j)} \frac{1}{\mu_{j,k}}\, v_k\, \phi_k|_{\Omega_j\setminus\Omega_j^\circ}\,.$$

Now note that if $\mu_{j,k} > 1$, then $\phi_k|_{\Omega_j \setminus \Omega_j^\circ} = 0$, because $k \in \mathrm{dof}(\Omega_j')$ for $j \neq j'$. Hence,

$$\Xi_j(v)|_{\Omega_j \setminus \Omega_j^\circ} = \sum_{k \in \mathrm{dof}(\Omega_j) \text{ s.t. } \mu_{j,k}=1} v_k \, \phi_k|_{\Omega_j \setminus \Omega_j^\circ} = \sum_{k \in \overline{\mathrm{dof}}(\Omega_j \setminus \Omega_j^\circ)} v_k \, \phi_k|_{\Omega_j \setminus \Omega_j^\circ},$$

and this is also the definition of $v|_{\Omega_j \setminus \Omega_j^\circ}$.

□

Next we define the local generalized eigenproblems for the GenEO coarse space.

Definition 4.18 (Generalized Eigenproblems in the Overlaps)**.** For each $j = 1, \ldots, N$, we define the following generalized eigenvalue problem

$$a_{\Omega_j}(p, v) = \lambda \, b_j(p, \, v), \quad \text{for all } \; v \in V_h(\Omega_j). \tag{4.17}$$

where $b_j(p, \, v) := a_{\Omega_j^\circ}(\Xi_j(p), \Xi_j(v))$, for all $p, \; v \in V_h(\Omega_j)$.

Remark 4.19. Although the form of the bilinear forms $b_j(\cdot, \cdot)$ seems somewhat artificial, we will see below that it arises naturally in the analysis. It is clear that the eigenvalues and eigenvectors will depend on the choice of the partition of unity in Definition 4.15.

The GenEO coarse space is now constructed (locally) as the span of a suitable subset of the eigenfunctions in (4.17). Finally, to obtain a global coarse space we apply the partition of unity operators.

Definition 4.20 (GenEO coarse space)**.** For each $j = 1, \ldots, N$, let $(p_j^k)_{k=1}^{m_j}$ be the eigenfunctions of the eigenproblem (4.17) in Definition 4.18 corresponding to the m_j smallest eigenvalues. Then,

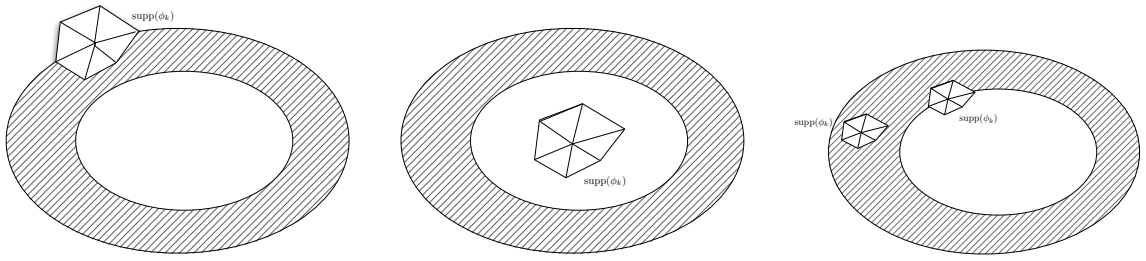
$$V_H := \mathrm{span}\{R_j^\top \Xi_j(p_j^k) : k = 1, \ldots, m_j; \; j = 1, \ldots, N\},$$

where Ξ_j are the partition of unity operators from Definition 4.15 and $R_j^\top$ are the extension operators defined in (4.4).

Consequently, we can also make explicit the final component in Definition 4.5 of the matrix form $\boldsymbol{M}_{AS,2}^{-1}$ of the additive Schwarz preconditioner, namely the prolongation matrix $\boldsymbol{R}_H^T$. The columns of the rectangular matrix $\boldsymbol{R}_H^T \in \mathbb{R}^{n \times \dim(V_H)}$ are simply the vector representations of the functions $\{R_j^\top \Xi_j(p_j^k) : k = 1, \ldots, m_j; \; j = 1, \ldots, N\}$ with respect to the finite element basis $\{\phi_k\}_{k=1}^n$. Clearly $\dim(V_H) = \sum_{j=1}^N m_j$ and a strategy for selecting m_j will be given below. This completes the definition of $\boldsymbol{M}_{AS,2}^{-1}$.

4.3.2 Analysis of the preconditioner

To confirm the robustness of the above coarse space and to bound the condition number of $\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}$ via Theorem 4.8 we will now show that there is a stable splitting for each $v \in V_h$ in the sense of Definition 4.7. First we will give some results on the local subspaces Ω_j, then we use them to show that the eigenproblems from Definition 4.18 are well defined and that the eigenpairs have some particular properties. In order to do this we define a subspace $\widetilde{V}_j$ of each $V_h(\Omega_j)$ on which the restriction of the local generalized eigenproblems satisfy the hypotheses of Lemma 4.11. This leads to local projectors onto subspaces of $V_h(\Omega_j)$ which satisfy stability estimates. These stability estimates will generalize to the whole of $V_h(\Omega_j)$ and enable us to split any $v \in V_h$ in a "C_0-stable" manner.



$k \in \beta_1^j$
$\operatorname{supp}(\phi_k) \not\subset \overline{\Omega}_j$

$k \in \beta_2^j$
$\operatorname{supp}(\phi_k) \subset \overline{\Omega_j \setminus \Omega_j^\circ}$

$k \in \beta_3^j$
$\operatorname{supp}(\phi_k) \subset \overline{\Omega}_j$,
$\operatorname{supp}(\phi_k) \not\subset \Omega_j \setminus \Omega_j^\circ$

Figure 4.1: Three types of finite element basis functions on each subdomain Ω_j. The hashed surface is the overlap Ω_j°.

Definition 4.21. We partition the set $\overline{\mathrm{dof}}(\Omega_j)$ of degrees of freedom in $V_h(\Omega_j)$ into three sets (see also Figure 4.1):

$\beta_1^j := \overline{\mathrm{dof}}(\Omega_j) \setminus \mathrm{dof}(\Omega_j)$ (the DOFs on the boundary of Ω_j),
$\beta_2^j := \mathrm{dof}(\Omega_j \setminus \Omega_j^\circ)$ (the interior DOFs in $\Omega_j \setminus \Omega_j^\circ$),
$\beta_3^j := \mathrm{dof}(\Omega_j) \setminus \mathrm{dof}(\Omega_j \setminus \Omega_j^\circ)$ (the DOFs in the overlap, incl. the inner boundary).

From these index sets we define subsets of functions of $V_h(\Omega_j)$

$$\mathcal{B}_1^j := \operatorname{span}\{\phi_k|_{\Omega_j}\}_{k\in\beta_1^j}, \quad \mathcal{B}_2^j := \operatorname{span}\{\phi_k|_{\Omega_j}\}_{k\in\beta_2^j} \quad \text{and} \quad \mathcal{B}_3^j := \operatorname{span}\{\phi_k|_{\Omega_j}\}_{k\in\beta_3^j},$$

such that

$$V_h(\Omega_j) = \mathcal{B}_1^j \oplus \mathcal{B}_2^j \oplus \mathcal{B}_3^j.$$

The following simple properties will be used frequently in the following.

Lemma 4.22. For any $1 \le j \le N$, the following properties are true

1. $\operatorname{supp}(v) \subset \Omega_j^\circ$, for all $v \in \mathcal{B}_1^j$,
2. $\mathcal{B}_1^j = \mathrm{Ker}(\Xi_j)$,
3. $\mathcal{B}_2^j = \{v \in V_h(\Omega_j) : v|_{\Omega_j^\circ} = 0\}$,
4. a_{Ω_j} is coercive on $\mathcal{B}_2^j$.

Proof.

1. For any basis function ϕ_k with $k \in \beta_1^j$, Lemma 4.4 implies that there is another subdomain $\Omega_{j'}$ with $\operatorname{supp}(\phi_k) \subset \overline{\Omega}_{j'}$, and so $\operatorname{supp}(\phi_k) \cap (\Omega_j \setminus \Omega_j^\circ) = \emptyset$.
2. Let $v \in V_h(\Omega_j)$. Then
$$v \in \mathrm{Ker}(\Xi_j) \quad \Leftrightarrow \quad v_k = 0, \ \text{ for all } \ k \in \mathrm{dof}(\Omega_j) \quad \Leftrightarrow \quad v = \sum_{k\in\beta_1^j} v_k \phi_k|_{\Omega_j} \in \mathcal{B}_1^j .$$
3. It is clear from the definition of $\mathcal{B}_2^j$ that $\mathcal{B}_2^j \subset \{v \in V_h(\Omega_j) : v|_{\Omega_j^\circ} = 0\}$. Conversely, if $v|_{\Omega_j^\circ} = 0$, then from the unisolvence property, $v_k = 0$, for all $k \in \overline{\mathrm{dof}}(\Omega_j^\circ) = \beta_1^j \cup \beta_3^j$, and therefore $\{v \in V_h(\Omega_j) : v|_{\Omega_j^\circ} = 0\} \subset \mathcal{B}_2^j$ also.

4. The previous property implies that $\mathcal{B}_2^j \subset V_{h,0}(\Omega_j)$ and so

$$a_{\Omega_j}(v,w) = a(R_j^\top v, R_j^\top w) \quad \text{for all } v,w \in \mathcal{B}_2^j.$$

The coercivity of $a_{\Omega_j}(\cdot,\cdot)$ on $\mathcal{B}_2^j$ follows from the coercivity of $a(\cdot,\cdot)$.

□

To carry out a robustness analysis we need to make the following two assumptions.

Assumption 4.23. For any $1 \le j \le N$, a_{Ω_j} is coercive on $\mathcal{B}_1^j$.

Assumption 4.24. For any $1 \le j \le N$, $a_{\Omega_j^\circ}$ is coercive on $\mathcal{B}_3^j$.

Note that by the first property in Lemma 4.22, Assumption 4.23 is equivalent to assuming that, for any $1 \le j \le N$, $a_{\Omega_j^\circ}$ is coercive on $\mathcal{B}_1^j$.

Remark 4.25. Assumptions 4.23 and 4.24 are not too restrictive. If all the element stiffness matrices are positive definite, then a_{Ω_j} and $a_{\Omega_j^\circ}$ are positive definite on the whole of $V_h(\Omega_j)$. For the Darcy equation or linear elasticity, the element stiffness matrices are not positive definite. However, any function $v \in \mathcal{B}_1^j$ satisfies $v_k = 0$, for $k \notin \beta_1^j$, and any function $v \in \mathcal{B}_3^j$ vanishes on the boundary of Ω_j (i.e. $v_k = 0$, for $k \in \beta_1^j$). Therefore, in the Darcy case and in the case of standard H^1-conforming finite elements, Assumptions 4.23 and 4.24 hold if each of the sets β_1^j and β_3^j contains at least one DOF. To make the assumptions hold for linear elasticity, the sets β_1^j and β_3^j need to contain enough DOFs to fix the rigid body modes in Ω_j°, i.e., at least $3(d-1)$ DOFs. Hence, for standard H^1-conforming finite elements, it is sufficient to have d non-collinear points (with associated DOFs for all components of the vector function) that lie on the outer boundary $\partial\Omega_j$, respectively in $\overline{\Omega_j^\circ} \setminus \partial\Omega_j$.

The final technical hurdle to construct a stable splitting is that we cannot apply the abstract Lemma 4.11 to the specific eigenproblems used in the construction of the GenEO coarse space V_H directly, because the bilinear forms $b_j(\cdot,\cdot) := a_{\Omega_j^\circ}(\Xi_j(\cdot), \Xi_j(\cdot))$ from Definition 4.17 are not necessarily positive definite on all of $V_h(\Omega_j) \times V_h(\Omega_j)$, for all $1 \le j \le N$. To complete the analysis we thus need to define a suitable subspace $\widetilde{V}_j \subset V_h(\Omega_j)$ such that b_j is positive definite on $\widetilde{V}_j \times \widetilde{V}_j$.

Definition 4.26. Let the spaces $\widetilde{V}_j$ and $\widetilde{W}_j$ be defined by

$$\widetilde{V}_j := \{v \in V_h(\Omega_j) : a_{\Omega_j}(v,w) = 0, \text{ for all } w \in \widetilde{W}_j\} \quad \text{where} \quad \widetilde{W}_j := \mathcal{B}_1^j \oplus \mathcal{B}_2^j\,.$$

Lemma 4.27. Under Assumption 4.23,

$$V_h(\Omega_j) = \widetilde{V}_j \oplus \widetilde{W}_j.$$

Proof. Since a_{Ω_j} is coercive on $\mathcal{B}_1^j$ (cf. Assumption 4.23) and on $\mathcal{B}_2^j$ (cf. Lemma 4.22 (4)) and since functions in $\mathcal{B}_1^j$ and $\mathcal{B}_2^j$ have disjoint supports, we also have that a_{Ω_j} is coercive on $\widetilde{W}_j$. It follows from the definition of $\widetilde{V}_j$ (via some simple linear algebra) that $\widetilde{V}_j \cap \widetilde{W}_j = \{0\}$ and that $\dim(\widetilde{V}_j) = \dim(V_h(\Omega_j)) - \dim(\widetilde{W}_j)$.

□

Remark 4.28. While this lemma shows that $\widetilde{V}_j$ and $\mathcal{B}_3^j$ contain the same degrees of freedom, it does not imply that $\widetilde{V}_j = \mathcal{B}_3^j$. Indeed having chosen values for the degrees of freedom in β_3^j, the corresponding function in $\widetilde{V}_j$ is the discrete PDE-harmonic extension to the whole of Ω_j while the corresponding function in $\mathcal{B}_3^j$ is the extension by zero. The discrete harmonic extension into $\Omega_j \setminus \Omega_j^\circ$ is always well defined because of the coercivity of a_{Ω_j} on $\mathcal{B}_2^j$ (cf. Lemma 4.22 (4)). The fact that the discrete harmonic extension onto $\mathcal{B}_1^j$ is well defined is a consequence of Assumption 4.23.

The role of Assumption 4.24 becomes clear in the next lemma.

Lemma 4.29. Under Assumptions 4.23 and 4.24, for $j = 1, ..., N$, the bilinear form $b_j(\cdot,\cdot) := a_{\Omega_j^\circ}(\Xi_j(\cdot), \Xi_j(\cdot))$ is positive definite on $\widetilde{V}_j \times \widetilde{V}_j$.

Proof. Let $v \in \widetilde{V}_j$ such that $\widetilde{b}_j(v, v) = 0$. We need to show that necessarily $v = 0$.

There exists a unique decomposition $v = v_1 + v_2 + v_3$, such that $v_i \in \mathcal{B}_i^j$. The second property in Lemma 4.22 states that $\mathcal{B}_1^j = \mathrm{Ker}(\Xi_j)$, and so

$$\Xi_j(v_1) = 0.$$

From the definition of Ξ_j it is obvious that $\Xi_j|_{\mathcal{B}_2^j} : \mathcal{B}_2^j \to \mathcal{B}_2^j$ is the identity, and so $\Xi_j(v_2) \in \mathcal{B}_2^j$ and in particular from the third property in Lemma 4.22

$$\mathrm{supp}\,(\Xi_j(v_2)) \cap \Omega_j^\circ = \emptyset.$$

From these two remarks and the definition of b_j it follows that

$$b_j(v, v) = a_{\Omega_j^\circ}(\Xi_j(v_3), \Xi_j(v_3)). \tag{4.18}$$

Moreover, from the definition of Ξ_j it is also obvious that $\Xi_j|_{\mathcal{B}_3^j} : \mathcal{B}_3^j \to \mathcal{B}_3^j$ is a bijection, and so $\Xi_j(v_3) \in \mathcal{B}_3^j$. Now, (4.18) and Assumption 4.24 imply that $\Xi_j(v_3) = 0$. The fact that $\Xi_j|_{\mathcal{B}_3^j}$ is a bijection in turn implies that $v_3 = 0$, and so $v \in \widetilde{W}_j$. From Lemma 4.27, we know that $\widetilde{V}_j \cap \widetilde{W}_j = \{0\}$, and so $v = 0$ which ends the proof. □

We can now apply Lemma 4.11 to the restriction of the GenEO eigenproblems to $\widetilde{V}_j \times \widetilde{V}_j$ and characterize the entire spectrum (including the infinite eigenvalues).

Lemma 4.30. For each $j = 1, ..., N$, consider the generalized eigenproblem (4.17) in Definition 4.18.

(i) There are $\dim(\widetilde{V}_j)$ finite eigenvalues $0 \leq \lambda_1^j \leq \lambda_2^j \leq \ldots \leq \lambda^j_{\dim(\widetilde{V}_j)} < \infty$ (counted according to multiplicity) with corresponding eigenvectors denoted by $\{p_j^k\}_{k=1}^{\dim(\widetilde{V}_j)}$ and normalized to form an orthonormal basis of $\widetilde{V}_j$ with respect to $b_j(\cdot,\cdot)$.

(ii) There are $\dim(\widetilde{W}_j)$ infinite eigenvalues $\lambda^j_{\dim(\widetilde{V}_j)+1} = \ldots = \lambda^j_{\dim(V_h(\Omega_j))} = \infty$ with associated eigenvectors denoted by $\{p_j^k\}_{\dim(\widetilde{V}_j)+1}^{\dim(V_h(\Omega_j))}$ forming a basis of $\widetilde{W}_j$.

Proof. Since $V_h(\Omega_j) = \widetilde{V}_j \oplus \widetilde{W}_j$ (cf. Lemma 4.27) and $a_{\Omega_j}(v, w) = b_j(v, w) = 0$, for all $v \in \widetilde{V}_j$ and $w \in \widetilde{W}_j$, the eigenproblem (4.17) can be decoupled into two eigenproblems: one on $\widetilde{V}_j$ and one on $\widetilde{W}_j$.

Since, according to Lemma 4.29, $b_j(\cdot,\cdot)$ is coercive on $\widetilde{V}_j \times \widetilde{V}_j$, we can apply Lemma 4.11 with $\widetilde{V} \mapsto \widetilde{V}_j$, $\widetilde{a} \mapsto a_{\Omega_j}$, and $\widetilde{b} \mapsto b_j$ to analyse the restriction of (4.17) to $\widetilde{V}_j$. This completes the proof of (i).

For the restriction of (4.17) to $\widetilde{W}_j$, we prove that all vectors in $\widetilde{W}_j$ are eigenvectors associated with the eigenvalue $+\infty$ in the sense of Definition 4.10. Let $v \in \widetilde{W}_j$. Then $\Xi_j(v)|_{\Omega_j^\circ} = 0$ and so in particular

$$a_{\Omega_j^\circ}(\Xi_j(v), \Xi_j(w)) = 0 \quad \text{for all } v,\ w \in \widetilde{W}_j\,. \tag{4.19}$$

Moreover, we have already seen in the proof of Lemma 4.27 that a_{Ω_j} is coercive on $\widetilde{W}_j$, and so

$$a_{\Omega_j}(v, v) \neq 0 \quad \text{for all } v \in \widetilde{W}_j \backslash \{0\}. \tag{4.20}$$

Due to (4.19) and (4.20), any $v \in \widetilde{W}_j$ is indeed an eigenvector to the eigenvalue $+\infty$ in the sense of Definition 4.10. We can use any set of linearly independent vectors in $\widetilde{W}_j$ to form a basis, e.g. $\{p_j^k\}_{k=\dim(\widetilde{V}_j)+1}^{\dim(V_h(\Omega_j))} = \{\phi_k|_{\Omega_j}\}_{k\in\beta_1^j\cup\beta_2^j}$.

□

We are now ready to define the crucial projection operators onto the local components of the GenEO coarse space that satisfy suitable stability estimates.

Lemma 4.31 (Local stability estimate)**.** Let $j \in \{1, ..., N\}$ and let $\{(p_j^k, \lambda_j^k)\}_{k=1}^{\dim(V_h(\Omega_j))}$ be as defined in Lemma 4.30. Suppose that $m_j \in \{1, \ldots, \dim(V_h(\Omega_j)) - 1\}$ such that $0 < \lambda_{m_j+1}^j < \infty$. Then, the local projection operator

$$\Pi_{m_j}^j v := \sum_{k=1}^{m_j} a_{\Omega_j^\circ}(\Xi_j(v), \Xi_j(p_j^k))\, p_j^k$$

satisfies

$$|\Pi_{m_j}^j v|_{a,\Omega_j} \leq |v|_{a,\Omega_j} \quad \text{and} \quad |v - \Pi_{m_j}^j v|_{a,\Omega_j} \leq |v|_{a,\Omega_j}\,, \quad \text{for all } v \in V_h(\Omega_j), \tag{4.21}$$

as well as the stability estimate

$$\left|\Xi_j(v - \Pi_{m_j}^j v)\right|_{a,\Omega_j^\circ}^2 \leq \frac{1}{\lambda_{m_j+1}^j} \left|v - \Pi_{m_j}^j v\right|_{a,\Omega_j}^2\,, \quad \text{for all } v \in V_h(\Omega_j). \tag{4.22}$$

Proof. The condition $\lambda_{m_j+1}^j < \infty$, ensures that $m_j \leq \dim(\widetilde{V}_j)$, so $\Pi_{m_j}^j$ maps to $\widetilde{V}_j$. Therefore, for all $v \in \widetilde{V}_j$, the estimates in (4.21) and (4.22) can be deduced from Lemma 4.11 again, with $\widetilde{V} \mapsto \widetilde{V}_j$, $\widetilde{a} \mapsto a_{\Omega_j}$, $\widetilde{b} \mapsto b_j$, and $m \mapsto m_j$.

To prove the result for all $v \in V_h(\Omega_j)$, we use again the fact that $V_h(\Omega_j) = \widetilde{V}_j \oplus \widetilde{W}_j$ and that $a_{\Omega_j}(v, w) = 0$, for all $v \in \widetilde{V}_j$ and $w \in \widetilde{W}_j$. Let $v = v_V + v_W \in V_h(\Omega_j)$ with $v_V \in \widetilde{V}_j$ and $v_W \in \widetilde{W}_j$. Then $\Pi_{m_j}^j v = \Pi_{m_j}^j v_V$ and so (4.21) follows due to the a_{Ω_j}-orthogonality of $\widetilde{V}_j$ and $\widetilde{W}_j$. Estimate (4.22) follows similarly from $\Xi_j(v_W)|_{\Omega_j^\circ} = 0$.

□

Lemma 4.32 (Stable decomposition)**.** Let $v \in V_h$ and suppose the definitions and notations of Lemma 4.31 hold. Then, the decomposition

$$z_0 := \sum_{j=1}^{N} \Xi_j(\Pi_{m_j}^j v|_{\Omega_j}), \qquad z_j := \Xi_j(v|_{\Omega_j} - \Pi_{m_j}^j v|_{\Omega_j}), \quad \text{for } j = 1, \ldots, N,$$

is C_0-stable with

$$C_0^2 = 2 + k_0(2k_0+1)\max_{1\le j\le N}\Big(1+\frac{1}{\lambda^j_{m_j+1}}\Big).$$

Proof. By definition $\|z_j\|^2_{a,\Omega_j} = |\Xi_j(v-\Pi^j_{m_j}v|_{\Omega_j})|^2_{a,\Omega_j^\circ} + |\Xi_j(v-\Pi^j_{m_j}v|_{\Omega_j})|^2_{a,\Omega_j\setminus\Omega_j^\circ}$. However, due to property (4.16) in Lemma 4.17, Ξ_j is the identity for restrictions of functions to $\Omega_j\setminus\Omega_j^\circ$, and so

$$\|z_j\|^2_{a,\Omega_j} = |\Xi_j(v-\Pi^j_{m_j}v|_{\Omega_j})|^2_{a,\Omega_j^\circ} + |v-\Pi^j_{m_j}v|_{\Omega_j}|^2_{a,\Omega_j\setminus\Omega_j^\circ}.$$

Now we can apply Lemma 4.31 to get

$$\|z_j\|^2_{a,\Omega_j} \le \Big(1+\frac{1}{\lambda^j_{m_j+1}}\Big)|v-\Pi^j_{m_j}v|_{\Omega_j}|^2_{a,\Omega_j} \le \Big(1+\frac{1}{\lambda^j_{m_j+1}}\Big)|v|^2_{a,\Omega_j},$$

where in the last step we have used (4.21).

□

With this stable decomposition we can now state our main result on the convergence of the two-level Schwarz preconditioner with the new GenEO coarse space. It follows immediately from Theorem 4.8 and Lemma 4.32.

Theorem 4.33 (Bound on the condition number)**.** Let Assumptions 4.1, 4.23, and 4.24 hold. Suppose that the coarse space V_H is given by Definition 4.20 and $\boldsymbol{M}^{-1}_{AS,2}$ is as defined in (4.5). Then we can bound the condition number for the two-level Schwarz method by

$$\kappa(\boldsymbol{M}^{-1}_{AS,2}\boldsymbol{A}) \le (1+k_0)\Big[2+k_0(2k_0+1)\max_{1\le j\le N}\Big(1+\frac{1}{\lambda^j_{m_j+1}}\Big)\Big],$$

where k_0 is given in Definition 4.5.

The only parameters that need to be chosen in our coarse space are the numbers m_j of eigenmodes on each subdomain Ω_j, $1\le j\le N$, to be included in the coarse space. We suggest the following choice which recovers the condition number estimate for problems with no strong coefficient variation.

Corollary 4.34. For any j, $1\le j\le N$, let

$$m_j := \min\left\{m : \lambda^j_{m+1} > \frac{\delta_j}{H_j}\right\}, \tag{4.23}$$

where δ_j is a measure of the width of the overlap Ω_j° and $H_j = \operatorname{diam}(\Omega_j)$. Then

$$\kappa(\boldsymbol{M}^{-1}_{AS,2}\boldsymbol{A}) \le (1+k_0)\Big[2+k_0(2k_0+1)\max_{1\le j\le N}\Big(1+\frac{H_j}{\delta_j}\Big)\Big].$$

Note that the number of subdomains and the coefficient variations do not appear in this bound on the condition number. This means that we have established rigorously that the algorithm is robust with respect to these two parameters. We will confirm this with some numerical tests in Section 4.5. The size of the coarse space induced by the criterion does however depend on the geometry of the coefficient variation in the overlaps and the choice of the partition of unity. In fact, for some problems it may happen that even for a very small criterion the number of eigenmodes which are selected is very large. This is the case for instance in the context of linear elasticity when one of the materials is almost incompressible (i.e. its Poisson ratio approaches 1/2), because then the bilinear form $a_{\Omega_j^\circ}(\Xi_j(\cdot),\Xi_j(\cdot))$ on the right hand side of eigenproblem (4.17) has very high energy.

4.3.3 Two variants

In this section we present two variants around the GenEO coarse space. First we change the two level additive Schwarz preconditioner for the hybrid Schwarz preconditioner and prove a convergence result for this preconditioner with the GenEO coarse space. Next we propose a slight modification of the GenEO eigenproblem and also give a convergence result.

Hybrid Schwarz preconditioner The hybrid Schwarz preconditioner is

$$\boldsymbol{M}_{hy}^{-1} = \boldsymbol{R}_0^\top \boldsymbol{A}_0^{-1} \boldsymbol{R}_0 + (\boldsymbol{I} - \boldsymbol{P}_0) \left(\sum_{i=1}^{N} \boldsymbol{R}_j^\top \boldsymbol{A}_j^{-1} \boldsymbol{R}_j \right) (\boldsymbol{I} - \boldsymbol{P}_0)^\top. \tag{4.24}$$

We already introduced hybrid preconditioners in the introduction (2.30). For our theoretical analysis we will look at $\boldsymbol{M}_{hy}^{-1}$ in the abstract Schwarz framework as an additive preconditioner:

- The local solvers are the same as for the Additive Schwarz preconditioner: $\boldsymbol{A}_j = \boldsymbol{R}_j \boldsymbol{A} \boldsymbol{R}_j^\top, \quad \forall i = 0, \ldots, N.$
- The coarse interpolation operator is simply $\boldsymbol{R}_0^\top$.
- The local interpolation operators are $(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{R}_j^\top$ for $j = 1, \ldots, N$.

The fact that the coarse corrections are now also applied multiplicatively with respect to the one level additive Schwarz preconditioner leads to an improved upper bound for the eigenvalues of $\boldsymbol{M}_{hy}^{-1}\boldsymbol{A}$.

Lemma 4.35. With k_0 as in Definition 4.5, the largest eigenvalue of $\boldsymbol{M}_{hy}^{-1}\boldsymbol{A}$ satisfies

$$\lambda_{max}(\boldsymbol{M}_{hy}^{-1}\boldsymbol{A}) \;\leq\; k_0.$$

Proof. As usual we use Rayleigh quotients to prove the result

$$\begin{aligned}
\langle \boldsymbol{M}_{hy}^{-1}\boldsymbol{A}\boldsymbol{u}, \boldsymbol{A}\boldsymbol{u}\rangle &= \langle \boldsymbol{A}\boldsymbol{P}_0\boldsymbol{u}, \boldsymbol{P}_0\boldsymbol{u}\rangle + \sum_{j=1}^{N} \langle \boldsymbol{R}_j^\top \boldsymbol{A}_j^{-1} \boldsymbol{R}_j \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}, (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}\rangle \\
&\leq \langle \boldsymbol{A}\boldsymbol{P}_0\boldsymbol{u}, \boldsymbol{P}_0\boldsymbol{u}\rangle + k_0 \langle \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}, (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{u}\rangle \\
&\leq k_0 \langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{u}\rangle,
\end{aligned}$$

where in the second line we used a result proved in [25, Section 4] and in the first and last lines we use the $\boldsymbol{A}$-orthogonality of projection $\boldsymbol{P}_0$. □

Because the local interpolation operators are $(I - P_0)R_j^\top$ instead of $R_j^\top$, Definition 4.7 of a stable splitting must be slightly adapted: in (4.6) the condition $v = \sum_{j=0}^{N} R_j^\top z_j$ must be replaced by $v = \sum_{j=1}^{N}(I - P_0)R_j^\top z_j + R_0^\top z_0$.

Lemma 4.36 (Stable Decomposition: Hybrid preconditioner)**.** Let $v \in V_h$, then with notation introduced in Lemma 4.32 the decomposition

$$\begin{aligned}
z_0 &:= v_0, \qquad \text{such that } R_0^\top v_0 = P_0 v, \\
z_j &:= \Xi_j(v|_{\Omega_j} - \Pi_{m_j}^j v|_{\Omega_j}), \quad \text{for } j = 1, \ldots, N,
\end{aligned}$$

is C_0-stable with

$$C_0^2 = 1 + k_0 \max_{1 \leq j \leq N} \Big(1 + \frac{1}{\lambda_{m_j+1}^j}\Big).$$

Proof. This splitting is from [19]. Let $u \in V_h$, first we need to prove that the z_j indeed provide a splitting of u:

$$\begin{aligned} R_0^\top z_0 + \sum_{j=1}^{N} (I - P_0) R_j^\top z_j &= P_0 v + (I - P_0) R_j^\top \Xi_j (v|_{\Omega_j} - \Pi_{m_j}^j v|_{\Omega_j}) \\ &= P_0 v + (I - P_0) R_j^\top \Xi_j (v|_{\Omega_j}), \end{aligned}$$

where the argument is that $(I - P_0) R_j^\top \Xi_j (\Pi_{m_j}^j v|_{\Omega_j}) = 0$ since

$$R_j^\top \Xi_j (\Pi_{m_j}^j v|_{\Omega_j}) \in \operatorname{span}\{R_j^\top \Xi_j (p_j^k);\, k = 1, \ldots, m_j\}$$

and

$$\operatorname{span}\{R_j^\top \Xi_j (p_j^k);\, k = 1, \ldots, m_j\} \subset V_H = \operatorname{range}(P_0) = \operatorname{Ker}(I - P_0).$$

Finally,

$$R_0^\top z_0 + \sum_{j=1}^{N} (I - P_0) R_j^\top z_j = P_0 v + (I - P_0) v = v$$

The stability property has pretty much been proved already in the proof of Lemma 4.32:

$$\|z_j\|_{a,\Omega_j}^2 \;\leq\; \Big(1 + \frac{1}{\lambda_{m_j+1}^j}\Big) |v|_{a,\Omega_j}^2,$$

so by definition of k_0

$$\sum_{j=1}^{N} \|z_j\|_{a,\Omega_j}^2 \;\leq k_0 \max_{1 \leq j \leq N} \Big(1 + \frac{1}{\lambda_{m_j+1}^j}\Big) |v|_a^2,$$

and finally

$$\sum_{j=0}^{N} \|z_j\|_{a,\Omega_j}^2 \;\leq \left[1 + k_0 \max_{1 \leq j \leq N} \Big(1 + \frac{1}{\lambda_{m_j+1}^j}\Big)\right] |v|_a^2.$$

□

Theorem 4.37 (Bound on the condition number: hybrid operator)**.** Let Assumptions 4.1, 4.23, and 4.24 hold. Assume that the coarse space V_H is given by Definition 4.20 and $\boldsymbol{M}_{hy}^{-1}$ is as defined in (4.24). Then we can bound the condition number of the preconditioned operator by

$$\kappa(\boldsymbol{M}_{hy}^{-1} \boldsymbol{A}) \;\leq\; k_0 \Big[1 + k_0 \max_{1 \leq j \leq N} \Big(1 + \frac{1}{\lambda_{m_j+1}^j}\Big)\Big],$$

where k_0 is given in Definition 4.5.

A different eigenproblem Another variant for GenEO is to replace the generalized eigenvalue problem in Definition 4.18 by the following

Definition 4.38 (Generalized Eigenproblems in the Overlaps: a Variant)**.** For each $j = 1, \ldots, N$, we define the following generalized eigenvalue problem

$$a_{\Omega_j}(p, v) = \lambda\, a_{\Omega_j}(\Xi_j(p), \Xi_j(v)), \quad \text{for all} \;\; v \in V_h(\Omega_j). \tag{4.25}$$

The difference is that the matrix in the right hand side of the generalized eigenvalue problem is no longer restricted to the overlap. The remainder of the definition of the coarse space is unchanged:

Definition 4.39 (GenEO coarse space: a Variant)**.** For each $j = 1, \ldots, N$, let $(p_k^j)_{k=1}^{m_j}$ be the eigenfunctions of the eigenproblem (4.25) in Definition 4.38 corresponding to the m_j smallest eigenvalues. Then,

$$V_H' := \text{span}\{R_j^\top \Xi_j(p_k^j) : k = 1, \ldots, m_j;\ j = 1, \ldots, N\}.$$

Thanks to this choice the technical Assumption 4.24 is no longer required. Next, we give the convergence theorems for the fully Additive and Hybrid preconditioners with these modified coarse spaces. The proofs are very similar and slightly more simple than with the original GenEO.

Theorem 4.40 (Bound on the condition numbers: modified coarse space)**.** Let Assumptions 4.1 and 4.23 hold. Suppose that the coarse space V_H is given by Definition 4.39 and $\boldsymbol{M}_{AS,2}^{-1}$ and $\boldsymbol{M}_{hy}^{-1}$ are as defined in (4.5) and (4.24). Then we can bound the condition number of the preconditioned operators by

$$\kappa(\boldsymbol{M}_{AS,2}^{-1}\boldsymbol{A}) \ \leq\ (1+k_0)\left[2 + k_0(2k_0+1)\max_{1\leq j\leq N}\Big(\frac{1}{\lambda_{m_j+1}^j}\Big)\right],$$

and

$$\kappa(\boldsymbol{M}_{hy}^{-1}\boldsymbol{A}) \ \leq\ k_0\left[1 + k_0\max_{1\leq j\leq N}\Big(\frac{1}{\lambda_{m_j+1}^j}\Big)\right],$$

where k_0 is given in Definition 4.5.

4.4 Implementation

In this section we would like to address implementation issues of the proposed algorithm involving the GenEO coarse space. In the sections above, we have worked with function spaces as they are more convenient in the analysis. However, as we will demonstrate below, our algorithm requires only abstract information of the problem in form of the element stiffness matrices and no further information on the mesh, the finite element spaces, or any coefficients. Indeed, for running the algorithm we need

(i) the list $\overline{\text{dof}}(\tau)$ of degrees of freedom associated with each element $\tau \in \mathcal{T}_h$,

(ii) the element stiffness matrix $\boldsymbol{A}^\tau = (a_\tau(\phi_l, \phi_k))_{k,\, l\in\overline{\text{dof}}(\tau)}$ associated with each element $\tau \in \mathcal{T}_h$.

Unless the overlapping subdomain partition is available a priori, we additionally need

(iii) the number ℓ of layers which determine the amount of overlap.

Before going into details, we note that as for the classical two-level overlapping Schwarz method (see, e.g. [112, Sect. 3]), our algorithm can be parallelized straightforwardly. In particular, the solution of the eigenproblems in the preprocessing step and the subdomain solves during each PCG iteration can be performed fully in parallel.

4.4.1 Preprocessing

We need the overlapping partition $\Omega = \bigcup_{j=1}^{N} \Omega_j$ in form of the list of elements associated with each subdomain Ω_j. To obtain this, we first create the connectivity graph of the elements (using the lists $\overline{\mathrm{dof}}(\tau)$ from (i)) and partition it into disjoint sets of elements which make up the non-overlapping subdomains Ω'_j using for instance METIS [54] or SCOTCH [13]. Then, for each (global) DOF k, we build the list

$$\mathrm{elem}(k) = \{\tau \in \mathcal{T}_h : k \in \overline{\mathrm{dof}}(\tau)\}$$

of elements where DOF k is active. This list realizes $\mathrm{supp}(\phi_k)$ without knowing the basis function ϕ_k itself. In a second step we add ℓ layers to each non-overlapping subdomain Ω'_j according to Definition 4.3, which finally results in a list of elements per (overlapping) subdomain Ω_j. From this, we construct

$$\overline{\mathrm{dof}}(\Omega_j) = \bigcup_{\tau \subset \bar{\Omega}_j} \overline{\mathrm{dof}}(\tau)$$

(cf. Definition 4.13). Then we can compute the set of *internal* degrees of freedom in Ω_j

$$\mathrm{dof}(\Omega_j) = \left\{ k \in \overline{\mathrm{dof}}(\Omega_j) : \bigcup_{\tau \in \mathrm{elem}(k)} \tau \subset \bar{\Omega}_j \right\}$$

(cf. Definition 4.15). Finally it is straightforward to get the list of elements that make up the overlapping zone Ω_j° for each $j = 1, \ldots, N$, namely $\{\tau \subset \bar{\Omega}_j : \tau \subset \bar{\Omega}_{j'}, j' \neq j\}$.

4.4.2 The eigenproblems

For each subdomain Ω_j, $j = 1, \ldots, N$ we use a local renumbering of the degrees of freedom $\overline{\mathrm{dof}}(\Omega_j)$ of $V_h(\Omega_j)$. By assembling the element stiffness matrices for these DOFs over the elements $\tau \subset \bar{\Omega}_j$, we get the subdomain "Neumann" matrix $\widetilde{\boldsymbol{A}}_j$. This is the matrix formulation of $a_{\Omega_j}(\cdot,\cdot) : V_h(\Omega_j) \times V_h(\Omega_j) \to \mathbb{R}$. For the same renumbering of DOFs, we assemble only over the elements $\tau \subset \bar{\Omega}_j^\circ$ in the overlap and obtain matrix $\widetilde{\boldsymbol{A}}_j^\circ$ associated with the bilinear form $a_{\Omega_j^\circ}(\cdot,\cdot) : V_h(\Omega_j) \to V_h(\Omega_j)$. Note that $\widetilde{\boldsymbol{A}}_j$ and $\widetilde{\boldsymbol{A}}_j^\circ$ have the same format, but $\widetilde{\boldsymbol{A}}_j^\circ$ usually contains a block of zeros corresponding to the degrees of freedom that are in the part of Ω_j which is not overlapped by other subdomains.

From Definition 4.15, we see immediately that the action of the operator Ξ_j can be coded by a diagonal matrix $\boldsymbol{X}_j$, where the diagonal entry corresponding to DOF k is equal to $1/\mu_{j,k}$.

With these notations, the eigenproblem given in Definition 4.18 reads: Find the eigenvectors $\boldsymbol{p}_j^k \in \mathbb{R}^{\#\overline{\mathrm{dof}}(\Omega_j)}$ and eigenvalues $\lambda_j^k \in \mathbb{R} \cup \{+\infty\}$ that satisfy

$$\widetilde{\boldsymbol{A}}_j \boldsymbol{p}_j^k = \lambda_j^k \, \boldsymbol{X}_j \widetilde{\boldsymbol{A}}_j^\circ \boldsymbol{X}_j \boldsymbol{p}_j^k. \tag{4.26}$$

To get the coarse basis functions, we need to solve these eigenproblems (at least we need sufficiently many eigenpairs corresponding to low frequent modes) and to then select m_j of these eigenfunctions for our coarse space. With the criterion suggested in (4.23), we need measures δ_j and H_j for the width of the overlapping zone and the subdomain diameter, respectively. If the mesh can be assumed to be quasi-uniform, we may replace the ratio δ_j/H_j by the number of layers of extension we applied in subdomain Ω_j divided by the number of layers Ω_j contains in total (which is available via the connectivity graph).

4.4.3 The preconditioner

Having selected the eigenvectors $\boldsymbol{p}_k^j$, the coarse basis functions are given by the vectors $\widetilde{\boldsymbol{R}}_j^T \boldsymbol{X}_j \boldsymbol{p}_j^k$, where the matrix $\widetilde{\boldsymbol{R}}_j^T$ maps the renumbered DOFs to the global DOFs and fills the rest of the vector with zeros. The columns of the matrix $\boldsymbol{R}_H^T$ are exactly the vectors $\widetilde{\boldsymbol{R}}_j^T \boldsymbol{X}_j \boldsymbol{p}_j^k$, where $j = 1, \ldots, N$, $k = 1, \ldots, m_j$. The coarse matrix $\boldsymbol{A}_H = \boldsymbol{R}_H \boldsymbol{A} \boldsymbol{R}_H^T$ can be efficiently assembled subdomain-wise by using the fact that the coarse basis functions corresponding to two subdomains only interact when the subdomains overlap. Thus, in a parallel regime, we basically only need next-neighbor communication.

As for the 'one level' part of the preconditioner we have made the list $\mathrm{dof}(\Omega_j)$ of internal degrees of freedom for subdomain Ω_j available in the preprocessing step. Then R_j is simply a Boolean matrix which renumbers local vectors into global vectors and the matrix counterpart $\boldsymbol{A}_j$ of $a_{\Omega_j}(\cdot,\cdot) : V_{h,0}(\Omega_j) \times V_{h,0}(\Omega_j) \to \mathbb{R}$ is computed by assembling the element matrices for elements τ in the ready made list $\{\tau \subset \hat{\mathrm{A}}\ \bar{\Omega}_j\}$.

Clearly, once the information above is stored and the matrices $\boldsymbol{A}_j$ are factorized, each application of $\boldsymbol{M}_{AS,2}^{-1}$ (within the PCG) can be carried out efficiently.

4.4.4 An alternative way of solving the eigenproblems

The size of the (algebraic) eigenproblem (4.26) to be solved in each subdomain can be reduced. By rearranging the local DOFs $\overline{\mathrm{dof}}(\Omega_j)$ with respect to the sets β_1^j (the boundary), β_2^j (the overlap), and β_3^j (the interior) (cf. Definition 4.21), the matrices $\widetilde{\boldsymbol{A}}_j$ and $\boldsymbol{B}_j := \boldsymbol{X}_j \widetilde{\boldsymbol{A}}_j^{\circ} \boldsymbol{X}_j$ take the following block form

$$\widetilde{\boldsymbol{A}}_j = \begin{pmatrix} \widetilde{\boldsymbol{A}}_j^{11} & 0 & \widetilde{\boldsymbol{A}}_j^{13} \\ 0 & \widetilde{\boldsymbol{A}}_j^{22} & \widetilde{\boldsymbol{A}}_j^{23} \\ (\widetilde{\boldsymbol{A}}_j^{13})^T & (\widetilde{\boldsymbol{A}}_j^{23})^T & \widetilde{\boldsymbol{A}}_j^{33} \end{pmatrix}, \qquad \boldsymbol{B}_j = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{B}_j^{33} \end{pmatrix},$$

where $\widetilde{\boldsymbol{A}}_j^{kl} = a_{\Omega_j}(\phi_m, \phi_n)_{n \in \beta_k^j, m \in \beta_l^j}$. The two zero blocks in $\widetilde{\boldsymbol{A}}_j$ are due the fact that the supports of functions in $\mathcal{B}_1^j$ and $\mathcal{B}_2^j$ are always disjoint. Since $\widetilde{\boldsymbol{A}}_j^{11}$ is the matrix version of the bilinear form $a_{\Omega_j^\circ}(\cdot,\cdot) : \mathcal{B}_1^j \times \mathcal{B}_1^j \to \mathbb{R}$, and since Assumption 4.23 states that $a_{\Omega_j^\circ}(\cdot,\cdot)$ is coercive on $\mathcal{B}_1$, it follows that the block $\widetilde{\boldsymbol{A}}_j^{11}$ is positive definite and thus invertible. Similarly, $\boldsymbol{A}_j^{22}$ is positive definite due to Lemma 4.22 (4). This means that the Schur complement $\boldsymbol{S}_j = \widetilde{\boldsymbol{A}}_j^{33} - \widetilde{\boldsymbol{A}}_j^{13}[\widetilde{\boldsymbol{A}}_j^{11}]^{-1}\widetilde{\boldsymbol{A}}_j^{13} - \widetilde{\boldsymbol{A}}_j^{23}[\widetilde{\boldsymbol{A}}_j^{22}]^{-1}\widetilde{\boldsymbol{A}}_j^{23}$ is well defined and we can reduce eigenproblem (4.26) to an eigenproblem for the Schur complement

$$\boldsymbol{S}_j \boldsymbol{p}_k^{j,3} = \lambda_j^k \boldsymbol{B}_j^{33} \boldsymbol{p}_k^{j,3}. \tag{4.27}$$

The two remaining blocks in $\boldsymbol{p}^j$ can then be computed from

$$\begin{aligned} \boldsymbol{p}_k^{j,1} &= -[\widetilde{\boldsymbol{A}}_j^{11}]^{-1}\widetilde{\boldsymbol{A}}_j^{13}\boldsymbol{p}_k^{j,3}, \\ \boldsymbol{p}_k^{j,2} &= -[\widetilde{\boldsymbol{A}}_j^{22}]^{-1}\widetilde{\boldsymbol{A}}_j^{23}\boldsymbol{p}_k^{j,3} \end{aligned}$$

(i.e. via discrete harmonic extension). The only difference is that with this version of the eigenproblem there are no infinite eigenvalues. Because we are only interested in the small eigenvalues we can solve eigenproblem (4.27) instead of (4.26). Due to the appearance of the Schur complement $\boldsymbol{S}_j$ and because we are interested only in the first few eigenpairs,

an iterative eigensolver could be applied, e.g., we could use the inverse power method [81], ARPACK [65] or the LOBPCG method [59], maybe using a suitable regularization of $\widetilde{\boldsymbol{A}}_{jj}^{33}$ or $\boldsymbol{S}_j$ as a preconditioner. This, however, will be the subject of future research and we will use a direct eigensolver in the next section. Note finally, that the blocks $\boldsymbol{p}_k^{j,2}$ never need to be calculated in practice as they are annihilated by the matrix $\boldsymbol{X}_j$.

4.5 Numerical results

We have introduced an algorithm for a wide range of problems. In this section we test its efficiency on the two- and three-dimensional Darcy equation and on the two- and three-dimensional linear elasticity equations with heterogeneous coefficients. We have already conducted, with success, a scalability and robustness test in the Introduction (Section 2.3.2). For all our numerical examples we have used FreeFem++ [50] to define the test cases and build all the finite element data. Throughout we have used standard piecewise linear ($\mathbb{P}_1$) finite elements. The eigenvalue problems were solved using LAPACK [1]. For the remainder (including the subdomain solves and the coarse solve) we have used Matlab. Throughout this section we compare three methods.

1. The first one is the one-level additive Schwarz method (referred to as AS), defined by the preconditioner $\boldsymbol{M}_{AS,1}^{-1} = \sum_{j=1}^{N} \boldsymbol{R}_j^T \boldsymbol{A}_j^{-1} \boldsymbol{R}_j$.
2. The second one (referred to as ZEM for Zero Energy Modes) is the two-level method given by (4.5) with the coarse space $V_H := \text{span}\{\boldsymbol{R}_j^T \Xi_j(\boldsymbol{q}_k^j)\}_{j,k}$ where the $\boldsymbol{q}_k^j$ span the kernel of the subdomain operator. For the Darcy equation these are the constant functions and for elasticity the rigid body modes. In the floating subdomains that do not touch the Dirichlet boundary, this basically coincides with choosing $m_j = \dim(\text{Ker}(a_{\Omega_j}))$ in our GenEO method.
3. The third method (referred to as GenEO) is the two-level method introduced here, with the number m_j, for $j = 1, \ldots, N$, chosen according to (4.23) (except for one test where we will explicitly state this). The partition of unity operators are chosen to be the ones in Remark 4.16 where the weights are the multiplicities of each degree of freedom.

For each of these methods we use the Preconditioned Conjugate Gradient (PCG) solver. As a stopping criterion we apply $\|v - \bar{v}\|_\infty < 10^{-6} \|\bar{v}\|_\infty$ where $\bar{v}$ is the solution of (4.2) obtained *via* a direct solver on the global problem (unless otherwise stated). Of course this criterion is not practical but in this context we have chosen it to ensure a fair comparison.

In the tables below, we provide the number of PCG iterations needed to reach convergence. We have also computed condition number estimates for each of the preconditioned matrices using the Rayleigh-Ritz procedure [95] on the Krylov subspaces within PCG. We do not give any detail on the maximal and minimal eigenvalue. However, we can report that adding/enriching the coarse space leads to larger minimal eigenvalues, whereas the maximal eigenvalue depends only on the geometry. This is in agreement with Lemma 4.6 and Theorem 4.8. Finally, we also display the dimension of the coarse space V_H in each case.

For both three-dimensional scalability test (Sections 4.5.3 and 4.5.4), we use the domain $\Omega = [0, L] \times [0, 1] \times [0, 1]$ and a regular tetrahedral mesh of $(10L+1 \times 11 \times 11)$ nodes which we divide into L subdomains, horizontally side by side. We will either use a regular partition into L unit cubes (Figure 4.2 (left)) or an automatic partition into L subdomains using Metis (Figure 4.2 (right)). In the two dimensional test cases (Sections 4.5.1 and 4.5.2), we will use Metis partitions of the unit square.

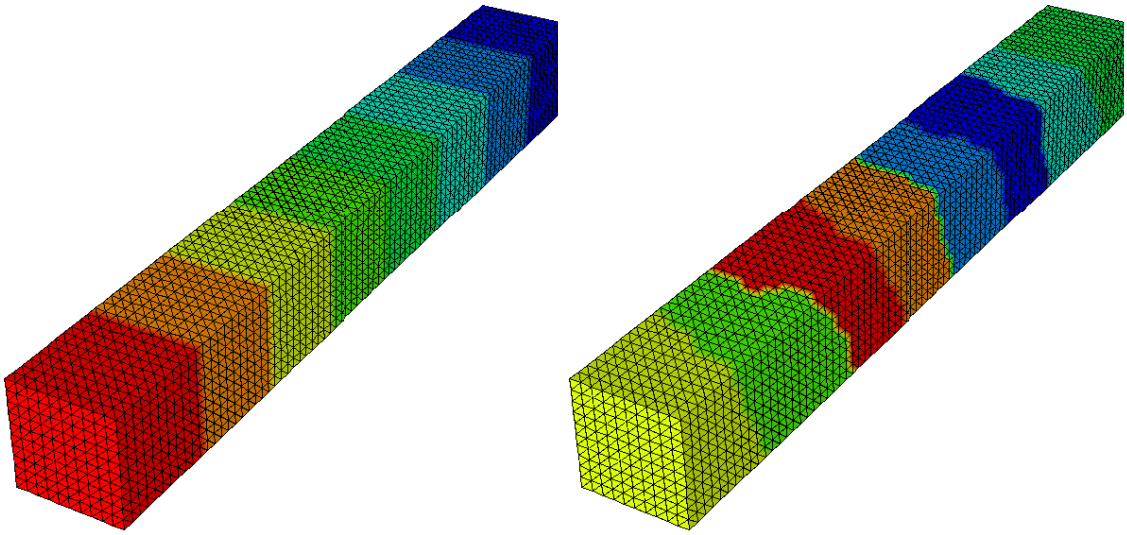


Figure 4.2: Partition of Ω into $L = 8$ subdomains – regular (left) and Metis (right)

4.5.1 The two-dimensional Darcy equation

We run a simulation for the Darcy equation $-\boldsymbol{\nabla} \cdot (\alpha \boldsymbol{\nabla} v) = 1$ in $\Omega = [0,1]^2$ with homogeneous Dirichlet boundary conditions on the whole of $\partial\Omega$. The mesh is 200×200 square elements further subdivided into triangles. The coefficient distribution is rather random since it is given by a QR code. This is shown on the left hand side of Figure 4.3 where in the yellow (or light) parts $\alpha = 1$ and in the pink (or dark) parts $\alpha = 1000$. The decomposition into subdomains is the 100 subdomain partition obtained *via* Metis [54] where we add one layer of overlap to each subdomain. This is illustrated on the right hand side of Figure 4.3. In Figure 4.4 we have plotted the condition number versus the size of the coarse space for this test. As a matter of comparison: without any coarse space (AS) the condition number is 9661. With just the weighted constant $\Xi_j(1_{|\Omega_j})$ per floating subdomain (ZEM) the condition number is 7324: this 62 dimensional coarse space is what we get for GenEO with a barely positive threshold $\tau = 0^+$. We have not plotted this on the graph purely for scaling issues. What this illustrates is that there is a good compromise to be found between the size of the coarse space and the efficiency of the method. An automatic optimal choice for $\mathcal{K}_j$ is a subject for future research.

4.5.2 The two-dimensional linear elasticity equations

In this subsection, we look at the two-dimensional linear elasticity equations with a Dirichlet boundary condition at $x = 0$ and Neumann conditions otherwise: Find $\boldsymbol{u} = (u_1, u_2)^{\mathrm{T}} \in H^1(\Omega)^2$ such that

$$-\mathrm{div}(\sigma(\boldsymbol{u})) = \boldsymbol{f}, \quad \text{in } \Omega,$$

$\boldsymbol{u} = (0,0)^{\mathrm{T}}$ on $\partial\Omega_D = \{(x,y) \in \partial\Omega : x = 0\}$ and $\sigma(\boldsymbol{u}) \cdot \boldsymbol{n} = 0$ on the rest of $\partial\Omega$, where the stress tensor $\sigma(\boldsymbol{u})$, the Lamé coefficients λ and μ and the right hand side are given by

$$\begin{cases} \sigma_{ij}(\boldsymbol{u}) = 2\mu\varepsilon_{ij}(\boldsymbol{u}) + \lambda\delta_{ij}\mathrm{div}(\boldsymbol{u}),\ \varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),\ i,j = 1,2 \\ \boldsymbol{f} = (0,g)^T, \\ \mu = \frac{E}{2(1+\nu)},\ \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}. \end{cases}$$

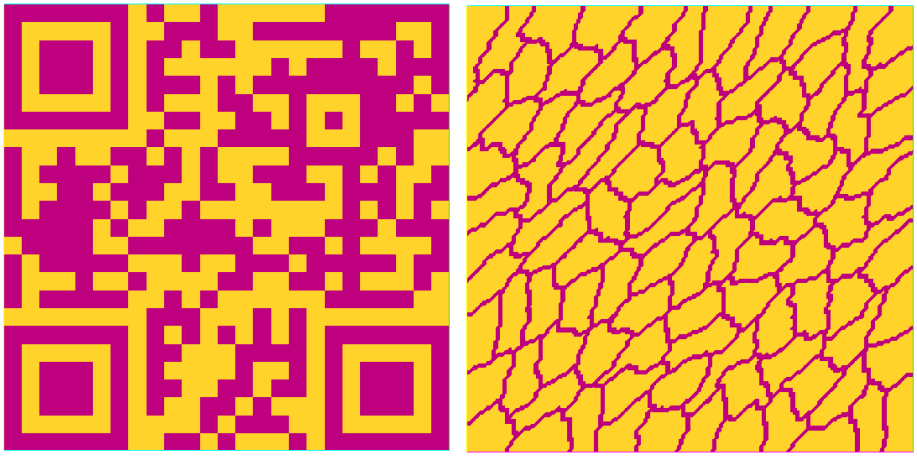


Figure 4.3: Left: coefficient distribution (pink or dark is high conductivity) – Right: Metis partition of the 200×200 mesh into 100 subdomains

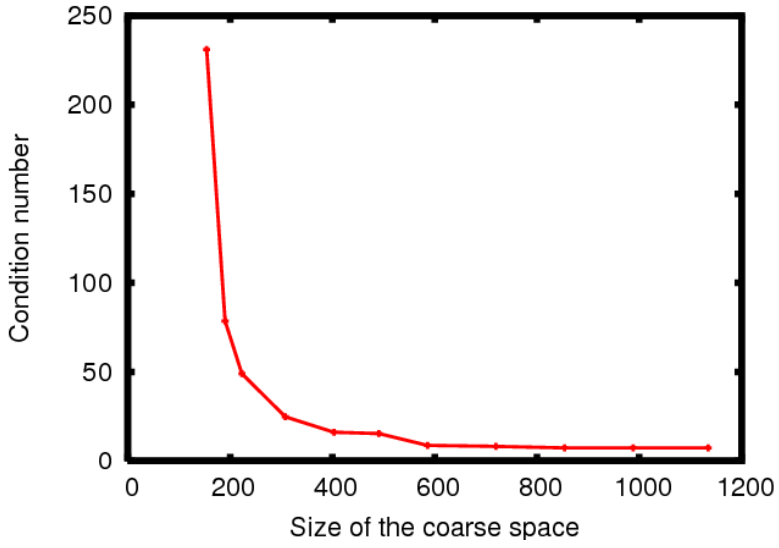


Figure 4.4: For the geometry given in Figure 4.3 we plot the condition number with respect to the coarse space size when the threshold successively takes the values $\tau \in [0.01; 0.05; 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9]$. We observe that the most troublesome eigenmodes are identified for quite a small value of the threshold and a reasonable size of the coarse space, then the condition number stagnates.

		AS	ZEM		GenEO	
sub	glob DOF	it	it	dim	it	dim
4	13122	90	94	12	36	36
16	13122	169	179	48	39	112
25	13122	222	157	75	40	166
64	13122	317	196	192	39	343

Table 4.1: 2D Elasticity: number of PCG iterations (it) and coarse space dimension (dim) vs. number (sub) of Metis subdomains for fixed problem size

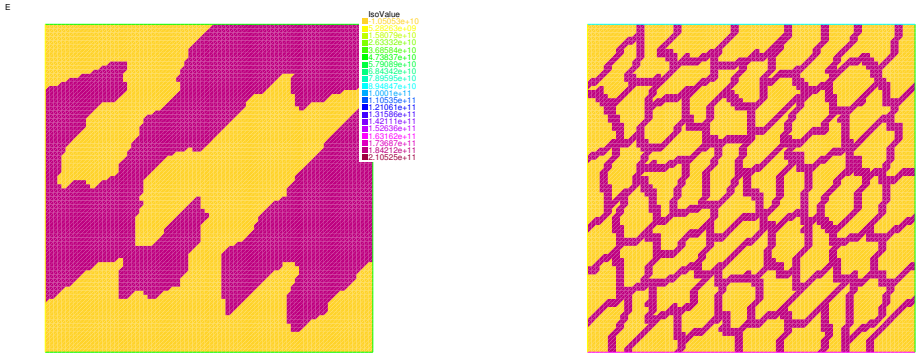


Figure 4.5: 2D Elasticity: coefficient distribution (left) – Metis decomposition into 64 subdomains (right)

In this case, the ZEM coarse space consists of three rigid body modes per subdomain. We choose $\Omega = [0,1]^2$ and use a structured simplicial mesh with 81×81 nodes. The coefficient distribution is sketched on the left hand side of Figure 4.5, where in the two regions (indicated by the two different colors) we take the parameters $(E_1, \nu_1) = (2 \cdot 10^{11}, 0.3)$ and $(E_2, \nu_2) = (2 \cdot 10^7, 0.45)$ for Young's modulus and Poisson's ratio.

We keep the problem size constant, but we make the number of subdomains vary. In all cases, we use a Metis partition and extend the non-overlapping subdomains by $\ell = 2$ layers. As shown in Figure 4.5 (right) for a decomposition into 64 subdomains there are many floating subdomains. Table 4.1 shows the iteration counts and coarse space dimensions for different Metis partitions (some parameters are defined in the table's caption). From the iteration counts we see that the GenEO method is scalable.

It is not surprising that the coarse space dimension grows with the number of subdomains because we construct local coarse basis functions per subdomain. Note however that for the case of 64 subdomains, the coarse space dimension of 343 is still comparable to the average dimension of 205 of a subdomain problem.

4.5.3 The three-dimensional Darcy equation

On the domain $\Omega \subset \mathbb{R}^3$ given above, we solve the following problem: Find $v \in H^1(\Omega)$ such that

$$-\boldsymbol{\nabla} \cdot (\kappa \boldsymbol{\nabla} v) = 0 \quad \text{in} \quad \Omega, \tag{4.28}$$

$v = 0$ on $\partial\Omega_D = \{(x, y, z) \in \partial\Omega : x = 0\}$ and $\kappa \boldsymbol{\nabla} v \cdot \boldsymbol{n} = 0$ on the rest of $\partial\Omega$, where $\boldsymbol{n}$ is the outward unit normal. The coefficient distribution alternates between two different

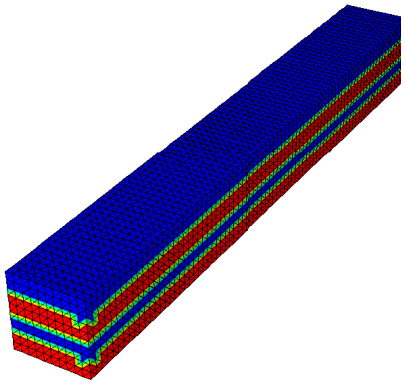


Figure 4.6: Coefficient distribution (four alternating layers)

	AS		ZEM			GenEO		
κ_2	it	*cond*	it	*cond*	dim	it	*cond*	dim
1	16	*229*	11	*6.3*	8	11	*8.4*	7
10^2	27	*230*	19	*22*	8	13	*8.4*	14
10^4	29	*230*	23	*210*	8	15	*8.4*	14
10^6	26	*230*	22	*230*	8	11	*8.4*	14

Table 4.2: 3D Darcy: number of PCG iterations (it), condition number (*cond) and coarse space dimension (dim) vs. jump in* κ *for* $\kappa_1 = 1$, $\ell = 1$ *added layer,* $L = 8$ *regular subdomains*

constant values κ_1 and κ_2 of κ on four horizontal layers (as shown in Figure 4.6).

First, we study the robustness of our algorithm with respect to the coefficient variation. We partition Ω into $L = 8$ (non-overlapping) regular subdomains. Each subdomain is then extended by $\ell = 1$ layer in order to create the overlapping partition. Table 4.2 shows the iteration counts and condition numbers for fixed value $\kappa_1 = 1$ and various κ_2. As expected, for our algorithm the condition number and the number of PCG iterations are robust with respect to the jump κ_2/κ_1. Furthermore, for $\kappa_2 = \kappa_1$, the algorithm automatically selects seven eigenmodes (one per floating subdomain) to build the coarse space, this leads essentially to the same choice as for the ZEM method except for the subdomain in which the Dirichlet boundary condition is active, where GenEO does not select any coarse mode.

The second test that we conduct is for the scalability with respect to the problem size and the number of subdomains. For simplicity, we make the problem parameter L vary. Recall that increasing L elongates the bar-shaped domain and at the same time increases the number of subdomains which equals L. Thus, the global number of degrees of freedom is also proportional to L. Table 4.3 gives the results for different problem sizes (we display the number of subdomains and the total number of degrees of freedom) and for regular and irregular partitions. For regular partitions we use (4.23); for irregular partitions, the choice of m_j becomes more tricky since there may be additional 'bad' eigenmodes close to the ratio δ_j/H_j that are due to the irregularity of the subdomains and not due to any coefficient variation. In particular, the ratio δ_j/H_j which is constant for regular partitions, as L gets increased, may differ significantly for two 'Metis' decompositions into L and L' subdomains with $L \neq L'$. In the regular case, (4.23) leads to $m_j = 2$ and $\lambda_3 = 0.5$. Thus, in order for the bound on the condition number given by Theorem 4.33 to be at least as

		Regular							
		AS		ZEM			GenEO		
L	glob DOF	it	*cond*	it	*cond*	dim	it	*cond*	dim
4	4840	14	*51*	15	*51*	4	10	*8.4*	6
8	9680	26	*230*	22	*230*	8	11	*8.4*	14
16	19360	51	*980*	36	*970*	16	13	*8.4*	30
32	38720	103	*4000*	61	*3900*	32	13	*8.4*	62
		Metis with criterion given by (4.29)							
		AS		ZEM			GenEO		
L	glob DOF	it	*cond*	it	*cond*	dim	it	*cond*	dim
4	4840	21	*67*	18	*63*	4	9	*3.0*	19
8	9680	36	*290*	29	*280*	8	9	*3.0*	40
16	19360	65	*1200*	45	*1200*	16	11	*3.1*	81
32	38720	123	*4900*	79	*4700*	32	11	*3.1*	171

Table 4.3: 3D Darcy: number of PCG iterations (it), condition number (*cond*) and coarse space dimension (dim) vs. problem size for $\kappa_1 = 1$, $\kappa_2 = 10^6$, $\ell = 1$ added layer, L subdomains

	AS		ZEM			GenEO		
ℓ	it	*cond*	it	*cond*	dim	it	*cond*	dim
1	26	*230*	22	*230*	8	11	*8.4*	14
2	22	*150*	18	*150*	8	9	*5.4*	14
3	16	*110*	15	*110*	8	9	*4.0*	14
4	15	*92*	13	*92*	8	7	*3.3*	14

Table 4.4: 3D Darcy: number of PCG iterations (it), condition number (*cond*) and coarse space dimension (dim) vs. number ℓ of layers added to each domain, for $L = 8$ regular subdomains, $\kappa_1 = 1$ and $\kappa_2 = 10^6$

strict in the irregular ('Metis') case we set

$$m_j \;:=\; \min\left\{m : \lambda^j_{m+1} > 0.5\right\}, \tag{4.29}$$

in each subdomain in Table 4.3. We note that the condition numbers in both the regular and irregular subdomain cases are stable and consistently low.

Finally, Table 4.4 studies the dependence on the amount of overlap, or equivalently on the number ℓ of layers added to each non-overlapping subdomain. We can see that for this example, increasing the amount of overlap improves convergence without increasing the dimension of the coarse space.

4.5.4 The three-dimensional linear elasticity equations

For this family of tests the equations are the following. Find $\boldsymbol{u} = (u_1, u_2, u_3)^{\mathrm{T}} \in H^1(\Omega)^3$ such that

$$-\mathrm{div}(\sigma(\boldsymbol{u})) = \boldsymbol{f}, \quad \text{in } \Omega,$$

		AS		ZEM			GenEO		
L	glob DOF	it	*cond*	it	*cond*	dim	it	*cond*	dim
4	14520	79	$2.4 \cdot 10^3$	54	$2.9 \cdot 10^2$	24	16	*10*	46
8	29040	177	$1.3 \cdot 10^4$	87	$1.0 \cdot 10^3$	48	16	*10*	102
16	58080	378	$1.5 \cdot 10^5$	145	$1.4 \cdot 10^3$	96	16	*10*	214

Table 4.5: 3D Elasticity: number of PCG iterations (it), condition number (*cond*), and coarse space dimension (dim) vs. number of regular subdomains, for $\ell = 1$ added layer, $g = 10$, $(E_1, \nu_1) = (2 \cdot 10^{11}, 0.3)$ and $(E_2, \nu_2) = (2 \cdot 10^7, 0.45)$.

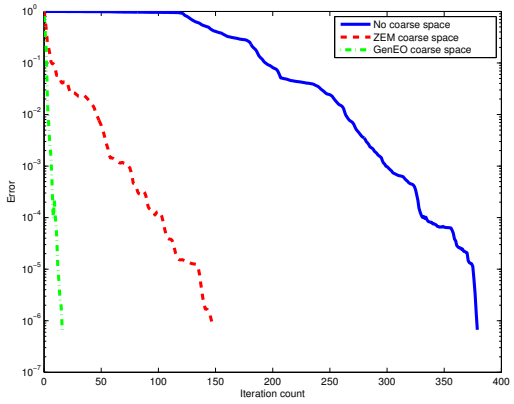


Figure 4.7: 3D Elasticity: Relative error vs. iteration count for $L = 16$ regular subdomains

$\boldsymbol{u} = (0,0,0)^{\mathrm{T}}$ on $\partial\Omega_D = \{(x,y,z) \in \partial\Omega : x = 0\}$ and $\sigma(\boldsymbol{u}) \cdot \boldsymbol{n} = 0$ on the rest of $\partial\Omega$, where the stress tensor $\sigma(\boldsymbol{u})$, the Lamé coefficients λ and μ and the right hand side are given by

$$\begin{cases} \sigma_{ij}(\boldsymbol{u}) = 2\mu\varepsilon_{ij}(\boldsymbol{u}) + \lambda\delta_{ij}\mathrm{div}(\boldsymbol{u}),\ \varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \boldsymbol{f} = (0,0,g)^T, \\ \mu = \frac{E}{2(1+\nu)},\ \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}. \end{cases}$$

Once more E and ν denote respectively Young's modulus and Poisson's ratio, and we will let both parameters vary discontinuously over the domain taking the values (E_1, ν_1) and (E_2, ν_2) alternating in four layers, as shown in Figure 4.6. Table 4.5 displays iteration counts, condition numbers, and coarse space dimensions for various partitions into regular subdomains (the parameter choices are given below the table). Note that for GenEO, we need only 16 PCG iterations in all cases. As an example, Figure 4.7 shows the convergence profile for the case where Ω is split into 16 regular subdomains.

4.6 Conclusion

In this Chapter we have introduced a coarse space for symmetric positive definite variational problems. In order to remain as general as possible, we did so using an abstract formulation. We rigorously proved a bound for the condition number of the overlapping two-level additive Schwarz preconditioner for this coarse space. This bound does not depend on any of the coefficients in the equations or on the way the domain is split into subdomains. Numerical results on two-dimensional and three-dimensional problems are in agreement with the fact that the method is robust with respect to heterogeneities

and rather irregular subdomains. We also gave details on how to implement the coarse space construction. This relies only on having access to finite element stiffness matrices and the underlying connectivity graph. No additional data is required and no additional elementary matrices need to be computed. This means that the method is quite easily applicable to simulations of actual physical problems and it is our ambition to do so.

Along the way we have identified promising leads to further improve the efficiency of the method. The first one is to take advantage of the fact that the partition of unity can be chosen differently since the proof holds as long as the partition of unity is defined by individual weights per interior degree of freedom in each subdomain. The second idea is to optimize the eigenvalue computations. Although this is a purely parallel task, this is the most costly part in the coarse space construction. Finally, the formulation of the GenEO coarse space makes it particularly well suited for a multilevel parallel implementation, which is of particular interest in cases where a two-level approach leads to excessively large coarse spaces. We begin to explore some of these leads in Chapter 7.

Chapter 5

Generalization of GenEO to substructuring methods

The content of this chapter was published in International Journal for Numerical Engineering [109] in collaboration Daniel J. Rixen. The method and convergence result were first presented in the note [108] with Victorita Dolean, Patrice Hauret, Frédéric Nataf, and Daniel J. Rixen.

Contents

5.1 Introduction

With substructuring methods if the domain consists of a few different materials it is possible to partition it in such a way that the interfaces match the boundaries of the different materials. In this case by applying well chosen weights to each of the subdomains

Table 5.1: Summary of Notations

Function space	*Description*	*Definition*
$W_h(\Omega)$	Global	solution space for (5.1)
$W_h(\Omega_i)$	Local	$\{u_{\|\Omega_i}; u \in W_h(\Omega)\}$ ((5.6); $D = \Omega_i$)
W_i	Local trace	$\{u_{\|\Gamma\cap\partial\Omega_i}; u \in W_h(\Omega)\}$ ((5.6); $D = \Gamma \cap \partial\Omega_i$)
W	Product trace	$W_1 \times \dots W_N$
$\hat{W}$	Global trace	$\{u_{\|\Gamma}; u \in W_h(\Omega)\}$ ((5.6) ; $D = \Gamma$)
Stiffness matrices (defined on)	*Matrix*	*Bilinear form*
Global ($W_h(\Omega)$)	$\hat{K}$ (5.3)	$\hat{a}$ (5.1)
Local ($W_h(\Omega_i)$)	K_i	a_{Ω_i} (5.7) for $D = \Omega_i$
Product space ($\prod_{i=1}^N W_h(\Omega_i)$)	K (5.11)	none
Lumped global ($W_h(\Omega)$)	$\hat{K}^{bb}$(5.18)	$\hat{a}^{bb}$
Lumped product space ($\prod_{i=1}^N W_h(\Omega_i)$)	K^{bb}(5.19)	a^{bb}
Schur complement (defined on)	*Matrix*	*Bilinear form*
Global ($\hat{W}$)	$\hat{S}$ (5.16)	$\hat{s}$
Local (W_i)	S_i (5.13)	s_i (5.22)
On the product space (W)	S (5.14)	s
Weighted local (W_i)	$\tilde{S}_i$	$\tilde{s}_i$ (5.23)
Right hand sides	*Notation*	
Condensed onto Γ	$\hat{f}_\Gamma$ (5.20)	
Condensed onto $\Gamma \cap \partial\Omega_i$	$f_{\Gamma,i}$ (5.21)	
Condensed on product space $\prod_{i=1}^N \Gamma \cap \partial\Omega_i$	f_Γ (5.21)	

the negative effect of heterogeneous coefficients can be annihilated [93, 58] even for non-smooth decompositions (where the interfaces are jagged) [57]. If this is not the case one may observe very bad convergence of the iterative solvers for the interface problem (see e.g. [2, 55]). It is also well known that bad aspect ratios of the domains [34] can also lead to poor convergence. This is what we work to improve: we aim to design FETI and BDD solvers for which the convergence rate does not depend on the choice of the decomposition into subdomains or on any of the coefficients in the equations.

In order to achieve this we will use the strategy introduced in the additive Schwarz framework by [105, 106] and [26] and presented in the previous chapter of this manuscript. This strategy is based on the abstract theory of the two-level additive Schwarz method [112]. The strategy is to write the Schwarz theory up to the point where it depends on the set of equations we are dealing with and where assumptions on the coefficient distribution with respect to the decomposition into subdomains are needed to write estimates which do not depend on the parameters. For the Darcy equation ($-\nabla\cdot\nabla(\alpha u) = b$) with the minimal coarse space (the constant functions) the Poincaré inequality and trace theorem are needed to complete the proof and they require quite strong assumptions. Instead, the authors in [26, 106, 105] and the previous chapter of this thesis propose to solve a generalized eigenvalue problem in each subdomain which selects the modes of the solution that satisfy the required estimates. The other modes, which do not satisfy the estimate, are used to build the coarse space and are basically taken care of with a direct solve in the coarse space. This is what we will refer to as the Schwarz-GenEO coarse space (Generalized Eigenvalues in the Overlaps). It leads to a two-level method with a convergence rate chosen a priori for problems described by a symmetric positive definite matrix.

As it turns out [70] also proposes to solve generalized eigenvalue problems to deal with heterogeneous coefficients in the BDDC and FETI-DP frameworks. More recently a multilevel extension of this work for BDDC was proposed in [103] with thorough numerical results. Although, at a glance, the generalized eigenvalue problems look similar to the ones in this chapter they are in fact the result of two different approaches. In [70] the global generalized eigenvalue problem which would need to be solved on the entire domain in order to achieve a targeted convergence rate is given and then it is made local by restricting it to each of the interfaces between two subdomains. The global condition number *indicator* is chosen to be the maximum over all the set of interfaces of a local estimate. In other words the global to local *conversion* of the estimate for the condition number is based on heuristics. The approach in this paper is different because it is inspired by previous work in the Schwarz framework. In particular the global to local *conversion* of the condition number estimate is justified theoretically. It relies very strongly on the abstract Schwarz theory with the result that each local generalized eigenvalue problem is posed on the boundary of a subdomain and not on the interface between two subdomains. With this choice the fact that the targeted condition number will be achieved is guaranteed theoretically.

FETI (Finite Element Tearing and Interconnecting) and BDD (Balancing Domain Decomposition) are two well known non overlapping domain decomposition methods. Balancing Domain Decomposition (BDD) is the work of [67] who added a coarse space to the preexisting Neumann Neumann method [14] to deal with singularities in the local problems. We will refer to the analysis of BDD in [112] which is very closely related to the analysis of the two-level Schwarz preconditioner. The FETI algorithm was first introduced in [36] and the convergence proof is due to [71, 111]. It is generalized in [58]. Coarse spaces for the FETI method are introduced first in [32] and further developed in [35, 33]. In both cases (BDD and FETI) the generalized eigenvalue problem which we solve is used

to prove a bound for the largest eigenvalue of the preconditioned operator. As usual the lower bound for the eigenvalues of the preconditioned operator is 1 regardless of the coarse space.

The rest of the Chapter is organized as follows. In Section 5.2 we introduce the notation which will be needed for both algorithms. In Section 5.3 we introduce the two-level GenEO preconditioner for the BDD algorithm and in Section 5.4 we introduce the two-level preconditioner for the FETI algorithm. The definitions of each of the coarse spaces with the corresponding generalized eigenvalue problems can be found in Definitions 5.15 and 5.31 respectively. These generalized eigenvalue problems are chosen specifically to ensure the properties in Lemmas 5.20 and 5.36 (*i.e.* the stability of the local solvers) are satisfied. As for the convergence results they are stated (and proved) in Theorems 5.23 and 5.38. Finally in section 5.5 we give a few numerical results.

5.2 Notation for FETI and BDD

For a given domain $\Omega \in \mathbb{R}^d$ and a finite dimensional Hilbert space $W_h(\Omega)$, given a symmetric, positive definite bilinear form,

$$\hat{a}(\cdot,\cdot) : W_h(\Omega) \times W_h(\Omega) \to \mathbb{R}, \tag{5.1}$$

and an element $\hat{g} \in W_h(\Omega)'$, we consider the problem of finding $u \in W_h(\Omega)$, such that

$$\hat{a}(u,v) = \hat{g}(v), \quad \forall\, v \in W_h(\Omega). \tag{5.2}$$

In order to introduce the BDD and FETI algorithms we will need to introduce notation for discrete operators at the global and local (on each subdomain) levels.

5.2.1 Problem setting

We begin by rewriting Problem (5.2) in an algebraic framework. As usual in the finite element setting, we start with a triangulation $\mathcal{T}_h$ of Ω: $\Omega = \bigcup_{\tau\in\mathcal{T}_h} \tau$ and a basis $\{\phi_k\}_{1\le k\le N}$ for the finite element space $W_h(\Omega)$.

Assumption 5.1. Given any element τ of the mesh $\mathcal{T}_h$, let $W_h(\tau) := \{u|_\tau : u \in W_h(\Omega)\}$. We assume that for each element $\tau \in \mathcal{T}_h$, there exists a symmetric positive semi-definite (spsd) bilinear form $a_\tau : W_h(\tau) \times W_h(\tau) \to \mathbb{R}$, such that

$$\hat{a}(u,v) = \sum_{\tau\in\mathcal{T}_h} a_\tau(u|_\tau, v|_\tau), \quad \forall\, u, v \in W_h(\Omega),$$

and an element $g_\tau \in W_h(\tau)'$ such that

$$\hat{g}(v) = \sum_{\tau\in\mathcal{T}_h} g_\tau(v|_\tau), \quad \forall\, v \in W_h(\Omega).$$

The stiffness matrix is assembled with the following entries

$$(\hat{K})_{kl} := \hat{a}(\phi_k,\phi_l) \left(= \sum_{\tau\in\mathcal{T}_h} a_\tau(\phi_{k|\tau}, \phi_{l|\tau}) \right), \quad \forall\, k,l = 1,\dots,n, \tag{5.3}$$

and the discrete right hand side $\hat{f} \in \mathbb{R}^n$ is defined by the entries

$$(\hat{f})_k := \hat{g}(\phi_k) \left(= \sum_{\tau\in\mathcal{T}_h} g_\tau(\phi_k|_\tau) \right), \quad \forall\, k = 1,\dots,n.$$

As is quite customary we identify vectors of degrees of freedom, which are in some spaces $\mathbb{R}^m$, with the associated finite element functions. Operators between the spaces are represented as matrices, and we frequently commit an abuse of notation by using matrices and operators interchangeably. With this abuse of notation the original problem (5.2) is equivalent to the linear system: find $u \in W_h(\Omega)$ such that

$$\hat{K}u = \hat{f}, \tag{5.4}$$

with $\hat{K}$ symmetric, positive definite (spd).

5.2.2 Local setting and notation

Local Setting We introduce a partition of the global domain Ω into N non-overlapping subdomains Ω_i which are resolved by the mesh

$$\bar{\Omega} = \bigcup_{i=1}^{N} \bar{\Omega}_i \quad \text{and} \quad \Omega_i \cap \Omega_j = \emptyset, \quad i \neq j,$$

and the resulting set of boundaries between subdomains

$$\Gamma := \bigcup_{i \neq i'} \bar{\Omega}_i \cap \bar{\Omega}_{i'}.$$

The reason why we have required the information on the non-assembled stiffness matrices is that we want to have access to local matrices for any choice of the partition into subdomains. In order to do this we also need to define local finite element spaces and local bilinear forms.

Assumption 5.2. The basis functions ϕ_k are continuous on Ω. In particular for any subset $D \subset \Omega$ the restriction $\phi_{k|D}$ of ϕ_k to D is well defined.

Definition 5.3 (Local finite element spaces)**.** For any subset $D \subset \Omega$ let the set of degrees of freedom in D be the set

$$dof(D) := \{k = 1, \ldots, n; \phi_{k|D} \neq 0_{|D}\}, \tag{5.5}$$

where $0_{|D} : D \to \mathbb{R}$ is identically zero. Then the finite element space on D is defined as

$$W_h(D) := \{u_{|D}; u \in W_h(\Omega)\} = \operatorname{span}\{\phi_{k|D}; k \in dof(D)\}. \tag{5.6}$$

The second equality in the definition of $W_h(D)$ is an immediate consequence.

Definition 5.4 (Local bilinear forms and local right hand sides)**.** For any open subset $D \subset \Omega$ which is resolved by the mesh $\mathcal{T}_h$, let the local bilinear form on D be

$$a_D : W_h(D) \times W_h(D) \to \mathbb{R}; \quad a_D(v, w) := \sum_{\tau \subset D} a_\tau(v_{|\tau}, w_{|\tau}), \tag{5.7}$$

and the local right hand side be the element

$$g_D \in W_h'(D); \quad g_D(v) := \sum_{\tau \subset D} g_\tau(v_{|\tau}). \tag{5.8}$$

For any $i = 1, \ldots, N$, the space of finite element functions on each Ω_i follows from (5.6) with $D = \Omega_i$:

$$W_h(\Omega_i) = \{u_{|\Omega_i}; u \in W_h(\Omega)\},$$

as well as the trace spaces for $D = \partial\Omega_i \cap \Gamma$:

$$W_i := W_h(\Gamma \cap \partial\Omega_i) = \{u_{|\Gamma\cap\partial\Omega_i}; u \in W_h(\Omega)\}.$$

Finally, we define the product space

$$W := \prod_{i=1}^{N} W_i.$$

We know from (5.6) that $W_i = \text{span}\{\phi_{k|\partial\Omega_i\cap\Gamma}; k \in dof(\partial\Omega_i \cap \Gamma)\}$, we make the further assumption that this set of functions is a basis of W_i.

Assumption 5.5. The set $\{\phi_{k|\partial\Omega_i\cap\Gamma}; k \in dof(\partial\Omega_i \cap \Gamma)\}$ is a basis of W_i.

Throughout the analysis, we will consider elements in the product space W. Each component $u_i \in W_i$ is defined on a part $\Gamma \cap \partial\Omega_i$ of the boundary and two contributions from two neighbouring subdomains do not necessarily match on the shared interface. This is a result of the partition of Ω into subdomains. Our finite element approximation of the elliptic problem is, however, based on functions in $W_h(\Omega)$ which are defined on the whole of Ω with one value per degree of freedom. We denote the space of restrictions of these functions to the set of internal boundaries Γ by $\hat{W}$:

$$\hat{W} := W_h(\Gamma) = \{u_{|\Gamma}; u \in W_h(\Omega)\} \left(= \text{span}\{\phi_{k|\Gamma}; k \in dof(\Gamma)\}\right). \tag{5.9}$$

Next we introduce interpolation (prolongation) operators $R_i^\top : W_i \to \hat{W}$ for $i = 1, \ldots, N$:

$$\forall\, u_i = \sum_{k\in dof(\Gamma\cap\partial\Omega_i)} \alpha_i^k \phi_{k|\Gamma\cap\partial\Omega_i}\ (\alpha_i^k \in \mathbb{R}); \quad R_i^\top u_i := \sum_{k\in dof(\Gamma\cap\partial\Omega_i)} \alpha_i^k \phi_{k|\Gamma}.$$

These are the natural interpolation operators represented by boolean matrices: the continuous global function $R_i^\top u_i \in \hat{W}$ shares the same values as u_i for degrees of freedom in $dof(\Gamma \cap \partial\Omega_i)$ and has no contributions from any other degrees of freedom. The corresponding restriction operator $R_i : \hat{W} \to W_i$ is defined as

$$\forall\, u = \sum_{k\in dof(\Gamma)} \alpha^k \phi_{k|\Gamma}\ (\alpha_k \in \mathbb{R}); \quad R_i u := \sum_{k\in dof(\Gamma\cap\partial\Omega_i)} \alpha^k \phi_{k|\Gamma\cap\partial\Omega_i}.$$

We note that $\hat{W} \not\subset W$ and $\hat{W} = \sum_{i=1}^{N} R_i^\top W_i$. It is obvious from the definition of $R_i^\top$ and Assumption 5.5 that for $i = 1, \ldots, N$ and $u_i \in W_i$:

$$u_i = 0_{|\Gamma\cap\partial\Omega_i} \Leftrightarrow R_i^\top u_i = 0_{|\Gamma}. \tag{5.10}$$

Stiffness matrices The local stiffness matrix $K_i : W_h(\Omega_i) \to W_h(\Omega_i)$ is the matrix associated with bilinear form a_{Ω_i} defined by (5.7) for $D = \Omega_i$. From these, the stiffness matrix on the product space is defined as

$$K : W_h(\Omega_1) \times \ldots W_h(\Omega_N) \to W_h(\Omega_1) \times \ldots W_h(\Omega_N); \quad K := \begin{pmatrix} K_1 & 0 & \ldots & 0 \\ 0 & K_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & K_N \end{pmatrix} \tag{5.11}$$

so that

$$Ku = (K_1 u_1, \ldots, K_N u_N)^\top, \quad \forall\, u = (u_1, \ldots, u_N)^\top \in W_h(\Omega_1) \times \ldots W_h(\Omega_N). \tag{5.12}$$

Schur complement matrices The degrees of freedom $dof(\Omega_i)$ in $W_h(\Omega_i)$ can be split into the set $b_i := dof(\Gamma \cap \partial\Omega_i)$ of degrees of freedom that are also in the trace space W_i and the remainder $I_i := dof(\Omega_i) \setminus dof(\Gamma \cap \partial\Omega_i)$. This way we can rewrite the local stiffness matrix in block formulation

$$K_i = \begin{pmatrix} K_i^{b_i b_i} & K_i^{b_i I_i} \\ K_i^{I_i b_i} & K_i^{I_i I_i} \end{pmatrix}.$$

The interior variables of any subdomain are then eliminated in work that can be parallelized across the subdomains. The resulting matrices are the local Schur complements

$$S_i : W_i \to W_i; \quad S_i := K_i^{b_i b_i} - K_i^{b_i I_i} (K_i^{I_i I_i})^{-1} K_i^{I_i b_i}, \quad i = 1, \ldots, N, \tag{5.13}$$

and the Schur complement on the product space is

$$S : \underbrace{W_1 \times \ldots W_N}_{W} \to \underbrace{W_1 \times \ldots W_N}_{W}; \quad S := \begin{pmatrix} S_1 & 0 & \ldots & 0 \\ 0 & S_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & S_N \end{pmatrix} \tag{5.14}$$

so that

$$Su = (S_1 u_1, \ldots, S_N u_N)^\top, \quad \forall\, u = (u_1, \ldots, u_N)^\top \in W. \tag{5.15}$$

The Schur complement S on the product space W admits the following counterpart $\hat{S}$ for functions in $\hat{W}$:

$$\hat{S} : \hat{W} \to \hat{W}; \quad \hat{S}u := \sum_{i=1}^{N} R_i^\top S_i R_i u. \tag{5.16}$$

We notice that this is the usual Schur complement for the global problem reduced to the set Γ of internal boundaries:

$$\hat{S} = \hat{K}^{bb} - \hat{K}^{bI} (\hat{K}^{II})^{-1} \hat{K}^{Ib}, \tag{5.17}$$

where $\hat{K}^{bb}$, $\hat{K}^{bI}$, $\hat{K}^{II}$ and $\hat{K}^{Ib}$ are the components in the bloc formulation of $\hat{K}$

$$\hat{K} = \begin{pmatrix} \hat{K}^{bb} & \hat{K}^{bI} \\ \hat{K}^{Ib} & \hat{K}^{II} \end{pmatrix}, \quad b := dof(\Gamma) \text{ and } I := dof(\Omega) \setminus dof(\Gamma). \tag{5.18}$$

Lumped matrices In the FETI literature the lumped version of the stiffness matrix is the extraction of the entries in the stiffness matrix which correspond to boundary degrees of freedom. We have already introduced $\hat{K}^{bb}$ and $K_i^{b_i b_i}$, let K^{bb} be the counterpart on the product space W:

$$K^{bb} : \underbrace{W_1 \times \ldots W_N}_{W} \to \underbrace{W_1 \times \ldots W_N}_{W}; \quad K^{bb} := \begin{pmatrix} K_1^{b_1 b_1} & 0 & \ldots & 0 \\ 0 & K_2^{b_2 b_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & K_N^{b_N b_N} \end{pmatrix}. \tag{5.19}$$

We notice that $\hat{K}^{bb} = \sum_{i=1}^{N} R_i^\top K_i^{b_i b_i} R_i$ and the next Lemma gives an important relation between lumped matrices and Schur complement matrices.

Lemma 5.6. For any $\hat{u} \in \hat{W}$ and any $u \in W$ the following inequalities hold

$$\langle \hat{S}\hat{u}, \hat{u} \rangle \leq \langle \hat{K}^{bb}\hat{u}, \hat{u} \rangle \quad \text{and} \quad \langle Su, u \rangle \leq \langle K^{bb}u, u \rangle.$$

Proof. Let $\hat{u} \in \hat{W}$. Then by definition of $\hat{S}$

$$\langle \hat{S}\hat{u}, \hat{u} \rangle = \langle (\hat{K}^{bb} - \hat{K}^{bI}(\hat{K}^{II})^{-1}\hat{K}^{Ib})\hat{u}, \hat{u} \rangle = \langle \hat{K}^{bb}\hat{u}, \hat{u} \rangle - \langle (\hat{K}^{II})^{-1}\hat{K}^{Ib}\hat{u}, \hat{K}^{Ib}\hat{u} \rangle.$$

The first inequality follows by noticing that $\langle (\hat{K}^{II})^{-1}\hat{K}^{Ib}\hat{u}, \hat{K}^{Ib}\hat{u} \rangle \geq 0$ because $(\hat{K}^{II})^{-1}$ is spd. For the second, let $u \in W$. Then by definition of S

$$\begin{aligned} \langle Su, u \rangle = \sum_{i=1}^{N} \langle S_i u_i, u_i \rangle &= \sum_{i=1}^{N} \langle (K_i{}^{b_i b_i} - K_i{}^{b_i I_i}(K_i{}^{I_i I_i})^{-1} K_i{}^{I_i b_i}) u_i, u_i \rangle \\ &= \langle K^{bb} u, u \rangle - \sum_{i=1}^{N} \langle (K_i{}^{I_i I_i})^{-1} K_i{}^{I_i b_i} u_i, K_i{}^{I_i b_i} u_i \rangle. \end{aligned}$$

And the second inequality follows by noticing that $\langle (K_i{}^{I_i I_i})^{-1} K_i{}^{I_i b_i} u_i, K_i{}^{I_i b_i} u_i \rangle \geq 0$ for any $i = 1, \ldots, N$ because $(K_i{}^{I_i I_i})^{-1}$ is spd. □

Right hand sides In order to reduce the problem to the set of interfaces between subdomains, we define the following right hand side

$$\hat{f}_\Gamma := \hat{f}^b - \hat{K}^{bI}(\hat{K}^{II})^{-1}\hat{f}^I, \tag{5.20}$$

which is the right hand side of the original problem (5.4) condensed onto the degrees of freedom in $\hat{W}$. As for the right hand side on the product space W, for each subdomain $i = 1, \ldots, N$: first let f_i be the local right hand side given by (5.8) with $D = \Omega_i$. Then condense it onto the interfaces following: $f_{\Gamma,i} := f_i^{b_i} - K_i^{b_i I_i}(K_i^{I_i I_i})^{-1} f_i^{I_i}$. (We have used the identification between the finite element representation of f_i and its vector representation.) Finally, the right hand side for the problem condensed onto the space W is

$$f_\Gamma = \begin{pmatrix} f_{\Gamma,1} \\ \cdots \\ f_{\Gamma,N} \end{pmatrix}. \tag{5.21}$$

Most of this notation is summed up in Table 5.1 at the beginning of the article. Some comments are given in subsection 5.2.4, along with an important lemma on which of these matrices are positive definite.

Remark 5.7. Assumption 5.1 is actually stronger than what we really need but enables the use of any partition into subdomains and allowed us to define each component of the algorithm thoroughly. For a given non overlapping partition into subdomains it is enough to have access to the local matrices K_i on each subdomain, the local right hand sides f_i, the local-global interpolation operators $R_i^\top$ and the information on the boundary of each subdomain $\Gamma \cap \partial\Omega_i$.

5.2.3 Partition of unity and weighted operators

An important role in the description of the BDD algorithms is played by a weighting (counting) function on W. As in the original GenEO algorithm [105, 106] this induces partition of unity operators Ξ_i which act directly on the degrees of freedom of the finite element functions.

Definition 5.8 (Partition of unity)**.** Let $\mu = (\mu_1, \ldots, \mu_N) \in W$ be a *discrete* partition of unity:

$$\sum_{i=1,\ldots,N} R_i^\top \mu_i = 1_{|\hat{W}}, \text{ where } 1_{|\hat{W}} \in \hat{W} \text{ and all vector entries are 1.}$$

Then for any function $u_i \in W_i$ written as

$$u_i = \sum_{k \in dof(\Gamma \cap \partial\Omega_i)} \alpha_i^k \, \phi_{k|\Gamma\cap\partial\Omega_i}, \quad \alpha_i^k \in \mathbb{R},$$

the local partition of unity operator $\Xi_i : W_i \to W_i$ is defined by:

$$\Xi_i(u_i) := \sum_k \mu_i^k \alpha_i^k \, \phi_{k|\Gamma\cap\partial\Omega_i},$$

where μ_i^k is the k-th entry in μ_i. The inverse $\Xi_i^{-1} : W_i \to W_i$ is defined by:

$$\Xi_i^{-1}(u_i) := \sum_k \frac{1}{\mu_i^k} \alpha_i^k \, \phi_{k|\Gamma\cap\partial\Omega_i}.$$

It is clear that the Ξ_i define a partition of unity from $\hat{W}$ onto the product space $W = W_1 \times \cdots \times W_N$ in the sense that

$$u = \sum_{i=1}^N R_i^\top \underbrace{\Xi_i(R_i u)}_{\in W_i}, \quad \forall\, u \in \hat{W}.$$

It is also clear that Ξ_i^{-1} is the inverse of Ξ_i since any $u_i \in W_i$ satisfies $\Xi_i^{-1}(\Xi_i(u_i)) = \Xi_i(\Xi_i^{-1}(u_i)) = u_i$.

Remark 5.9. Two common choices for μ are the multiplicity scaling where μ_i^k is chosen as $(\#\{i = 1, \ldots, N; k \in dof(\Gamma \cap \partial\Omega_i)\})^{-1}$ and the K-scaling where μ depends on the diagonal entries of the stiffness matrices [93, 58]. In the numerical result section we mostly use K-scaling.

We introduce the local bilinear forms which correspond to the local Schur complements S_i as follows. For $i = 1, \ldots, N$ define

$$s_i : W_i \times W_i \to \mathbb{R}, \; s_i(u_i, v_i) := \langle S_i u_i, v_i \rangle; \quad \forall u_i, v_i \in W_i. \tag{5.22}$$

Next we use the partition of unity operators to define weighted versions of the Schur complements which will be instrumental in defining the BDD algorithm.

Definition 5.10 (Weighted Schur complements)**.** For any $i = 1, \ldots, N$, let $\tilde{s}_i : W_i \times W_i \to \mathbb{R}$ be the bilinear form defined by

$$\tilde{s}_i(u_i, v_i) := s_i(\Xi_i^{-1}(u_i), \Xi_i^{-1}(v_i)); \quad \forall\, u_i,\, v_i \in W_i, \tag{5.23}$$

where s_i is the local Schur complement, and Ξ_i^{-1} is the inverse partition of unity operator introduced in Definition 5.8.
Next, let the matrix $\tilde{S}_i : W_i \to W_i$ be the matrix counterpart of $\tilde{s}_i$:

$$\langle \tilde{S}_i u_i, v_i \rangle := \tilde{s}_i(u_i, v_i).$$

5.2.4 Summary of the notation and complements

We have introduced quite a lot of notation. Table 5.1 at the beginning of the article sums up most of the notation which will appear in the description of the algorithms and the reference to where it is first introduced. Some of the operators are introduced for the first time ($\hat{a}^{bb}$, a^{bb}, $\hat{s}$ and s) as the bilinear forms associated with a matrix. More precisely, let $\hat{a}^{bb}$ and $\hat{s}$ be defined as

$$\hat{a}^{bb} : \hat{W} \times \hat{W} \to \mathbb{R};\ \hat{a}^{bb}(\hat{u}, \hat{v}) := \langle \hat{K}^{bb}\hat{u}, \hat{v}\rangle \quad \text{and} \quad \hat{s} : \hat{W} \times \hat{W} \to \mathbb{R};\ \hat{s}(\hat{u}, \hat{v}) := \langle \hat{S}\hat{u}, \hat{v}\rangle,$$

for any $\hat{u}$ and $\hat{v} \in \hat{W}$, and let a^{bb} and s be defined as

$$a^{bb} : W \times W \to \mathbb{R};\ a^{bb}(u, v) := \langle K^{bb}u, v\rangle \quad \text{and} \quad s : W \times W \to \mathbb{R};\ s(u, v) := \langle Su, v\rangle,$$

for any u and $v \in W$.

The operators with a $\hat{\cdot}$ always correspond to functions defined either on the whole of Ω or the whole of Γ. The subscript i always refers to a local operator defined on a subdomain Ω_i or its boundary. Operators without a $\hat{\cdot}$ or a subscript i are defined on the product spaces. Finally operators $\tilde{S}_i$ are weighted by the inverse partition of unity operators.

In many cases the local stiffness matrices K_i are not spd on all floating subdomains. (A floating subdomain is a subdomain which does not *touch* the Dirichlet part of the boundary). For example, in the case of the Darcy equation, the kernel of K_i for a floating subdomain is the set of constant functions. In the case of linear elasticity, the kernel of K_i is the set of rigid body motions. It is easy to see that these kernels induce kernels for the corresponding Schur complements S_i as well as their weighted counterparts $\tilde{S}_i$ and, possibly, the lumped matrices $K_i^{b_i b_i}$.The next lemma makes precise which matrices are positive definite. They are all symmetric positive semi definite.

Lemma 5.11. The stiffness matrix K, lumped stiffness matrix K^{bb} and Schur complement S, which correspond to the product spaces, can be singular. Their respective counterparts, $\hat{K}$, $\hat{K}^{bb}$ and $\hat{S}$, on the original spaces of functions $W_h(\Omega)$ and $\hat{W}$ are symmetric positive definite. Finally, under Assumption 5.5 each of the local matrices $R_i\hat{K}^{bb}R_i^\top$ and $R_i\hat{S}R_i^\top$ is also symmetric positive definite.

Proof. The fact that $\hat{K}$ and $\hat{S}$ are positive definite is clear because the original problem is well posed. The positive definiteness of $\hat{K}^{bb}$ follows from Lemma 5.6 and the positive definiteness of $\hat{S}$: let $u \in \hat{W}$

$$\langle \hat{K}^{bb}u, u\rangle = 0 \Rightarrow \langle \hat{S}u, u\rangle = 0 \Rightarrow u = 0.$$

The positive definiteness of $R_i\hat{S}R_i^\top$ and $R_i\hat{K}^{bb}R_i^\top$ is obvious from the positive definiteness of $\hat{K}$ and $\hat{S}$ and (5.10) which is a direct consequence of Assumption 5.5. □

Remark 5.12. Note that in nearly all practical cases K^{bb} is also symmetric positive definite.

We are now ready to introduce the BDD preconditioner.

5.3 Balancing Domain Decomposition

The problem which we solve is the original problem (5.4) reduced to the set Γ of interfaces between subdomains: find $u \in \hat{W}$ such that

$$\hat{S}u = \hat{f}_\Gamma. \tag{5.24}$$

5.3.1 One level BDD in the abstract Schwarz framework

The only thing that is needed in order to define the one-level preconditioner is a solver on each subdomain. Then we will precondition the global problem (5.24) with a sum of these local solves. The usual BDD strategy is to use the weighted Schur complements $\tilde{S}_i$ introduced in Definition 5.10 to build local problems. Then each local solve is the solution of a Neumann problem: $\tilde{S}_i^\dagger$.

Definition 5.13 (One level preconditioner)**.** For each $i = 1, \dots, N$, let $\tilde{P}_i$ and P_i be defined as

$$\tilde{P}_i := \tilde{S}_i^\dagger R_i \hat{S} \quad \text{and} \quad P_i := R_i^\top \tilde{P}_i, \tag{5.25}$$

where $\tilde{S}_i^\dagger$ is a pseudo inverse of $\tilde{S}_i$. Equivalently for any $u \in \hat{W}$, $\tilde{P}_i u$ is the unique vector in range$(\tilde{S}_i^\dagger)$ which satisfies

$$\tilde{s}_i(\tilde{P}_i u, v_i) = \hat{s}(u, R_i^\top v_i), \quad \forall\, v_i \in W_i. \tag{5.26}$$

The one-level preconditioner is the sum of local solves $\sum_{i=1}^N R_i^\top \tilde{S}_i^\dagger R_i$ so the one-level preconditioned operator is $\sum_{i=1}^N P_i$.

The next lemma gives a lower bound for the eigenvalues of the one-level preconditioned operator. It does not depend on the specific choice of the pseudo inverse or on any coarse space.

Essentially what we do is check that a stable splitting assumption (Assumption 2.2 in [112]) holds on the whole of $\hat{W}$. Then we give the result of Lemma 2.5 in [112] which is that this implies a lower bound for the condition number of the one-level preconditioned operator. One of the assumptions in [112] is that the local bilinear forms ($\tilde{S}_i$ in this case) be positive definite. Here they are only positive semi definite but the proof goes through in the exact same way so we don't give it again.

Lemma 5.14 (Stable splitting – Lower bound for the eigenvalues of the preconditioned operator)**.** For any $u \in \hat{W}$ there exists a stable splitting $(v_1, \dots, v_N)$ of u onto $W = W_1 \times \cdots \times W_N$:

$$u = R_1^\top v_1 + \cdots + R_N^\top v_N;\ v_i \in W_i \ \text{ and } \sum_{i=1}^N \tilde{s}_i(v_i, v_i) \leq \hat{s}(u, u). \tag{5.27}$$

This implies that the one-level preconditioned operator satisfies

$$\hat{s}(u, u) \leq \hat{s}\left(\sum_{i=1}^N P_i u, u\right) \ \text{ for any } u \in \hat{W}. \tag{5.28}$$

Proof. Let $u \in \hat{W}$. The fact that, by definition, the operators Ξ_i define a partition of unity allows us to write an obvious splitting of u onto W:

$$(v_i := \Xi_i(R_i u), \quad \forall\, i = 1, \dots, N) \text{ Â} \quad \Rightarrow \quad u = \sum_{i=1}^N R_i^\top v_i \quad .$$

We prove (5.27) for this splitting using only the definitions of $\tilde{s}_i$ and $\hat{s}$:

$$\sum_{i=1}^N \tilde{s}_i(v_i, v_i) = \sum_{i=1}^N s_i(\Xi_i^{-1}(\Xi_i(R_i u)), \Xi_i^{-1}(\Xi_i(R_i u)) = \sum_{i=1}^N s_i(R_i u, R_i u) = \hat{s}(u, u).$$

The second part of the lemma is the result of Lemma 2.5 in [112], we refer the reader to there for the proof. □

The fact that (5.28) provides a lower bound for the eigenvalues of the preconditioned operator $\sum_{i=1}^N P_i$ is easy to see: suppose u is an eigenvector associated with eigenvalue λ, then

$$\sum_{i=1}^N P_i u = \lambda u \Rightarrow \hat{S}\sum_{i=1}^N P_i u = \lambda \hat{S} u \Rightarrow \hat{s}(\sum_{i=1}^N P_i u, u) = \lambda \hat{s}(u,u),$$

and (5.28) implies that $\lambda \geq 1$.

In other words the lower bound for the eigenvalues of the preconditioned operator does not depend on the choice of the coarse space. This is a big difference with the Additive Schwarz method where the proof of a lower bound depends very strongly on the choice of the coarse space and on restrictive assumptions on the coefficient distribution. This is why the Schwarz-GenEO strategy in [106] is precisely to build an enriched coarse space for which the stable splitting property and thus a lower bound for the spectrum of the preconditioned operator hold regardless of the partition into subdomains and the coefficient distribution. Luckily, the upper bound for the eigenvalues of the Additive Schwarz operator depends only on the number of neighbours of each subdomain enabling the proof of a bound for the condition number of the preconditioned operator.

Here the situation is reversed: Lemma 5.14 gives a lower bound for the eigenvalues of the preconditioned operator which does not depend on the choice of the coarse space thanks to the adequate weighting of the local solvers. However the upper bound requires more work and with the usual coarse space it can only be independent of the coefficients in the equation if some assumptions on the coefficient distribution are satisfied. The GenEO strategy will enable us to waive all of these assumptions.

5.3.2 GenEO coarse space for BDD

The abstract Schwarz theory tells us that the upper bound for the eigenvalues of the preconditioned operator is implied by the stability of the local solvers $\tilde{s}_i$ on the local subspaces once the coarse components have been removed (this is made explicit in Lemma 5.20). This is where the GenEO strategy comes in. We solve a generalized eigenvalue problem which identifies the 'bad' modes: in this case those for which we cannot ensure that the local solver is stable for a constant independent of the coefficients in the equations. These 'bad' modes are then used to span the coarse space, and the local solvers are stable on all remaining local components (the 'good' components). More precisely, the next two definitions introduce the generalized eigenvalue problem, the coarse space and the corresponding two-level BDD-GenEO preconditioners.

Definition 5.15 (GenEO coarse space for BDD)**.** For each subdomain $i = 1, \ldots, N$, find the eigenpairs $(p_i^k, \lambda_i^k) \in W_i \times \mathbb{R}^+$ of the generalized eigenvalue problem:

$$\boxed{\tilde{s}_i(p_i^k, v_i) = \lambda_i^k \, \hat{a}^{bb}(R_i^\top p_i^k, R_i^\top v_i)} \quad \text{for any } v_i \in W_i. \tag{5.29}$$

Next, given a threshold $\mathcal{K}_i > 0$ for each subdomain, define the coarse space as

$$W_0 = \text{span}\{R_i^\top p_i^k \,;\, \lambda_i^k < \mathcal{K}_i,\, i = 1, \ldots, N\} \quad \left(\subset \hat{W}\right). \tag{5.30}$$

Let the interpolation operator $R_0^\top$ be the matrix whose columns are the coarse basis functions $\{R_i^\top p_i^k \,;\, \lambda_i^k < \mathcal{K}_i,\, i = 1, \ldots, N\}$. Finally, let the coarse solver be the exact solver on W_0:

$$S_0 := R_0 \hat{S} R_0^\top,$$

and P_0 be the $\hat{S}$-orthogonal projection operator defined by

$$P_0 := R_0^\top S_0^\dagger R_0 \hat{S}. \tag{5.31}$$

This definition gives rise to a few immediate remarks.

Remark 5.16.

(i) The operator $R_0^\top$ is a mapping between the coordinates of a vector from W_0 in the set of coarse basis functions and its representation in $\hat{W}$ ($\text{range}(R_0^\top) \subset \hat{W}$). Its transpose R_0 is a restriction operator which maps an element in $\hat{W}$ to the coordinates of its l_2 projection onto W_0 in the set of coarse basis functions.

(ii) Eigenvalue 0 for eigenproblem (5.29) is associated with the kernel of $\tilde{s}_i$ so in some sense the coarse space will take care of the fact that $\tilde{s}_i$ is not necessarily coercive. Note that if the coarse space includes only the kernel of $\tilde{s}_i$, one obtains the usual coarse space for BDD.

(iii) In the definition of P_0 we used a pseudo inverse $S_0^\dagger$ because the columns of $R_0^\top$ are not necessarily linearly independent. The pseudo inverse is defined up to an element in $\text{Ker}(R_0^\top)$ and the specific choice of the pseudo inverse makes no difference because the application of $S_0^\dagger$ is followed by an application of $R_0^\top$.

(iv) The fact that P_0 is an $\hat{S}$-orthogonal projection can be proved easily using the definitions of P_0 and S_0 and it is equivalent to the fact that P_0 is self adjoint with respect to S_0.

We are now ready to introduce the BDD-GenEO preconditioner. There are mainly two ways to add the second level once that we have chosen the coarse space: either we use the balanced preconditioner (5.33) with the preconditioned conjugate gradient (PCG) algorithm or we use the projected preconditioned conjugate gradient (PPCG) algorithm in the space $\text{range}(I - P_0)$ with the deflated preconditioner (5.32). Both alternatives will lead to essentially identical convergence bounds. In fact for certain starting vectors in exact arithmetic they produce the same iterates. Comparison (both theoretical and in terms of implementation) between (5.33), (5.32) and some other variants can be found in [110] or, more specifically for FETI and BDD, in [56]. The deflated preconditioner is the more natural since it simply restricts the problem to a smaller subspace. However it suffers from robustness problems when the coarse solves are not sufficiently accurate. The balanced preconditioner is more robust but its application is slightly more expensive.

Definition 5.17 (Two-level preconditioners)**.** Recall that, according to (5.25) and (5.31), we have defined $P_i = R_i^\top \tilde{S}_i^\dagger R_i \hat{S}$ for any $i = 1, \ldots, N$ and $P_0 = R_0^\top S_0^\dagger R_0 \hat{S}$. Then define the deflated preconditioned operator as

$$P_{def} := \sum_{i=1}^{N} (I - P_0)^\top P_i (I - P_0), \tag{5.32}$$

and the balanced preconditioned operator as

$$P_{bal} := P_0 + \sum_{i=1}^{N} (I - P_0)^\top P_i (I - P_0). \tag{5.33}$$

In the remainder of this subsection we show that the BDD-GenEO coarse space leads to an upper bound for the eigenvalues of the preconditioned operators which does not depend on the number of subdomains or the coefficients in the equations. Instead it depends on the thresholds $\mathcal{K}_i$ which were introduced to select the coarse basis functions. First we give some properties of the family of generalized eigenvectors (Lemma 5.18). Then we use these properties to show that the local bilinear forms are stable on the deflated local subspaces (Lemma 5.20) and the upper bound follows from there (Lemma 5.22).

Lemma 5.18. For a given subdomain $i = 1, \dots, N$, the eigenpairs (p_i^k, λ_i^k) of generalized eigenproblem (5.29) can be chosen so that the set $\{p_i^k\}_k$ of eigenvectors is an orthonormal basis of W_i with respect to the inner product induced $\hat{a}^{bb}(R_i^\top \cdot, R_i^\top \cdot)$. This can be written as

$$\hat{a}^{bb}(R_i^\top p_i^k, R_i^\top p_i^k) = 1; \quad \text{and} \quad \hat{a}^{bb}(R_i^\top p_i^k, R_i^\top p_i^{k'}) = 0,\ k \neq k'.$$

An orthogonality type property with respect to $\tilde{s}_i$ (which is not necessarily coercive) also holds:

$$\tilde{s}_i(p_i^k, p_i^{k'}) = 0,\ k \neq k'.$$

Proof. Lemma 5.11 tells us that $R_i \hat{K}^{bb} R_i^\top$ is positive definite on W_i so we may indeed speak of a $\hat{a}^{bb}(R_i^\top \cdot, R_i^\top \cdot)$ orthonormal basis of W_i. The proof is an application of Lemma 2.13. □

Remark 5.19. The fact that the generalized eigenproblem (5.29) is equivalent to a non-generalized eigenproblem implies that all eigenvalues are finite. Because both matrices are symmetric positive semi definite, the eigenvalues are also non negative: for any k, $0 \leq \lambda_i^k < +\infty$.

The next lemma states that the local solvers are stable and strongly relies on the definition of the GenEO coarse space. In fact the purpose of the GenEO strategy is specifically to ensure that Lemma 5.20 holds. This corresponds to Assumption 2.4 in [112].

Lemma 5.20 (Stability of the local solvers)**.** Suppose the pseudo inverse $\tilde{S}_i^\dagger$ in Definition 5.13 is chosen such that $\text{range}(\tilde{S}_i^\dagger) = \text{span}\{p_i^k; \lambda_i^k > 0\}$. Then for any $i = 1, \dots, N$, the local solvers are stable in the sense

$$\hat{s}(R_i^\top u_i, R_i^\top u_i) \leq \frac{1}{\mathcal{K}_i} \tilde{s}_i(u_i, u_i), \quad \forall\, u_i \in \text{range}(\tilde{P}_i(I - P_0)),$$

where the $\mathcal{K}_i$ are the thresholds that were used to select eigenvectors for the coarse space in Definition 5.15.

Proof. We may indeed choose $\text{range}(\tilde{S}_i^\dagger) = \text{span}\{p_i^k; \lambda_i^k > 0\}$ because the pseudo inverse of an operator is defined up to an element in the kernel of this operator. Precisely there are an infinity of pseudo inverse and we may choose the range of $\tilde{S}_i^\dagger$ among all the spaces which satisfy $\text{range}(\tilde{S}_i^\dagger) \oplus \text{Ker}(\tilde{S}_i) = W_i$. Here, $\text{Ker}(\tilde{S}_i) = \text{span}\{p_i^k; \lambda_i^k = 0\}$ and the set of all p_i^k is a basis of W_i so our choice fits this limitation.

Next we prove that

$$\text{range}(\tilde{P}_i(I - P_0)) \left(= \text{range}(\tilde{S}_i^\dagger R_i \hat{S}(I - P_0))\right) \subset \{p_i^k;\ \lambda_i^k \geq \mathcal{K}_i\}.$$

We will use the following linear algebra identity:

$$\text{Ker}((I - P_0)^\top \hat{S} R_i^\top) \oplus^\perp \text{range}(R_i \hat{S}(I - P_0)) = W_i, \tag{5.34}$$

where the symbol $\perp$ refers to the l_2 orthogonality between both spaces and $\oplus$ means that the sum is direct. By definition (5.31) of P_0, $(I - P_0)^\top = I - \hat{S} R_0^\top S_0^\dagger R_0$ so

$$\mathrm{range}(\hat{S} R_0^\top) \subset \mathrm{Ker}((I - P_0)^\top).$$

In particular, for a given $i = 1, \ldots, N$: $\mathrm{span}\{\hat{S} R_i^\top p_i^k; \lambda_i^k < \mathcal{K}_i\} \subset \mathrm{Ker}((I - P_0)^\top)$, which implies

$$\mathrm{span}\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\} \subset \mathrm{Ker}((I - P_0)^\top \hat{S} R_i^\top). \tag{5.35}$$

Next we use another linear algebra identity: W_i is finite dimensional so

$$\mathrm{span}\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\} \oplus^\perp \left\{\mathrm{span}\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\}\right\}^\perp = W_i. \tag{5.36}$$

According to Lemma 5.18 the $\{p_i^k\}_k$ form a $R_i \hat{K}^{bb} R_i^\top$-orthonormal basis of W_i so

$$\langle p_i^k, R_i \hat{K}^{bb} R_i^\top p_i^{k'} \rangle = 0,\ \forall k \neq k'.$$

This implies that $\mathrm{span}\{R_i \hat{K}^{bb} R_i^\top p_i^k;\ \lambda_i^k \geq \mathcal{K}_i\} \subset \left\{\mathrm{span}\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\}\right\}^\perp$. The equality between these subsets follows by a dimensional argument: the set $\{p_i^k\}_k$ forms a basis of W_i and $R_i \hat{K}^{bb} R_i^\top$ is spd so

$$\mathrm{rank}\{R_i \hat{K}^{bb} R_i^\top p_i^k; \lambda_i^k \geq \mathcal{K}_i\} = \mathrm{rank}\{p_i^k; \lambda_i^k \geq \mathcal{K}_i\} = \mathrm{rank}\left\{\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\}^\perp\right\},$$

and in turn the inclusion becomes an equality:

$$\mathrm{span}\{R_i \hat{K}^{bb} R_i^\top p_i^k; \lambda_i^k \geq \mathcal{K}_i\} = \left\{\mathrm{span}\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\}\right\}^\perp.$$

Injecting this into (5.36) implies

$$\mathrm{span}\{p_i^k;\ \lambda_i^k < \mathcal{K}_i\} \oplus^\perp \mathrm{span}\{R_i \hat{K}^{bb} R_i^\top p_i^k; \lambda_i^k \geq \mathcal{K}_i\} = W_i. \tag{5.37}$$

Putting together (5.34), (5.35) and (5.37) we get

$$\mathrm{range}(R_i \hat{S}(I - P_0)) \subset \mathrm{span}\{R_i \hat{K}^{bb} R_i^\top p_i^k; \lambda_i^k \geq \mathcal{K}_i\},$$

where the argument is:

$$(E_1 \oplus^\perp E_2 = E_3 \oplus^\perp E_4 \,\mathrm{and}\, E_1 \subset E_3) \Rightarrow E_4 \subset E_2,$$

for any vector spaces $E_1, \ldots, E_4$.

By definition of eigenproblem (5.29), $\lambda_i^k R_i \hat{K}^{bb} R_i^\top p_i^k = \tilde{S}_i p_i^k$ so

$$\mathrm{range}(R_i \hat{S}(I - P_0)) \subset \mathrm{span}\{\tilde{S}_i p_i^k; \lambda_i^k \geq \mathcal{K}_i\}.$$

Finally, for the specific choice of the pseudo inverse $\tilde{S}_i^\dagger$ it follows that

$$\mathrm{range}(\tilde{S}_i^\dagger R_i \hat{S}(I - P_0)) \left(= \mathrm{range}(\tilde{P}_i(I - P_0))\right) \subset \mathrm{span}\{p_i^k;\ \lambda_i^k \geq \mathcal{K}_i\}.$$

Now we prove the inequality in the lemma. Any $u_i \in \mathrm{range}(\tilde{P}_i(I - P_0))$ can be written as $u_i = \sum_{\{k;\lambda_i^k \geq \mathcal{K}_i\}} \alpha_i^k p_i^k$ for some coefficients $\alpha_i^k \in \mathbb{R}$. From Lemma 5.6, it is obvious that

$$\hat{s}(R_i^\top u_i, R_i^\top u_i) \leq \hat{a}^{bb}(R_i^\top u_i, R_i^\top u_i) = \hat{a}^{bb}\left(R_i^\top \sum_{\{k;\lambda_i^k \geq \mathcal{K}_i\}} \alpha_i^k p_i^k \quad , \quad R_i^\top \sum_{\{k;\lambda_i^k \geq \mathcal{K}_i\}} \alpha_i^k p_i^k\right).$$

Using successively the first orthogonality property in Lemma 5.18, the definition of the eigenproblem and the second orthogonality property in Lemma 5.18 we get

$$\begin{aligned}\hat{s}(R_i^\top u_i, R_i^\top u_i) &\leq \sum_{\{k;\lambda_i^k \geq \mathcal{K}_i\}} {\alpha_i^k}^2 \hat{a}^{bb}(R_i^\top p_i^k, R_i^\top p_i^k) \\ &= \sum_{\{k;\lambda_i^k \geq \mathcal{K}_i\}} \frac{1}{\lambda_i^k} {\alpha_i^k}^2 \tilde{s}_i(p_i^k, p_i^k) \\ &\leq \frac{1}{\mathcal{K}_i} \sum_{\{k;\lambda_i^k \geq \mathcal{K}_i\}} {\alpha_i^k}^2 \tilde{s}_i(p_i^k, p_i^k) \\ &= \frac{1}{\mathcal{K}_i} \tilde{s}_i(u_i, u_i).\end{aligned}$$

□

Remark 5.21 (Local stability, Exact solvers, and Choice of the eigenproblem)**.** The bilinear form on the left hand side of the inequality in the lemma is $\hat{s}(R_i^\top \cdot, R_i^\top \cdot)$. This is the so called exact solver on subdomain i for the global problem given by $\hat{S}$. The exact solvers are by definition the solvers which are used to build the Additive Schwarz preconditioner. For the problem $\hat{S}u = \hat{f}_\Gamma$ the Additive Schwarz preconditioner would be $\sum_{i=1}^N R_i^\top \hat{S} R_i$. If these exact solvers were used instead of $\tilde{S}_i$ the upper bound for the eigenvalues of the preconditioned operator would depend only on a constant related to the number of neighbours (introduced in the next lemma). The nice bound that we have for the lowest eigenvalue of the preconditioned operator would no longer hold though. The most straightforward generalized eigenproblem which arises from the theory is

$$\tilde{s}_i(p_i^k, v_i) = \lambda_i^k \, \hat{s}(R_i^\top p_i^k, R_i^\top v_i) \quad \text{for any} \quad v_i \in W_i, \tag{5.38}$$

so the eigensolve operates some sort of spectral comparison between the exact solver (on the right) and the one which we actually use (on the left). We then isolate the modes for which the chosen preconditioner is not a good enough approximation in the coarse space and use a direct solve on these modes. It is however expensive to assemble and to solve (5.38). This is is why in this article we have chosen to go through only with eigenproblem (5.29) where $\hat{s}$ is replaced by $\hat{a}^{bb}$. For a coarse space based on Eigenproblem (5.38) the theory goes through to the exact same final estimate simply by replacing $\hat{a}^{bb}$ by $\hat{s}$ in the proofs.

The following lemma gives a consequence of the stability of the local solvers. It is very narrowly related to Lemma 2.6 in [112].

Lemma 5.22 (Upper bound for the eigenvalues of the preconditioned operator)**.** The stability of each of the local solvers which was proved in Lemma 5.20 implies

$$\hat{s}\left(\sum_{i=1}^N P_i u, u\right) \leq \mathcal{N} \max_{1 \leq i \leq N} \left(\frac{1}{\mathcal{K}_i}\right) \hat{s}(u, u) \quad \forall \, u \in \text{range}(I - P_0),$$

where $\mathcal{N}$ is the maximal number of neighbours of a subdomain (including itself) in the sense:

$$\mathcal{N} := \max_{1 \leq i \leq N} \left(\#\{j; R_j R_i^\top \neq 0\} \right).$$

Proof. This is basically the proof of Lemma 2.6 in [112] but where we have chosen not to rely on strengthened Cauchy Schwarz inequalities. Instead we make the number of neighbours of a subdomain appear explicitly. Let $u \in \text{range}(I - P_0)$, then

$$\begin{aligned}\hat{s}(P_i u, P_i u) &= \hat{s}(R_i^\top \tilde{P}_i u, R_i^\top \tilde{P}_i u) \\ &\leq \frac{1}{\mathcal{K}_i} \tilde{s}_i(\tilde{P}_i u, \tilde{P}_i u) \quad \text{(Lemma 5.20)} \\ &= \frac{1}{\mathcal{K}_i} \hat{s}(u, R_i^\top \tilde{P}_i u) \quad \text{(definition of } \tilde{P}_i \text{ (5.26))} \\ &= \frac{1}{\mathcal{K}_i} \hat{s}(u, P_i u).\end{aligned}$$

We use the fact that $P_i = R_i^\top \tilde{P}_i$ and the definition of $\hat{s}$ to write

$$\hat{s}(P_i u, u) = \sum_{j=1}^{N} s_j(R_j R_i^\top \tilde{P}_i, R_j u) = \sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j R_i^\top \tilde{P}_i, R_j u).$$

We apply the Cauchy Schwarz inequality first for s_j then for the Euclidean inner product to this and inject the previous result (in the last step)

$$\begin{aligned}\hat{s}(P_i u, u) &\leq \sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j R_i^\top \tilde{P}_i, R_j R_i^\top \tilde{P}_i)^{1/2} s_j(R_j u, R_j u)^{1/2} \\ &\leq \left[\sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j R_i^\top \tilde{P}_i, R_j R_i^\top \tilde{P}_i)\right]^{1/2} \left[\sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j u, R_j u)\right]^{1/2} \\ &= \hat{s}(P_i u, P_i u)^{1/2} \left[\sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j u, R_j u)\right]^{1/2} \\ &\leq \left(\frac{1}{\mathcal{K}_i} \hat{s}(u, P_i u)\right)^{1/2} \left[\sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j u, R_j u)\right]^{1/2}.\end{aligned}$$

Raising to the square and simplifying by $\hat{s}(P_i u, u)$ yields

$$\hat{s}(P_i u, u) \leq \frac{1}{\mathcal{K}_i} \left[\sum_{\{j; R_j R_i^\top \neq 0\}} s_j(R_j u, R_j u)\right].$$

Finally summing these inequalities over i gives the result. □

5.3.3 Main theorem: convergence bound for BDD with GenEO

We are now ready to give the estimates for the condition number of BDD with the GenEO coarse space.

Theorem 5.23 (Main theorem for BDD with the GenEO coarse space)**.** The condition number for BDD solved in $\text{range}(I - P_0)$ with the deflated operator (5.32) satisfies

$$\kappa\left(P_{def}\right) \leq \mathcal{N} \max_{1 \leq i \leq N} \left(\frac{1}{\mathcal{K}_i}\right). \tag{5.39}$$

As for the condition number of the balanced operator (5.33) with the GenEO coarse space, it satisfies

$$\kappa\left(P_{bal}\right) \leq \max\left\{1, \mathcal{N} \max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\right\}. \tag{5.40}$$

These bounds depend only on the chosen thresholds $\mathcal{K}_i$ which we use to select eigenvectors for the coarse space in Definition 5.15 and on the maximal number $\mathcal{N}$ of neighbours of a subdomain:

$$\mathcal{N} = \max_{1\leq i\leq N}\left(\#\{j; R_j R_i^\top \neq 0\}\right).$$

Proof. The proof of this theorem is the proof of Theorem 2.13 in [112]. The fact that the local solvers ($\tilde{S}_i^\dagger$ here) are not spd does not play a role in the proof. The idea is to prove the following bounds:

$$\hat{s}(u,u) \leq \hat{s}(P_{def}u,u) \leq \mathcal{N} \max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\hat{s}(u,u); \quad u \in \text{range}(I-P_0), \tag{5.41}$$

and

$$\hat{s}(u,u) \leq \hat{s}(P_{bal}u,u) \leq \max\left\{1, \mathcal{N} \max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\right\}\hat{s}(u,u); \quad u \in \hat{W}. \tag{5.42}$$

Following Lemma C.1 in the appendix of [112] these bounds imply the bounds for the condition numbers. They are proved using Lemma 5.14 and Lemma 5.22 combined with the fact that P_0 is an $\hat{s}$-orthogonal projection. □

Remark 5.24. The fact that $\mathcal{K}_i$ can be chosen such that $\left(\mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\right) < 1$ in (5.41) is not a contradiction: in this case the space range$(I-P_0)$ is simply empty.

5.4 Finite Element Tearing and Interconnecting

We use the following references to introduce FETI: the book by Toselli and Widlund [112], Tezaur's dissertation [111] and the article by Klawonn and Widlund [58]. A second level was introduced for FETI in [32], and further developed in [35, 33].

5.4.1 The FETI formulation

In the BDD section we built the coarse space for problem (5.24) which we simply recall here: find $\hat{u} \in \hat{W}$ such that $\hat{S}\hat{u} = \hat{f}_\Gamma$, where $\hat{W}$ is the space of functions defined on the interface Γ. Instead the FETI formulation of the problem is on the product space W with an additional matching constraint at the interfaces. This constraint is ensured using matrix

$$B = (B_1, B_2, \ldots, B_N); \quad Bu = \sum_{i=1,\ldots,N} B_i u_i, \; \forall\, u \in W, \tag{5.43}$$

which is constructed from entries $0, 1, -1$ such that the components u_i of a vector u in the product space W coincide on Γ when $Bu = 0$. More precisely each line in B corresponds to one continuity constraint for one degree of freedom and two of the subdomains to which it belongs: each line in B contains one 1 and one -1 while all other entries are zero. Denoting by λ the vector of Lagrange multipliers which is used to enforce the constraint $Bu = 0$ we obtain a saddle point formulation of the problem: find $(u, \lambda) \in W \times U$ such that

$$\begin{pmatrix} S & B^\top \\ B & 0 \end{pmatrix}\begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} f_\Gamma \\ 0 \end{pmatrix}. \tag{5.44}$$

We note that the solution λ of (5.44) is unique only up to an additive element of $\mathrm{Ker}(B^\top)$ however the solution u to our problem does not depend on the choice of λ so this is not an issue in practice. For the theoretical study we introduce the space

$$U := \mathrm{range}(B) = \mathrm{Ker}(B^\top)^\perp,$$

and will search for $\lambda \in U$. Given a basis for $\mathrm{Ker}(S)$ which consists of n_K vectors, an important role is played by the prolongation operator $\mathcal{R}_N^\top : \mathbb{R}^{n_K} \to W$ which columns are these basis functions. The transpose $\mathcal{R}_N$ is a restriction operator which maps an element in W to the coordinates of its l_2-orthogonal projection onto $\mathrm{Ker}(S)$ in the same basis. We have used the subscript N because $\mathrm{Ker}(S)$ is often referred to as the *Natural* coarse space for FETI. Going back to the system, the solution of the first equation in (5.44) can be written as

$$u = S^\dagger(f_\Gamma - B^\top\lambda) + \mathcal{R}_N^\top\alpha, \quad \text{for some} \quad \alpha \in \mathrm{range}(\mathcal{R}_N), \tag{5.45}$$

if the right-hand side associated to the operator S is such that

$$f_\Gamma - B^\top\lambda \perp \mathrm{Ker}(S) \Leftrightarrow \mathcal{R}_N(f_\Gamma - B^\top\lambda) = 0, \tag{5.46}$$

or with notation inspired by the usual FETI notation:

$$G_N^\top\lambda = \mathcal{R}_N f_\Gamma, \quad G_N := B\mathcal{R}_N^\top. \tag{5.47}$$

Injecting (5.45) into the second equation in (5.44) we get

$$BS^\dagger B^\top\lambda - G_N\alpha = BS^\dagger f_\Gamma, \quad \text{for some} \quad \alpha \in \mathrm{range}(\mathcal{R}_N).$$

We may again rewrite the problem using a saddle point formulation as

$$\begin{pmatrix} F & -G_N \\ G_N^\top & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \alpha \end{pmatrix} = \begin{pmatrix} d \\ e \end{pmatrix}, \tag{5.48}$$

where

$$F := BS^\dagger B^\top,\ d := BS^\dagger f_\Gamma,\ e := \mathcal{R}_N f_\Gamma, \text{ and again } G_N = B\mathcal{R}_N^\top. \tag{5.49}$$

In order to homogenize the second equation and bring the problem down to a single equation we decompose λ into $\lambda = \tilde{\lambda} + \lambda_N$ where $G_N^\top\tilde{\lambda} = 0$ and $G_N^\top\lambda_N = e$. Then we introduce a projection operator $\mathcal{P}_N$ as follows: let $Q : U \to U$ be a self-adjoint matrix which is positive definite on $\mathrm{range}(G_N)$, then define

$$\mathcal{P}_N : U \to U; \quad \mathcal{P}_N := I - QG_N(G_N^\top QG_N)^{-1}G_N^\top. \tag{5.50}$$

Remark 5.25. It is straightforward to prove that $\mathcal{P}_N$ is a projection operator from U onto $\mathrm{Ker}(G_N^\top)$ and that its transpose $\mathcal{P}_N^\top = I - G_N(G_N^\top QG_N)^{-1}G_N^\top Q$ is a Q-orthogonal projection. It is however less obvious to prove that the inverse $(G_N^\top QG_N)^{-1}$ is well defined. This can be derived from the fact that Q is positive definite on $\mathrm{range}(G_N)$ so $G_N^\top QG_N\beta = 0$ implies $G_N\beta = 0 \Leftrightarrow B\mathcal{R}_N^\top\beta = 0$. In other words $R_N^\top\beta \in \mathrm{Ker}(S) \cap \mathrm{Ker}(B)$ and this intersection is zero because the problem is well posed.[1] Finally $\beta = 0$ and $(G_N^\top QG_N)^{-1}$ is well defined.

1. In case the global operator $\hat{K}$ is singular, a solution exists for the original problem if $\hat{f}$ is in the range of $\hat{K}$. In that case the natural coarse space becomes singular but the FETI approach can still be applied [92].

The system which we solve is the projected system into the space

$$V_N := \mathrm{Ker}(G_N^\top) = \mathrm{range}(\mathcal{P}_N). \tag{5.51}$$

For the choice $\lambda_N := QG_N(G_N^\top QG_N)^{-1}\mathcal{R}_N f_\Gamma$ (which fulfills the condition $G_N^\top \lambda_N = e$) the problem is: find $\tilde{\lambda} \in V_N$ and $\alpha \in \mathrm{range}(\mathcal{R}_N)$ such that

$$F\tilde{\lambda} - G_N\alpha = d - F\lambda_N. \tag{5.52}$$

Testing this against elements in V_N yields the final form of the problem before preconditioning

$$\mathcal{P}_N^\top F\tilde{\lambda} = \mathcal{P}_N^\top(d - F\lambda_N), \tag{5.53}$$

whereas testing against function in $\mathrm{range}(I - \mathcal{P}_N)$ allows us to define the component α of the solution completely with respect to $\tilde{\lambda}$:

$$(I - \mathcal{P}_N^\top)G_N\alpha = (I - \mathcal{P}_N^\top)(F\tilde{\lambda} - d + F\lambda_N) \Leftrightarrow \alpha = (G_N^\top QG_N)^{-1}G_N^\top Q(F\lambda - d),$$

where we simply used a multiplication by $(G_N^\top QG_N)^{-1}G_N^\top Q$ to write the equivalence. Next we introduce the two usual FETI preconditioners.

5.4.2 Usual preconditioners for FETI

We first need to introduce diagonal scaling matrices $D_i : W_i \to W_i$ for each $i = 1, \ldots, N$. These are the matrix counterparts of the partition of unity operators Ξ_i used in the BDD section. Then let $D : W \to W$ be the diagonal scaling matrix $D := \begin{pmatrix} D_1 & 0 & \ldots & 0 \\ 0 & D_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & D_N \end{pmatrix}$, on the product space. We will consider two different preconditioners for (5.53): the Dirichlet preconditioner with the subscript D and the lumped preconditioner with the subscript L [36]. When scaled, those preconditioners can be written as the following operators on U [58]:

$$\mathcal{M}_D^{-1} = \left[D^{-1}B^\top(BD^{-1}B^\top)^\dagger\right]^\top S\left[D^{-1}B^\top(BD^{-1}B^\top)^\dagger\right] \tag{5.54}$$

$$\mathcal{M}_L^{-1} = \left[D^{-1}B^\top(BD^{-1}B^\top)^\dagger\right]^\top K^{bb}\left[D^{-1}B^\top(BD^{-1}B^\top)^\dagger\right]. \tag{5.55}$$

We use the subscript $*$ to refer to either of these preconditioners generically: if $*$ denotes D then $\mathcal{M}_*^{-1} = \mathcal{M}_D^{-1}$ is the Dirichlet preconditioner and if $*$ denotes L then $\mathcal{M}_*^{-1} = \mathcal{M}_L^{-1}$ is the Lumped preconditioner. When the diagonal scaling matrix D is chosen to be the diagonal of the local operator matrix K, the scaling in the preconditioners (5.54,5.55) are equivalent to so-called super-lumped scaling (or K-scaling) originally proposed in [93].

Remark 5.26. In (5.54,5.55) we have used a pseudo inverse where the usual FETI theory uses an inverse. This has no impact on what follows. Indeed, $(BD^{-1}B^\top)^\dagger$ is defined up to an additive element in $\mathrm{Ker}(BD^{-1}B^\top)$ and we have the inclusion $\mathrm{Ker}(BD^{-1}B^\top) \subset \mathrm{Ker}(B^\top)$ since

$$\lambda \in \mathrm{Ker}(BD^{-1}B^\top) \Rightarrow D^{-1}B^\top\lambda \in \mathrm{Ker}(B) \Rightarrow B^\top\lambda = Dv \,\text{for some}\, v \in \mathrm{Ker}(B),$$

and $\mathrm{Ker}(B) = (\mathrm{range}(B^\top))^\perp$ so $v^\top B^\top\lambda = v^\top Dv = 0 \Rightarrow v = 0 \Rightarrow \lambda \in \mathrm{Ker}(B^\top)$. The operator $(BD^{-1}B^\top)^\dagger$ is applied to elements in $\mathrm{range}(B) = \mathrm{Ker}(B^\top)^\perp$ so this application is well

defined. Moreover the application of $(BD^{-1}B^\top)^\dagger$ is followed by an application of $B^\top$ so $D^{-1}B^\top(BD^{-1}B^\top)^\dagger$ is uniquely defined independently of the choice of the pseudo inverse. This pseudo inverse can be avoided by defining scaling matrices directly on the space of Lagrange multipliers which is done for instance in the redundant Lagrange multiplier section of [58]. For sensible choices both approaches can lead to identical preconditioners and in practical implementations the scaling matrices are actually never computed explicitly as is explained in [93].

Using the subscript $*$ for either D or L, the preconditioned operator is $\mathcal{M}_*^{-1}\mathcal{P}_N^\top F$. Because we solve the system using a projected conjugate gradient method we require that the search directions remain in V_N. Therefore we actually solve: find $\lambda \in V_N$ such that

$$\mathcal{P}_N\mathcal{M}_*^{-1}\mathcal{P}_N^\top F\lambda = \mathcal{P}_N\mathcal{M}_*^{-1}\mathcal{P}_N^\top(d - F\lambda_N). \tag{5.56}$$

Because of the projection step (5.53) and the choice $\lambda_N := QG_N(G_N^\top QG_N)^{-1}\mathcal{R}_N f$ this is already a two-level preconditioner where the coarse space is $\mathrm{Ker}(\mathcal{P}_N) = \mathrm{range}(QG_N) = \mathrm{range}(QB\mathcal{R}_N^\top)$. The PPCG solver is initialized with λ_N and the entire solution space is $\lambda_N + V_N$. We will refer to $\mathcal{P}_N$ as the natural coarse space projector.

The theoretical study of the preconditioner is related to operator

$$P_D : W \to W; \quad P_D := D^{-1}B^\top(BD^{-1}B^\top)^\dagger B, \tag{5.57}$$

where $D : W \to W$ is the diagonal scaling matrix already introduced. This is a projection that is orthogonal in the scaled l_2 inner product $x^\top Dy$ $(x, y \in W)$. The next two lemmas follow essentially by noticing that $BP_Du = Bu$. They are Lemmas 4.1 and 4.3 in [58]. We give the proofs for sake of completeness because they are short.

Lemma 5.27. For any $\mu \in U$ there exists $\tilde{u} \in \mathrm{range}(P_D)$ such that $\mu = B\tilde{u}$.

Proof. By definition of U there exists $u \in W$ such that $\mu = Bu$. Now take $\tilde{u} = P_Du$, $B\tilde{u} = Bu = \mu$. □

Lemma 5.28. Let $u \in W$, then

$$P_Du = u - E_Du, \tag{5.58}$$

where $E_Du : W \to W$ is an averaging operator defined by its components as: $(E_Du)_i = R_i \sum_{j=1}^N R_j^\top D_j u_j$.

Proof. We start by noticing that $B(u - P_Du) = 0$. This means that $u - P_Du$ matches at the interfaces and thus its weighted average satisfies $E_D(u - P_Du) = u - P_Du$. A sufficient condition to ensure that the result holds is now $E_DP_Du = 0$.

By definition of E_D, E_DP_Du is a D-weighted average of the values of P_Du which correspond to the same global dof. One way to compute the averaged value for global dof k is to first compute $DP_Du = B^\top(BD^{-1}B^\top)^\dagger Bu$ and then sum the contributions from the different subdomains for which k is a degree of freedom. This is the same as computing an l_2 scalar product between $B^\top(BD^{-1}B^\top)^\dagger Bu$ and the function $e_x \in W$ which is zero everywhere except at the degrees of freedom which correspond to global dof k. By definition $Be_x = 0$. The orthogonality of $\mathrm{Ker}(B)$ and $\mathrm{range}(B^\top)$ allows us to conclude that $\langle Be_x, B^\top(BD^{-1}B^\top)^\dagger Bu\rangle = 0$ and thus $E_DP_Du = 0$. □

This last lemma allows us to prove that two suitable choices for Q in the projection operator $\mathcal{P}_N$ are $\mathcal{M}_D^{-1}$ and $\mathcal{M}_L^{-1}$.

Lemma 5.29. Both preconditioners $\mathcal{M}_D^{-1}$ and $\mathcal{M}_L^{-1}$ defined by (5.54) and (5.55) are self adjoint on U and positive definite on range(G_N). Consequently they are possible choices for matrix Q in the natural projection operator defined by (5.50).

Proof. We will only prove positive definiteness. Any $\lambda \in \text{range}(G_N)$ can be writteb as $\lambda = Bz$ for some $z \in \text{Ker}(S)$. Moreover, according to Lemma 5.6, $\lambda \in \text{Ker}(\mathcal{M}_L^{-1})$ implies $\lambda \in \text{Ker}(\mathcal{M}_D^{-1})$ so whether $*$ denotes D or L we get $\lambda = Bz \in \text{Ker}(\mathcal{M}_D^{-1})$. Using the definitions of $\mathcal{M}_D^{-1}$ and P_D as well as Lemma 5.28

$$0 = \langle \mathcal{M}_D^{-1} Bz, Bz\rangle = \langle SP_D z, P_D z\rangle = \langle S(z - E_D z), z - E_D z\rangle.$$

Now we have $z \in \text{Ker}(S)$ and $z - E_D z \in \text{Ker}(S)$ so necessarily $E_D z \in \text{Ker}(S)$. By definition $E_D z \in \text{Ker}(B)$ (it is the D-weighted average of z). The problem is well posed so $\text{Ker}(S) \cap \text{Ker}(B) = 0$. Finally $z = 0$ and $\mathcal{M}_*^{-1}$ is positive definite on range(G_N). □

We have just given two possible choices which complete the definition of the natural coarse space projector and thus the definitions of the spaces V_N and V_N'. The main result which we prove holds for these particular choices. For $*$ denoting either D or L, we introduce the notation:

$$\mathcal{P}_{*,N} := I - \mathcal{M}_*^{-1} G_N (G_N^\top \mathcal{M}_*^{-1} G_N)^{-1} G_N^\top \tag{5.59}$$

and

$$V_{*,N} = \text{range}(\mathcal{P}_{*,N}), \quad V'_{*,N} = \text{range}(\mathcal{P}_{*,N}^\top). \tag{5.60}$$

The next lemma states a crucial property for the preconditioners which is that they are positive definite.

Lemma 5.30. The preconditioners $\mathcal{P}_{*,N}\mathcal{M}_*^{-1} : V'_{*,N} \to V_{*,N}$ are symmetric positive definite for $*$ denoting either D or L.

Proof. Again, we only prove positive definiteness. Consider any $\mu \in V'_{*,N}$ with $\langle \mathcal{P}_{*,N}\mathcal{M}_*^{-1}\mu, \mu\rangle = \langle \mathcal{M}_*^{-1}\mu, \mu\rangle = 0$. By Lemma 5.27, $\mu = B\tilde{u}$ for some $\tilde{u} \in \text{range}(P_D)$. Operator P_D is a projection so $P_D\tilde{u} = \tilde{u}$, and we obtain

$$0 = \langle \mathcal{M}_*^{-1} B\tilde{u}, B\tilde{u}\rangle = \begin{cases} |D^{-1}B^\top (BD^{-1}B^\top)^\dagger B\tilde{u}|_S^2 & = |P_D\tilde{u}|_S^2 = |\tilde{u}|_S^2 \text{ if } * = D, \\ |D^{-1}B^\top (BD^{-1}B^\top)^\dagger B\tilde{u}|_{K^{bb}}^2 & = |P_D\tilde{u}|_{K^{bb}}^2 = |\tilde{u}|_{K^{bb}}^2 \text{ if } * = L. \end{cases}$$

According to Lemma 5.6, $|\tilde{u}|_{K^{bb}}^2 = 0$ implies $|\tilde{u}|_S^2 = 0$ so, whether $*$ denotes D or L, we get that $\tilde{u} \in \text{Ker}(S)$. By definition of $\mathcal{R}_N$, $\text{Ker}(S) = \text{range}(\mathcal{R}_N^\top)$ and in turn $\mathcal{M}_*^{-1} B\tilde{u} = \mathcal{M}_*^{-1}\mu \in \text{range}(\mathcal{M}_*^{-1} G_N)$.

The definition of $V'_{*,N}$ can be rewritten as

$$V'_{*,N} = \text{range}(\mathcal{P}_{*,N}^\top) = \text{Ker}(G_N^\top \mathcal{M}_*^{-1}) = \text{range}(\mathcal{M}_*^{-1} G_N)^\perp,$$

which together with $\mu \in V'_{*,N}$ and $\mathcal{M}_*^{-1}\mu \in \text{range}(\mathcal{M}_*^{-1} G_N)$ implies:

$$0 = \langle \mu, \mathcal{M}_*^{-1}\mu\rangle.$$

Finally, $\tilde{u} \in \text{range}(\mathcal{R}_N^\top)$ implies $\mu \in \text{range}(G_N)$ and $\mathcal{M}_*^{-1}$ is positive definite on range(G_N) so $\mu = 0$. □

5.4.3 Two level FETI preconditioner with the GenEO coarse space

The proof of an upper bound for the spectrum of the preconditioned FETI system usually relies on strong assumptions on the set of equations at hand and the coefficient distribution. Once again we build a coarse space which allows us to waive all of these assumptions. The coarse space is defined next along with the two-level FETI preconditioners (deflated and balanced). We use again the subscript 0 to refer to the coarse space. In order to avoid confusion with the BDD case we use calligraphic notation for the projection operator $\mathcal{P}_{*,0}$.

Definition 5.31 (GenEO coarse spaces for FETI)**.** Let $*$ denote either D (for Dirichlet) or L (for Lumped). For each subdomain $i = 1, \ldots, N$, find the eigenpairs $(q_i^k, \Lambda_i^k) \in W_i \times \mathbb{R}^+$ of the generalized eigenvalue problem:

$$\boxed{S_i\, q_i^k = \Lambda_i^k\, (B_i^\top \mathcal{M}_*^{-1} B_i)\, q_i^k.} \tag{5.61}$$

where $\mathcal{M}_*^{-1}$ is the preconditioner defined either by (5.54) or (5.55). Next, given a threshold $\mathcal{K}_i > 0$ for each subdomain, define the coarse space as

$$U_{*,0} = \text{span}(\{\mathcal{M}_*^{-1} B_i q_i^k\,;\, 0 < \Lambda_i^k < \mathcal{K}_i, i = 1, \ldots, N\}). \tag{5.62}$$

Let the interpolation operator $G_{*,0}$ be the matrix whose columns are the coarse basis functions $\{\mathcal{M}_*^{-1} B_i q_i^k; 0 < \Lambda_i^k < \mathcal{K}_i,\ i = 1, \ldots, N\}$. Let the coarse solver be the exact solver on $U_{*,0}$:

$$F_{*,0} := G_{*,0}^\top (\mathcal{P}_{*,N}^\top F \mathcal{P}_{*,N}) G_{*,0},$$

and let $\mathcal{P}_{*,0}$ be the $(\mathcal{P}_{*,N}^\top F \mathcal{P}_{*,N})$-orthogonal projection operator defined by

$$\mathcal{P}_{*,0} := I - G_{*,0} F_{*,0}^\dagger G_{*,0}^\top (\mathcal{P}_{*,N}^\top F \mathcal{P}_{*,N}). \tag{5.63}$$

Then the two-level preconditioners (respectively deflated and balanced) for F are

$$\mathcal{P}_{*,N} \mathcal{P}_{*,0} \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top \quad \text{and} \quad \mathcal{P}_{*,N} \mathcal{P}_{*,0} \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top + \mathcal{P}_{*,N} G_{*,0} F_{*,0}^\dagger G_{*,0}^\top \mathcal{P}_{*,N}^\top. \tag{5.64}$$

The operator $G_{*,0}$ is a mapping between the coordinates of a vector from $U_{*,0}$ in the set of coarse basis functions and its representation in U. Its transpose $G_{*,0}^\top$ is a restriction operator which maps an element in W to the coordinates of its l_2 projection onto $W_{*,0}$ in the set of coarse basis functions. The main difference with the coarse space for BDD is that we have left out the zero eigenvalues which correspond to the kernel of S because they are already taken care of by the natural coarse space through $\mathcal{P}_N$.

Remark 5.32. One common point with the BDD GenEO eigenvalue problem is that one of the operators (S_i) is a non assembled operator on the local space W_i whereas the other $(B_i^\top \mathcal{M}_*^{-1} B_i)$ is an assembled operator restricted to the local space W_i. This time the words assembled and restricted are to be understood in the FETI context and rely on the mappings B_i between the degrees of freedom in W_i and the Lagrange multipliers in U. In the same way as for BDD, the role of the GenEO eigenvalue problem for FETI can be interpreted as finding the modes necessary for describing the discrepancy between the interface behavior as seen from a single domain (left hand side of (5.61)), and the assembled interface operator F^{-1}, approximated by $\mathcal{M}_*^{-1}$ (right hand side of (5.61)). The idea is then to introduce those differences, which will not be well accounted for by the preconditioner, into the coarse space.

Once again in proving our estimate for the condition number we will take advantage of the orthogonality type properties which result from the generalized eigenvalue problem.

Lemma 5.33. Let $*$ denote either D or L. For a given subdomain $i = 1, \ldots, N$, the eigenpairs (q_i^k, Λ_i^k) of the generalized eigenproblem (5.61) can be chosen so that the set $\{q_i^k\}_k$ of eigenvectors is an orthonormal basis of W_i with respect to the inner product induced by $B_i^\top \mathcal{M}_*^{-1} B_i$. This can be written as

$$\langle \mathcal{M}_*^{-1} B_i q_i^k, B_i q_i^k \rangle = 1; \quad \text{and} \quad \langle \mathcal{M}_*^{-1} B_i q_i^k, B_i q_i^{k'} \rangle = 0,\ k \neq k'.$$

An orthogonality type property with respect to S_i (which is not necessarily coercive) also holds:

$$\langle S_i q_i^k, q_i^{k'} \rangle = 0,\ k \neq k'.$$

Proof. We proved in Lemma 5.29 that $\mathcal{M}_*^{-1}$ is spd on $\text{range}(G_N) = \text{Ker}(\mathcal{P}_N^\top)$. We also proved in Lemma 5.30 that $\mathcal{M}_*^{-1}$ is spd on $V_N' = \text{range}(\mathcal{P}_N^\top)$. So $\mathcal{M}_*^{-1}$ is spd on $\text{Ker}(\mathcal{P}_N^\top) \oplus \text{range}(\mathcal{P}_N^\top) = U$. Finally by definition of B_i, $B_i u_i = 0$ implies $u_i = 0$ so $B_i^\top \mathcal{M}_*^{-1} B_i$ is symmetric positive definite on W_i and the result is well known. □

In the next lemma we give some useful properties of the projections.

Lemma 5.34.

(i) $\text{range}(\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top) \subset \text{range}(\mathcal{P}_{*,N}^\top)$.

(ii) $\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top = \mathcal{P}_{*,N}^\top \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top$

(iii) $\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top$ and $\mathcal{P}_{*,N} \mathcal{P}_{*,0}$ are projections.

Proof. (i) By definition of $\mathcal{P}_{*,0}$ (5.63): $\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top = \mathcal{P}_{*,N}^\top (I - F \mathcal{P}_{*,N} G_0 F_0^\dagger G_0^\top)$.

(ii) It follows from (i) and the fact that $\mathcal{P}_{*,N}^\top$ is a projection that $\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top = \mathcal{P}_{*,N}^\top \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top$.

(iii) Then $\mathcal{P}_{*,0}^\top$ is also a projection so $\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top = \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top$.

□

For two spd matrices M_1 and M_2 of same size, the spectrum of $M_1 M_2$ is identical to the spectrum of $M_2 M_1$. Following this idea we decide to look at the problem in reverse: *Is F a good preconditioner for $\mathcal{M}_*^{-1}$?* The reason why we do this is that then we recognize an abstract Schwarz type preconditioner $F = \sum_{i=1}^N B_i S_i^\dagger B_i^\top$. In this framework, the local subspaces are the W_i and the local solvers are the pseudo inverses $S_i^\dagger$ of the local bilinear forms S_i. The prolongation operators are the $B_i : W_i \to U$ and the restriction operators are the $B_i^\top : U \to W_i$. Taking advantage of the abstract Schwarz framework, in Lemmas 5.35 and 5.37 we will prove the same estimates as in the BDD subsection for F viewed as the preconditioner and $\mathcal{M}_*^{-1}$ viewed as the matrix problem. In the proof of our final theorem it will become apparent that the se estimates makes it possible to prove the condition number of FETI with the two-level preconditioners given by (5.64). In the next Lemma, applying the exact same strategy as in Lemma 5.14 we give an estimate related to a lower bound for the eigenvalues of the preconditioned operator $F \mathcal{P}_{*,N} \mathcal{P}_{*,0} \mathcal{M}_*^{-1}$. This bound does not depend on the choice of the coarse space.

Lemma 5.35 (Stable splitting – Lower bound for the eigenvalues of the preconditioned operator)**.** For any $\mu \in V_{*,N}'$ there exists a stable splitting $(v_1, \ldots, v_N) \in W_1 \times \cdots \times W_N$ of μ :

$$\mu = B_1 v_1 + \ldots B_N v_N;\ v_i \in W_i \text{ and } \sum_{i=1}^N \langle S_i v_i, v_i \rangle \leq \langle \mathcal{M}_*^{-1} \mu, \mu \rangle. \tag{5.65}$$

This implies

$$\langle \mathcal{M}_*^{-1}\mu,\mu\rangle \leq \langle F\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu,\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle \text{ for any } \mu\in \operatorname{range}(\mathcal{P}_{*,0}^\top\mathcal{P}_{*,N}^\top).$$

Proof. Let $\mu \in V'_{*,N}$ and let $v_i = D_i^{-1}B_i^\top(BD^{-1}B^\top)^\dagger\mu$ for each $i = 1,\dots,N$. This provides a splitting of μ:

$$\sum_{i=1}^N B_i v_i = \sum_{i=1}^N B_i D_i^{-1}B_i^\top(BD^{-1}B^\top)^\dagger\mu = (BD^{-1}B^\top)(BD^{-1}B^\top)^\dagger\mu = \mu,$$

since $\mu\in\operatorname{range}(BD^{-1}B^\top)=\operatorname{range}(B)=U$. Moreover, the splitting is stable:

$$\begin{aligned}\sum_{i=1}^N\langle S_i v_i,v_i\rangle &= \sum_{i=1}^N\langle S_iD_i^{-1}B_i^\top(BD^{-1}B^\top)^\dagger\mu, D_i^{-1}B_i^\top(BD^{-1}B^\top)^\dagger\mu\rangle\\ &= \langle SD^{-1}B^\top(BD^{-1}B^\top)^\dagger\mu, D^{-1}B^\top(BD^{-1}B^\top)^\dagger\mu\rangle\\ &= \langle \mathcal{M}_D^{-1}\mu,\mu\rangle,\\ &\leq \langle \mathcal{M}_*^{-1}\mu,\mu\rangle,\end{aligned}$$

by Lemma 5.6. This is exactly (5.65). Now let $\mu\in\operatorname{range}(\mathcal{P}_{*,0}^\top\mathcal{P}_{*,N}^\top)$, then $\langle\mathcal{M}_*^{-1}\mu,\mu\rangle = \langle\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu,\mu\rangle$. Moreover, the fact that the v_i provide a splitting implies

$$\begin{aligned}\langle\mathcal{M}_*^{-1}\mu,\mu\rangle &= \langle\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu,\sum_{i=1}^N B_iv_i\rangle\\ &= \sum_{i=1}^N\langle\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, B_i(S_i^\dagger S_i)v_i\rangle\\ &= \sum_{i=1}^N\langle S_iv_i, S_i^\dagger B_i^\top\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle.\end{aligned}$$

Then we apply the Cauchy Schwarz inequality twice, first in the S_i inner product and then in the l_2 inner product and finish by using (5.65)

$$\begin{aligned}\langle\mathcal{M}_*^{-1}\mu,\mu\rangle &\leq \sum_{i=1}^N\left[\langle S_iv_i,v_i\rangle^{1/2}\langle S_iS_i^\dagger B_i^\top\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, S_i^\dagger B_i^\top\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle^{1/2}\right]\\ &\leq \left[\sum_{i=1}^N\langle S_iv_i,v_i\rangle\right]^{1/2}\left[\sum_{i=1}^N\langle S_iS_i^\dagger B_i^\top\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, S_i^\dagger B_i^\top\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle\right]^{1/2}\\ &\leq \langle\mathcal{M}_*^{-1}\mu,\mu\rangle^{1/2}\langle\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu,\sum_{i=1}^N B_iS_i^\dagger B_i^\top\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle^{1/2}.\end{aligned}$$

The result follows by raising to the square, simplifying by $\langle\mathcal{M}_*^{-1}\mu,\mu\rangle$ and recognizing $F=\sum_{i=1}^N B_iS_i^\dagger B_i^\top$. □

The next lemma is the FETI counterpart of lemma 5.20 and the proof follows the exact same steps. We prove a crucial result which relies very strongly on the choice of the coarse space. In fact the coarse space was chosen specifically to ensure that this estimate holds.

Lemma 5.36 (Stability of the local solvers)**.** Let $*$ denote either D or L. For each $i = 1, \dots, N$, let the pseudo inverse $S_i^\dagger$ be chosen such that $\text{range}(S_i^\dagger) = \text{span}\{q_i^k; \Lambda_i^k > 0\}$. Then the following estimate for the local solver holds

$$\langle \mathcal{M}_*^{-1} B_i u_i, B_i u_i \rangle \leq \frac{1}{\mathcal{K}_i} \langle S_i u_i, u_i \rangle, \quad \forall u_i \in \text{range}(S_i^\dagger B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top), \tag{5.66}$$

where the $\mathcal{K}_i$ are the thresholds that were used to select eigenvectors for the coarse space in Definition 5.31.

Proof. First we prove that $\text{range}(S_i^\dagger B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top) \subset \text{span}\{q_i^k; \Lambda_i^k \geq \mathcal{K}_i\}$. We will use the following linear algebra identity

$$\text{Ker}(\mathcal{P}_{*,N} \mathcal{P}_{*,0} \mathcal{M}_*^{-1} B_i) \oplus^\perp \text{range}(B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top) = W_i, \tag{5.67}$$

where the symbol $\perp$ refers to the l_2 orthogonality between both spaces and $\oplus$ means that the sum is direct. According to item (ii) in Lemma 5.34, $\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top = \mathcal{P}_{*,N}^\top \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top$. This implies $\mathcal{P}_{*,N} \mathcal{P}_{*,0} = \mathcal{P}_{*,N} \mathcal{P}_{*,0} \mathcal{P}_{*,N}$. So $\text{Ker}(\mathcal{P}_{*,N}) \subset \text{Ker}(\mathcal{P}_{*,N} \mathcal{P}_{*,0})$. It is also obvious that $\text{Ker}(\mathcal{P}_{*,0}) \subset \text{Ker}(\mathcal{P}_{*,N} \mathcal{P}_{*,0})$. Using the definitions of these projections ((5.59) and (5.63)) this can be rewritten as

$$\text{Ker}(\mathcal{P}_{*,N} \mathcal{P}_{*,0}) \supset (\text{Ker}(\mathcal{P}_{*,N}) \cup \text{Ker}(\mathcal{P}_{*,0})) \supset \left(\text{range}(G_{*,0}) \cup \text{range}(\mathcal{M}_*^{-1} G_N) \right).$$

By definition of $G_{*,0}$ and G_N, in particular, for each $i = 1, \dots, N$,

$$\text{span}\{\mathcal{M}_*^{-1} B_i q_i^k; \Lambda_i^k < \mathcal{K}_i\} \subset \text{Ker}(\mathcal{P}_{*,N} \mathcal{P}_{*,0}),$$

so

$$\text{span}\{q_i^k; \Lambda_i^k < \mathcal{K}_i\} \subset \text{Ker}(\mathcal{P}_{*,N} \mathcal{P}_{*,0} \mathcal{M}_*^{-1} B_i). \tag{5.68}$$

Following the same procedure as to prove (5.37) in Lemma 5.20, the first orthogonality property in Lemma 5.33 implies that

$$\text{span}\{q_i^k; \Lambda_i^k < \mathcal{K}_i\} \oplus^\perp \text{span}\{B_i^\top \mathcal{M}_*^{-1} B_i q_i^k; \Lambda_i^k \geq \mathcal{K}_i\} = W_i. \tag{5.69}$$

Putting (5.67), (5.68) and (5.69) together tells us that

$$\text{range}(B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top) \subset \text{span}\{B_i^\top \mathcal{M}_*^{-1} B_i q_i^k; \Lambda_i^k \geq \mathcal{K}_i\}.$$

Next the definition of eigenproblem (5.61), $S_i \, q_i^k = \Lambda_i^k \, (B_i^\top \mathcal{M}_*^{-1} B_i) \, q_i^k$, yields

$$\text{range}(B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top) \subset \text{span}\{S_i q_i^k; \Lambda_i^k \geq \mathcal{K}_i\}.$$

Finally for the specific choice of the pseudo inverse $S_i^\dagger$ it is obvious that

$$\text{range}(S_i^\dagger B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top) \subset \text{span}\{q_i^k; \Lambda_i^k \geq \mathcal{K}_i\}.$$

Now it is easy to prove (5.66) using the orthogonality type properties in Lemma 5.33 and the definition of the eigenproblem. Any $u_i \in \text{range}(S_i^\dagger B_i^\top \mathcal{M}_*^{-1} \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top)$ can be written as $u_i = \sum_{\{k; \Lambda_i^k \geq K_i\}} \alpha_i^k q_i^k$ for some coefficients $\alpha_i^k \in \mathbb{R}$, so:

$$\begin{aligned}
\langle \mathcal{M}_*^{-1} B_i u_i, B_i u_i \rangle &= \sum_{\{k; \Lambda_i^k \geq K_i\}} {\alpha_i^k}^2 \langle \mathcal{M}_*^{-1} B_i q_i^k, B_i q_i^k \rangle \\
&= \sum_{\{k; \Lambda_i^k \geq K_i\}} \frac{1}{\Lambda_i^k} {\alpha_i^k}^2 \langle S_i q_i^k, q_i^k \rangle \\
&\leq \frac{1}{\mathcal{K}_i} \sum_{\{k; \Lambda_i^k \geq K_i\}} {\alpha_i^k}^2 \langle S_i q_i^k, q_i^k \rangle \\
&= \frac{1}{\mathcal{K}_i} \langle S_i u_i, u_i \rangle
\end{aligned}$$

□

The next lemma is a direct consequence. It is the FETI counterpart of Lemma 5.22 and gives an estimate related to an upper bound for the eigenvalues of the preconditioned operator. The relationship will become apparent in the proof of the final theorem.

Lemma 5.37 (Upper bound for the eigenvalues of the preconditioned operator)**.** The following estimate holds

$$\langle F\mathcal{M}_*^{-1}\lambda, \mathcal{M}_*^{-1}\lambda\rangle \leq \mathcal{N} \max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\langle \mathcal{M}_*^{-1}\lambda,\lambda\rangle \text{ for any } \lambda \in \text{range}(\mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top), \tag{5.70}$$

where $\mathcal{N}$ is, as for BDD [2], the maximal number of neighbours of a subdomain (including itself) in the sense

$$\mathcal{N} = \max_{1\leq i\leq N}\left(\#\{j; R_j R_i^\top \neq 0\}\right).$$

Proof. In order to simplify notation lets write $\tilde{\mathcal{P}}_{*,i} := S_i^\dagger B_i^\top \mathcal{M}_*^{-1}$ and $\mathcal{P}_{*,i} := B_i\tilde{\mathcal{P}}_{*,i}$. Let $\lambda \in \text{range}(P_{*,0}^\top \mathcal{P}_{*,N}^\top)$, then

$$\begin{aligned}
\langle \mathcal{M}_*^{-1}\mathcal{P}_{*,i}\lambda, \mathcal{P}_{*,i}\lambda\rangle &= \langle \mathcal{M}_*^{-1}B_i\tilde{\mathcal{P}}_{*,i}\lambda, B_i\tilde{\mathcal{P}}_{*,i}\lambda\rangle \\
&\leq \frac{1}{\mathcal{K}_i}\langle S_i\tilde{\mathcal{P}}_{*,i}\lambda, \tilde{\mathcal{P}}_{*,i}\lambda\rangle \quad \text{(Lemma 5.36)} \\
&= \frac{1}{\mathcal{K}_i}\langle \mathcal{M}_*^{-1}\lambda, B_i\tilde{\mathcal{P}}_{*,i}\lambda\rangle \quad \text{(definition of } \tilde{\mathcal{P}}_{*,i}) \\
&= \frac{1}{\mathcal{K}_i}\langle \mathcal{M}_*^{-1}\lambda, \mathcal{P}_{*,i}\lambda\rangle
\end{aligned} \tag{5.71}$$

Taking a close look at the definition of the preconditioners in (5.54) and (5.55) we notice that they can be written as a sum of local contributions:

$$\mathcal{M}_*^{-1} = \sum_{j=1}^N \mathcal{M}_{*,j}^{-1}; \qquad \mathcal{M}_{*,j}^{-1} := \left[D_j^{-1}B_j^\top (BD^{-1}B^\top)^\dagger\right]^\top S_j \left[D_j^{-1}B_j^\top (BD^{-1}B^\top)^\dagger\right],$$

and if $\langle M_{*,j}^{-1}B_i u_i, B_i u_i\rangle \neq 0$ then $R_j R_i^\top \neq 0$. [3] A consequence of this is that

$$\langle \mathcal{M}_*^{-1}\lambda, \mathcal{P}_{*,i}\lambda\rangle = \langle \mathcal{M}_*^{-1}\lambda, B_i\tilde{\mathcal{P}}_{*,i}\lambda\rangle = \sum_{\{j; R_j R_i^\top \neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\lambda, B_i\tilde{\mathcal{P}}_{*,i}\lambda\rangle.$$

We apply the Cauchy Schwarz inequality for $\mathcal{M}_{*,j}^{-1}$ and then for the Euclidean inner product

2. in the article the definition of a neighbour was slightly different: $\mathcal{N} = \max_{1\leq i\leq N}\left(\#\{j; B_j^\top B_i \neq 0\}\right)$. It was brought to our attention that in the case of non redundant Lagrange multipliers we had comitted a mistake since it is possible to have $B_j^\top B_i = 0$ at the same time as $R_j R_i^\top \neq 0$ (see next footmark for the line in the proof where this appears).

3. We have changed this line compared to the article (in agreement with the previous footnote).

to this and inject the previous result

$$\begin{aligned}
\langle \mathcal{M}_*^{-1}\lambda, \mathcal{P}_{*,i}\lambda\rangle &\leq \sum_{\{j;R_jR_i^\top\neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\lambda,\lambda\rangle^{1/2}\langle \mathcal{M}_{*,j}^{-1}\mathcal{P}_{*,i}\lambda, \mathcal{P}_{*,i}\lambda\rangle^{1/2} \\
&\leq \left[\sum_{\{j;R_jR_i^\top\neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\lambda,\lambda\rangle\right]^{1/2}\left[\sum_{\{j;R_jR_i^\top\neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\mathcal{P}_{*,i}\lambda, \mathcal{P}_{*,i}\lambda\rangle\right]^{1/2} \\
&= \left[\sum_{\{j;R_jR_i^\top\neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\lambda,\lambda\rangle\right]^{1/2} \langle \mathcal{M}_*^{-1}\mathcal{P}_{*,i}\lambda, \mathcal{P}_{*,i}\lambda\rangle^{1/2} \\
&\leq \left[\sum_{\{j;R_jR_i^\top\neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\lambda,\lambda\rangle\right]^{1/2} \left[\frac{1}{\mathcal{K}_i}\langle \mathcal{M}_*^{-1}\lambda, \mathcal{P}_{*,i}\lambda\rangle\right]^{1/2} \qquad \text{(from (5.71))}.
\end{aligned}$$

Raising to the square and simplifying by $\langle \mathcal{M}_*^{-1}\lambda, \mathcal{P}_{*,i}\lambda\rangle$ yields

$$\langle \mathcal{M}_*^{-1}\lambda, \mathcal{P}_{*,i}\lambda\rangle \leq \frac{1}{\mathcal{K}_i}\sum_{\{j;R_jR_i^\top\neq 0\}} \langle \mathcal{M}_{*,j}^{-1}\lambda,\lambda\rangle.$$

Finally summing these inequalities over i and noticing that $\sum_{i=1}^N \mathcal{P}_{*,i} = F\mathcal{M}_*^{-1}$ ends the proof. □

We are now ready to prove the main theorem for the GenEO FETI algorithm which is similar to Theorem 5.23.

Theorem 5.38 (Main theorem for FETI with the GenEO coarse space)**.** Let $*$ denote either L for Lumped or D for Dirichlet. The condition number for FETI solved in range$(\mathcal{P}_{*,N}\mathcal{P}_{*,0})$ with the deflated operator satisfies

$$\kappa\left(\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^\top\mathcal{P}_{*,N}^\top F\right) \leq \mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right). \tag{5.72}$$

As for the balanced two-level preconditioner with the GenEO coarse space in range$(\mathcal{P}_{*,N})$, it satisfies

$$\kappa\left(\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^\top\mathcal{P}_{*,N}^\top F + \mathcal{P}_{*,N}G_{*,0}F_{*,0}^\dagger G_{*,0}^\top\mathcal{P}_{*,N}^\top F\right) \leq \max\left\{1, \mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\right\}. \tag{5.73}$$

These bounds depend only on the chosen thresholds $\mathcal{K}_i$ we use to select eigenvectors for the coarse space in Definition 5.31 and on the maximal number $\mathcal{N}$ of neighbours of a subdomain (including itself):

$$\mathcal{N} = \max_{1\leq i\leq N}\left(\#\{j; R_jR_i^\top \neq 0\}\right).$$

Proof. From Lemma C.1 in the appendix of [112], in order to prove (5.72), it is sufficient to show that, for any $\lambda \in$ range$(\mathcal{P}_{*,N}\mathcal{P}_{*,0})$, the following holds:

$$\langle(\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^\top\mathcal{P}_{*,N}^\top)^{-1}\lambda,\lambda\rangle \leq \langle F\lambda,\lambda\rangle \leq \mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\langle(\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^\top\mathcal{P}_{*,N}^\top)^{-1}\lambda,\lambda\rangle. \tag{5.74}$$

Lemma 5.30 tells us that the inverse is well defined. First of all note that the fact that $\mathcal{K}_i$ can be chosen such that $\left(\mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\right) < 1$ in (5.74) is not a contradiction: in this case the space range$(\mathcal{P}_{*,N}\mathcal{P}_{*,0})$ is simply empty. Next we prove (5.74): let $\mu \in$ range$(\mathcal{P}_{*,0}^{\top}\mathcal{P}_{*,N}^{\top})$, Lemma 5.35 tells us that

$$\langle \mathcal{M}_*^{-1}\mu, \mu\rangle \leq \langle F\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, \mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle.$$

Then, using the fact that $\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^{\top}\mathcal{P}_{*,N}^{\top}\mu = \mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu$, this is equivalent to

$$\langle (\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^{\top}\mathcal{P}_{*,N}^{\top})^{-1}\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, \mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle \leq \langle F\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, \mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle.$$

In turn, range$(\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^{\top}\mathcal{P}_{*,N}^{\top}) =$ range$(\mathcal{P}_{*,N}\mathcal{P}_{*,0})$ implies

$$\langle (\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mathcal{P}_{*,0}^{\top}\mathcal{P}_{*,N}^{\top})^{-1}\lambda, \lambda\rangle \leq \langle F\lambda, \lambda\rangle, \quad \forall\, \lambda \in \text{range}(\mathcal{P}_{*,N}\mathcal{P}_{*,0}),$$

which is the lower bound in (5.74).

For the upper bound we use the result from Lemma 5.37 which is that

$$\langle F\mathcal{M}_*^{-1}\mu, \mathcal{M}_*^{-1}\mu\rangle \leq \mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\langle \mathcal{M}_*^{-1}\mu, \mu\rangle, \quad \forall\, \mu \in \text{range}(\mathcal{P}_{*,0}^{\top}\mathcal{P}_{*,N}^{\top}).$$

We know that $\mathcal{M}_*^{-1}\mu = \mathcal{P}_{*,N}\mathcal{M}_*^{-1}\mu$ and projection $\mathcal{P}_{*,0}$ is $(\mathcal{P}_{*,N}^{\top}F\mathcal{P}_{*,N})$-orthogonal so

$$\langle F\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, \mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle \leq \langle F\mathcal{M}_*^{-1}\mu, \mathcal{M}_*^{-1}\mu\rangle,$$

and in turn

$$\langle F\mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu, \mathcal{P}_{*,N}\mathcal{P}_{*,0}\mathcal{M}_*^{-1}\mu\rangle \leq \mathcal{N}\max_{1\leq i\leq N}\left(\frac{1}{\mathcal{K}_i}\right)\langle \mathcal{M}_*^{-1}\mu, \mu\rangle.$$

In the same way as for the lower bound we may then show the upper bound in (5.74). This ends the proof for the condition number of the deflated preconditioned operator (5.72). The proof for the balanced operator (5.73) is similar to the BDD case, it relies simply on the fact that the projection operator $\mathcal{P}_{*,0}$ is $(\mathcal{P}_{*,N}^{\top}F\mathcal{P}_{*,N})$-orthogonal. □

5.5 Numerical results for two dimensional elasticity (FETI)

We give here a few numerical results to confirm the estimate for the condition number in the FETI case. The system of equations which we solve is related to two dimensional linear elasticity where the domain is clamped on the left hand side and subject to gravity. An important feature of the methods which we presented is that, given a FETI code, they do not demand a lot of implementation work: all the mathematical objects which are used to build the coarse space already appear in the algorithms.

All the results that follow were obtained using Freefem++ [50] to build the problem matrices and visualize solutions and Matlab for the solving procedure. The test problems we present here are only small tests which we use to validate our theoretical results. Of course, a full validation of the efficiency of the method would require larger scale tests with an optimized code. Full reorthogonalization at each iteration is used in PPCG. The meshes are regular with quadrilateral elements and the finite element discretization of the two dimensional elasticity equation uses standard $\mathbb{P}_1$ (linear) functions. There are two parameters in the linear elasticity system of equations: Young's modulus E and Poisson's

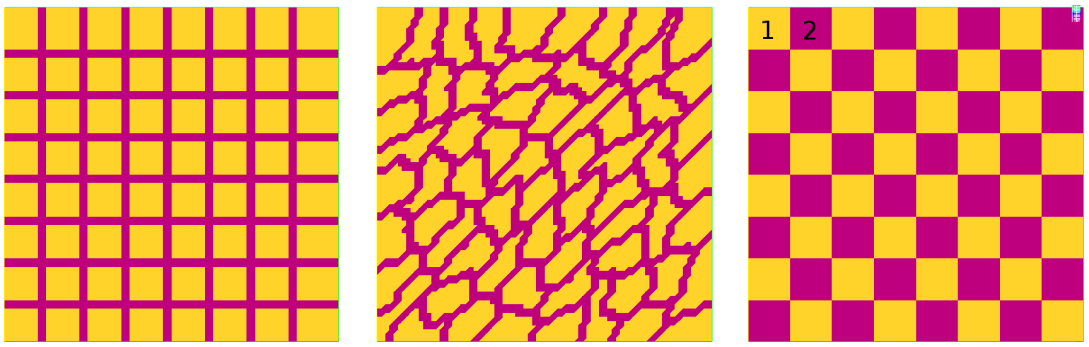


Figure 5.1: Decomposition of the unit square into 64 regular subdomains (left) – Decomposition of the unit square into 64 subdomains using Metis (middle) – Checkerboard coefficient distribution (right)

ratio ν. Each time an iteration count is given, the stopping criterion is that the relative primal residual at the final iteration k reach 10^{-4}:

$$\frac{\| \sum_{i=1}^{N} R_i^\top S_i D_i^{-1} B_i^\top (B D^{-1} B^\top)^\dagger \mathcal{P}_{*,0}^\top \mathcal{P}_{*,N}^\top (d - F\lambda_k)\|_2}{\|\hat{f}_\Gamma\|_2} < 10^{-4}.$$

The fact that this is indeed the primal residual is explained in [90] and proved for instance in [69].

5.5.1 Checkerboard coefficient distribution

We discretize a square of size 1×1 using 81×81 nodes. We use two different decompositions of this unit square: a regular decomposition into 8×8 regular subdomains (Figure 5.1 – left) and a decomposition into 64 subdomains obtained using Metis [54] (Figure 5.1 – middle). Throughout this subsection, the scaling matrices are chosen to be the K-scaling matrices [93, 58], meaning that in the definitions of the preconditioners (5.54) and (5.55) we set

$$D_i = diag(K_i). \tag{5.75}$$

The criterion for selecting which modes are used to build the coarse space is set to

$$\mathcal{K}_i = 0.1; \quad \forall\, i = 1, \ldots, N,$$

so the condition number should satisfy $\kappa \leq 10 \times \mathcal{N}$ where $\mathcal{N}$ is the maximal number of neighbours.

The partition resolves the heterogeneities

It is well known by now that in the case of a regular decomposition into subdomains which resolves the jumps in the coefficients and the Dirichlet preconditioner, the use of the K-scaling matrices (5.75) is sufficient to ensure good convergence. We check here that in these cases the (automatic) GenEO strategy is to do nothing special which is to say that no extra modes are selected to build the additional coarse space U_0. Table 5.2 gives the results for the regular partition (Figure 5.1 – left) into subdomains and a constant coefficient distribution $(E; \nu) = (10^7; 0.4)$ as well as a *checkerboard* coefficient distribution (Figure

	Dirichlet			Lumped				
				FETI-GenEO			FETI-1	
Coefficients	κ	$\#U_0$	it	κ	$\#U_0$	it	κ	it
Constant	9.5	0	15	11.1	15	17	86	24
Checkerboard	6.3	0	13	9.7	49	19	93	25

Table 5.2: Checkerboard (64 regular subdomains) κ : condition number; $\#U_0$: size of the GenEO coarse space; it: number of iterations – For the Dirichlet preconditioner the GenEO coarse space is empty so FETI-GenEO and FETI-1 are identical

5.1 – right) where the coefficients take the values $(E_1;\nu_1) = (10^7;0.4)$ and $(E_2;\nu_2) = (10^{12};0.3)$. We have solved each of these problems with the Dirichlet preconditioner and the Lumped preconditioner with and without the GenEO coarse space (we re fer to these cases as FETI-GenEO and FETI-1 respectively). For each test we give the condition number κ of the preconditioned operator, the size of the GenEO coarse space $\#U_0$ (if there is one) and the number it of iterations needed to reach convergence. The first thing that we notice is that in all four cases where the GenEO coarse space is used the estimate for the condition number is satisfied. In the Dirichlet preconditioner case, no modes where selected to build the coarse space which is what we expected since the K-scaling alone is known to be efficient. With the Lumped preconditioner case only few modes were selected (less than one per subdomain). This test indicates that the GenEO coarse space circumvents the fact that the lumped preconditioner does not properly predict the corrections needed on the interface for checkerboard problems.

The partition does not resolve the heterogeneities

This time we use the automatic partition into 64 subdomains obtained using METIS [54] (Figure 5.1 – middle). The coefficient distribution is still the checkerboard distribution shown on the right hand side of Figure 5.1 so the subdomain interfaces do not coincide with the jumps in the coefficients. The coefficients are a fixed $(E_1;\nu_1) = (10^7;0.4)$ and a variable (E_2,ν_2) one. Table 5.3 gives the results for different values of (E_2,ν_2). The middle line shows a case where the coefficients are constant throughout the subdomain ($(E_2,\nu_2) = (E_1;\nu_1)$). Once again we observe that in all cases the condition number satisfies the estimate and that it hardly varies with the jumps in the coefficients. In the worse case the number of modes used to build the coarse space is 370 (less than 6 modes per subdomain on average). Because of bad numerical conditioning there are a few cases where the FETI-1 residual never reaches 10^{-4}, instead it stagnates. In this case we report the iteration count before the *plateau* and the corresponding residual. Figure 5.2 shows a comparison between the convergence curves with and without the additional GenEO coarse space where this phenomenon can be observed. Figure 5.3 shows the spectrum of the preconditioned operators with and without the additional coarse space. The spectrum is represented in the complex plane but the imaginary part is always almost zero (imaginary parts result from numerical errors in the eigensolver). The zeros in the spectrum correspond to the coarse modes (either natural or GenEO) as well as the null space of $B^\top$. Whether the GenEO coarse space is used or not, the first non zero eigenvalue of the preconditioned operator is 1 which is what is expected.

	Dirichlet Preconditioner					Lumped Preconditioner				
	FETI-GenEO			FETI-1		FETI-GenEO			FETI-1	
$(E_2;\nu_2)$	κ	$\#U_0$	it	κ	it	κ	$\#U_0$	it	κ	it
$(10^{12};0.3)$	10.4	126	18	$1.5 \cdot 10^6$	$142^{(1)}$	11.7	186	19	$6.2 \cdot 10^6$	$154^{(2)}$
$(10^7;0.4)$	10.5	26	18	447	31	12.2	99	23	$2.1 \cdot 10^3$	58
$(10^2;0.49)$	12.2	182	21	$5.3 \cdot 10^6$	$170^{(3)}$	16.3	370	23	$4.0 \cdot 10^7$	$198^{(4)}$

(1) the relative residual reaches a plateau at $2 \cdot 10^{-4}$ after 142 iterations.
(2) the relative residual reaches a plateau at $3 \cdot 10^{-4}$ after 154 iterations.
(3) the relative residual reaches a plateau at $2 \cdot 10^{-3}$ after170 iterations.
(4) the relative residual reaches a plateau at $1 \cdot 10^{-3}$ after 198 iterations.

Table 5.3: Checkerboard (64 Metis subdomains) $(E_1;\nu_1) = (10^7;0.4)$; κ : condition number; $\#U_0$: size of the GenEO coarse space; it: number of iterations. When $(E_2;\nu_2) = (10^7;0.4)$ there are no jumps in the coefficients.

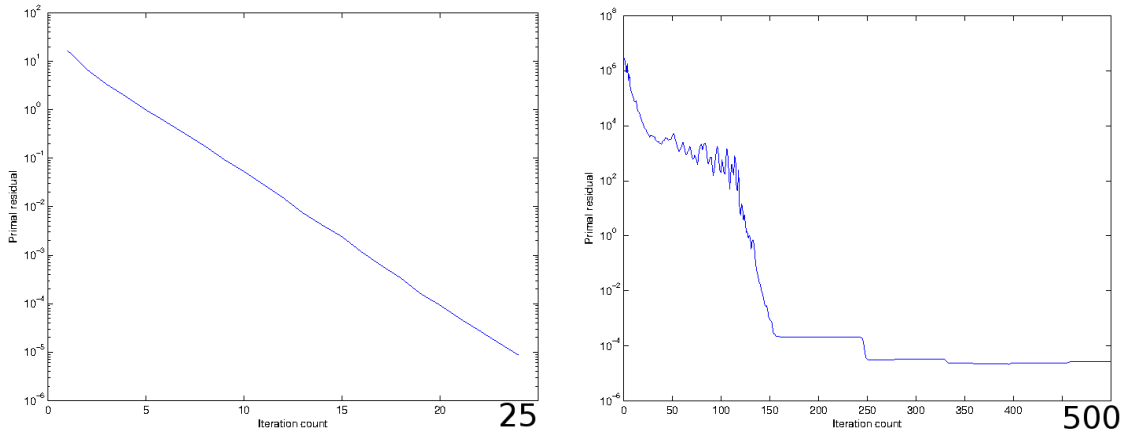


Figure 5.2: Checkerboard coefficient distribution – Convergence curve: primal residual versus iteration count – Left: with GenEO, Right : without GenEO – Lumped preconditioner for the Metis decomposition into 64 subdomains – $(E_1;\nu_1) = (10^7;0.4)$ and $(E_2;\nu_2) = (10^{12};0.3)$.

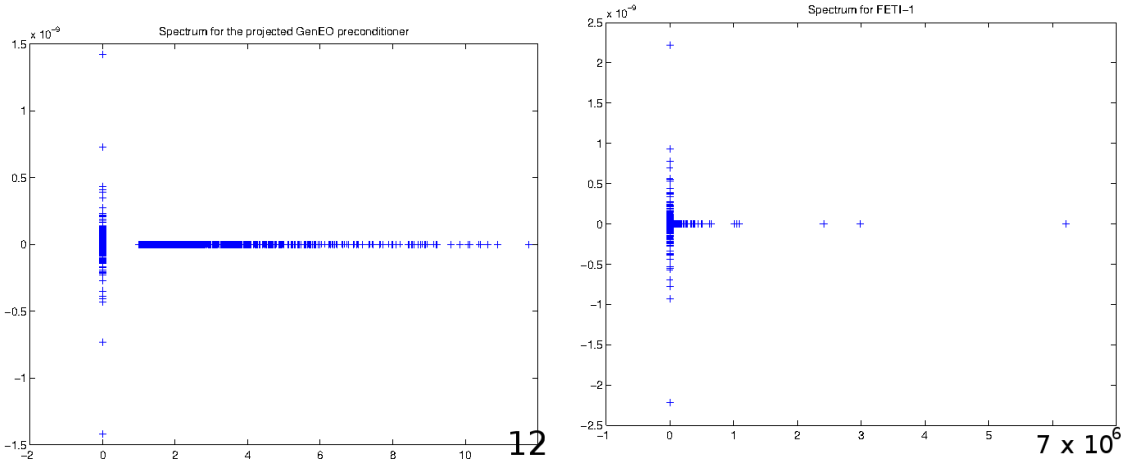


Figure 5.3: Checkerboard coefficient distribution – Spectrum of the preconditioned operator – Left: with GenEO, Right : without GenEO – Lumped preconditioner for the Metis decomposition into 64 subdomains – $(E_1;\nu_1) = (10^7;0.4)$ and $(E_2;\nu_2) = (10^{12};0.3)$.

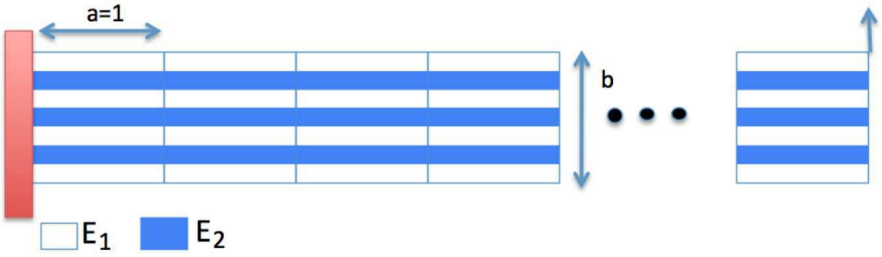


Figure 5.4: Discontinuities along the interfaces

5.5.2 Discontinuities along the interfaces

In this subsection we focus only on the GenEO coarse space for the Dirichlet preconditioner and we conduct a more extensive study. We use a partition into N regular subdomains of a rectangle of size $N \times b$ where b is the aspect ratio of each subdomain (see Figure 5.4). The discretization of each subdomain is $n_{el} \times n_{el}$ rectangular elements so that each element has the same aspect ratio as the subdomain to which it belongs. The coefficient distribution consists of a constant value $\nu = 0.3$ of Poisson's ratio and 7 layers of E (4 soft layers, 3 hard layers, see again Figure 5.4). Throughout this subsection we use again the K-scaling matrices (5.75) which is in fact, for this case, equivalent to choosing multiplicity scaling since the coefficient jumps are only along the interfaces.

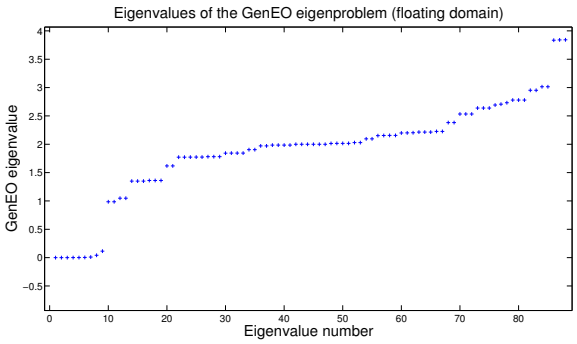
The parameters are: $b = 1$ (aspect ratio), $n_{el} = 21$ (number of elements per direction per subdomain) and $E_1/E_2 = 10^{-5}$ (jump in the coefficient). The spectrum is shown in Figure 5.5 along with the first 11 generalized eigenvectors and corresponding eigenvalues. We observe that there is a gap in the spectrum of the generalized eigenproblem after the 9-th generalized eigenvalue since $\lambda^9 = 0.11$ and $\lambda^{10} = 0.98$. For this reason a judicious choice of the threshold for selecting eigenvectors which are put into the coarse space is for instance

$$\mathcal{K}_i = 0.15,$$

we will use this in all following numerical tests. With this criteria, the GenEO eigenproblem for a floating subdomain will provide 9 modes: the first three are rigid body modes included in the usual FETI natural coarse space, and 6 deformation modes that are included in the GenEO coarse space. As can be seen in Figure 5.5 those deformation modes represent the behavior of the subdomain when the hard layers deform the soft ones. The 9 modes can be seen as a basis to describe the nearly rigid motion of the hard layers (3 modes for each of the 3 layers, amounting to 9 modes) and the basis spanned by those modes represent the behavior of the domain as if the hard layers were its backbones. In some sense the GenEO coarse space can be interpreted in this case as a *skeleton* of the overall problem describing the dominant behavior of the structure according to its hard layers.

Next we actually solve the problem for different numbers of subdomains, different aspect ratios and different discretizations. The results are shown in Table 5.4. The two level method with the GenEO coarse space is robust throughout all of these tests: the condition number varies between 1.34 and 4.51 only, which is indeed lower than the upper bound given by the theory, $\mathcal{N}/\mathcal{K}_i = 20$, $\mathcal{N}$ being equal to three in this simple decomposition. Further the following observations are noteworthy:

– When the number of domains increases, the classical FETI-1 method sees its number of iteration increase significantly, whereas equipped with the GenEO coarse space,



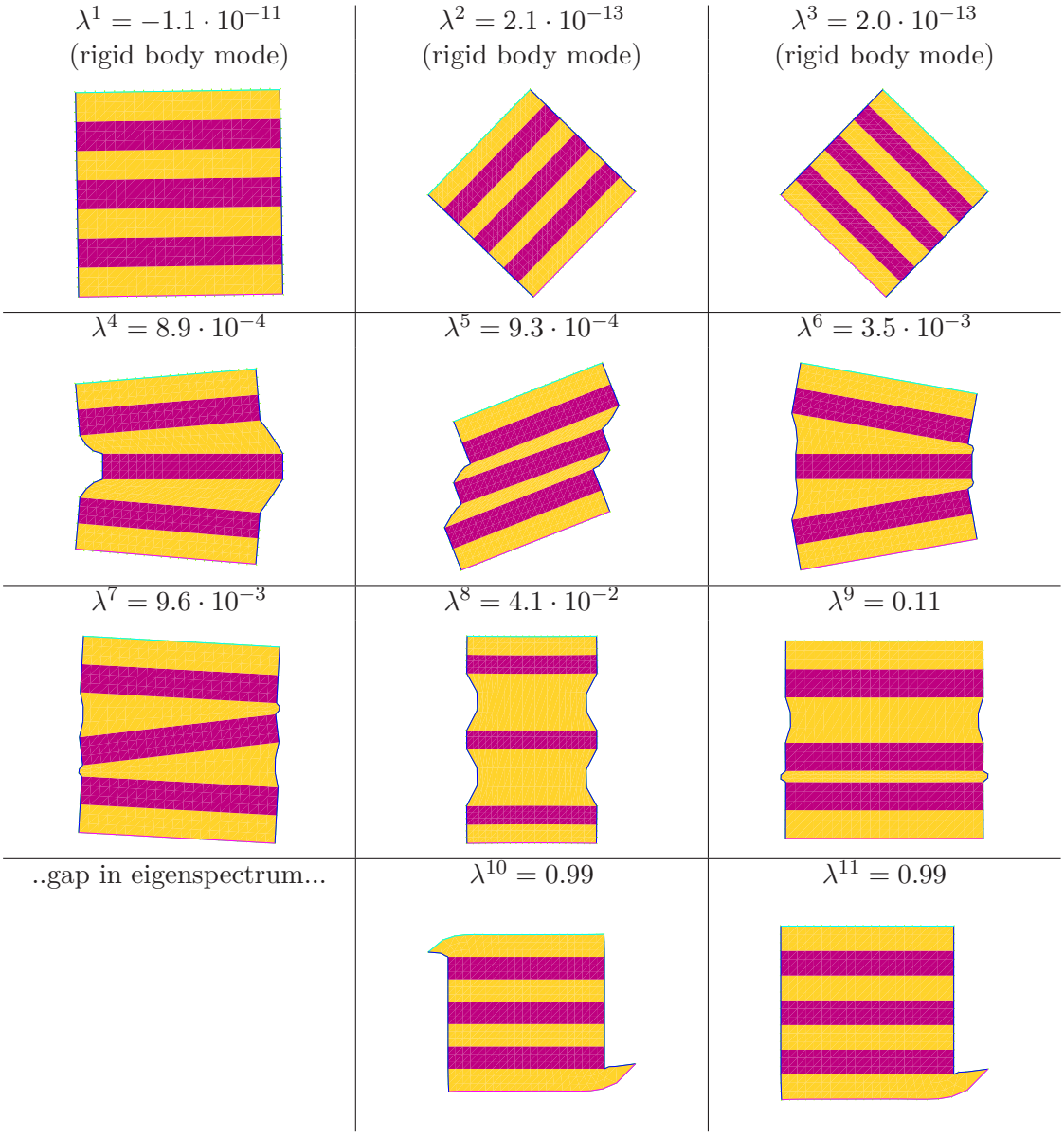


Figure 5.5: Eigenvalues and eigenmodes of the GenEO generalized eigenproblem for the geometry given in Figure 5.4 – dark or pink: hard material, light or yellow: soft material – The first eigenmodes (rigid body modes) are part of the natural coarse space, and the next 6 are selected for the GenEO coarse space.

Various number of subdomains (N), fixed aspect ratio ($b = 1$), fixed discretization ($n_{el} = 21$),fixed jump in coefficients ($E_1/E_2 = 10^{-5}$), the problem size increases with N

	FETI-GenEO			FETI-1	
N subdomains	κ	$\#U_0$	it	κ	it
4	3	14	5	$1.4 \cdot 10^3$	20
8	1.34	38	5	$1.9 \cdot 10^3$	39
16	1.34	86	4	$2.1 \cdot 10^3$	75
32	1.35	182	4	$2.2 \cdot 10^3$	137
64	1.35	374	4	$2.2 \cdot 10^3$	190

Various aspect ratios (b), fixed number of subdomains ($N = 8$), fixed discretization ($n_{el} = 21$),fixed jump in coefficients ($E_1/E_2 = 10^{-5}$)

	FETI-GenEO			FETI-1	
aspect ratio b	κ	$\#U_0$	it	κ	it
5	2.33	43	6	$1.7 \cdot 10^5$	$47^{(*)}$
2	1.42	40	5	$1.0 \cdot 10^4$	43
1	1.34	38	5	$1.9 \cdot 10^3$	40
1/2	4.51	27	9	446	33
1/5	4.07	14	11	70	22

$^{(*)}$ the relative residual reaches a plateau at $2 \cdot 10^{-3}$ after 47 iterations.

Various discretizations (n_{el}), fixed aspect ratios ($b = 1$), fixed number of subdomains ($N = 8$), fixed jump in coefficients ($E_1/E_2 = 10^{-5}$), the problem size increases with n_{el}.

	FETI-GenEO			FETI-1	
n_{el} elements	κ	$\#U_0$	it	κ	it
21	1.34	38	5	$1.92 \cdot 10^3$	39
42	1.42	38	5	$1.93 \cdot 10^3$	40
70	1.46	38	5	$1.94 \cdot 10^3$	40
84	1.47	38	5	$1.94 \cdot 10^3$	40

Table 5.4: Three tests for the geometry in Figure 5.4 – κ : condition number; $\#U_0$: size of the GenEO coarse space; it: number of iterations

the number of iteration remains small. The dimension of the GenEO coarse spaces is roughly proportional to the number of domains in this case.

– The classical FETI method convergences very slowly when the height of the domain is large compared to its width ($b = 5$). For that case the GenEO strategy generates only a small number of modes (43 in total) and converges very fast.
– For this layered structure, the preconditioned interface problem of FETI-1 has a condition number that barely depends on the number of elements per domain, and the number of iterations is nearly invariant with respect to the discretization step. When equipped with the GenEO coarse space, a small number of modes is included in the coarse space (38 GenEO modes, independent of the discretization step), and the number of iteration is very small

It is thus remarkable that the GenEO coarse space can handle automatically (once a proper threshold $\mathcal{K}$ has been chosen) the difficult cases of bad aspect ratios and heterogeneities along the interface.

5.5.3 Discontinuities along and across interfaces

In this subsection we consider the case of Figure 5.6 where the only difference with the previous subsection is that we have added jumps across the interfaces in subdomains 3 and 6 by inverting the soft and hard layers. The parameters are as follows: $n_{el} = 21$

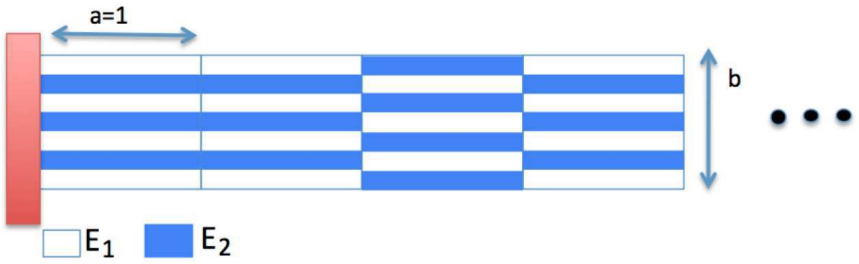


Figure 5.6: Discontinuities across and along interfaces (subdomains 3 and 6)

	FETI-GenEO			FETI-1	
scaling (D_i)	κ	it	$\#U_0$	κ	it
K-scaling	3.71	9	46	$7.0 \cdot 10^4$	55
multiplicity	3.89	7	173	$4.5 \cdot 10^4$	$189^{(*)}$

$^{(*)}$ the relative residual reaches a plateau at $1.5 \cdot 10^{-3}$ after 189 iterations.

Table 5.5: Geometry given in Figure 5.6 (discontinuities across and along the interfaces), $n_{el} = 21$, $N = 8$, $E_1/E_2 = 10^{-5}$ – κ : condition number; $\#U_0$: size of the GenEO coarse space; it: number of iterations

elements in each direction and each subdomain, $N = 8$ subdomains, $\nu = 0.3$ for Poisson's ratio, $E_1/E_2 = 10^{-5}$ for the magnitude of the jump in the coefficient, $b = 1$ for the aspect ratio of the subdomains and $\mathcal{K}_i = 0.15$ for the threshold on the GenEO eigenvalues. This is a known hard problem for FETI even with the Dirichlet preconditioner (which we use here again). In this case we show in Table 5.5 that with the K-scaling matrices (5.75) the number of *bad* eigenmodes is largely reduced compared to the case where multiplicity scaling is used (here multiplicity scaling reduces to setting all entries of each D_i to $1/2$). In deed with K-scaling we have selected 46 modes which is only 8 more than for the same case but without the extra jumps across the interfaces (see Table 5.4 – top – $N = 8$ subdomains). With the multiplicity scaling the GenEO strategy selects 173 modes. In fact, with K-scaling fewer modes are necessary because jumps across the interfaces are already accounted for in the preconditioner. The additional modes are needed to take into account the jumps across the interfaces. This confirms that GenEO compensates for the discrepancy between the preconditioner and the actual inverse of F: when inadequate weighting is used the preconditioner is less effective and hence a larger coarse space is needed. The condition numbers for both types of scaling are almost equal when the GenEO coarse space is introduced, which confirms the theory.

5.5.4 Choice of the threshold

Finally, we study the method on a unit square square discretized with a simplicial mesh consisting of 101×101 nodes and $\mathbb{P}_1$ finite elements. The local components of the diagonal scaling matrix D in the preconditioner are chosen to be the K-scaling matrices $D_i = diag(K_i)$. The coefficient distribution is given in Figure 5.7 along with two partitions of the domain into 25 subdomains. In both cases the interfaces do not match the jumps in the coefficients. The results are shown in Figure 5.6 where κ is the condition number of the preconditioned operator, $\#U_0$ is the number of bad eigenmodes selected in Definition 5.31 using the threshold $\mathcal{K}_i$. As is expected the condition number decreases when the threshold

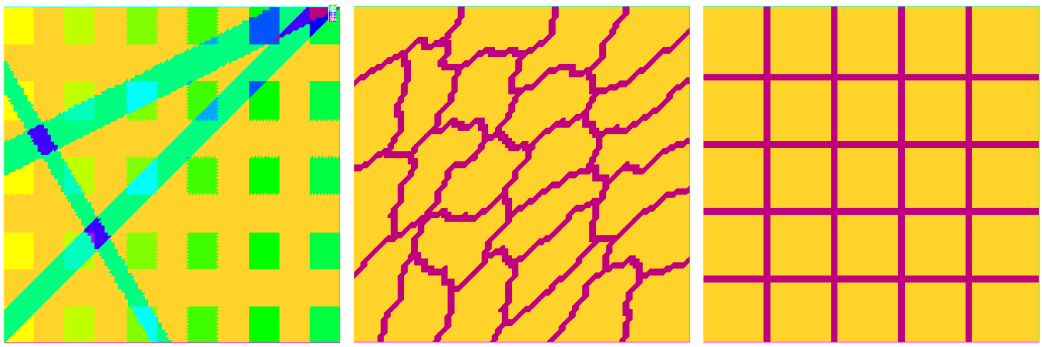


Figure 5.7: Left: Coefficients (Young's modulus $10^7 < E < 3 \cdot 10^{13}$; Poisson's ratio $0.3 < \nu < 0.4$) – Middle: Metis partition into 25 subdomains (1896 interface degrees of freedom) of the unit square – Right: Regular partition into 25 subdomains (1736 interface degrees of freedom)

	Metis partition		Regular partition	
$\mathcal{K}_i$	κ	$\#U_0$	κ	$\#U_0$
0	$2.9 \cdot 10^6$	0	$1.4 \cdot 10^5$	0
0.05	18.59	114	12.61	14
0.1	10.36	122	9.01	19
0.5	2.50	225	2.93	95
1	1.56	509	1.32	238
4	1.87	3295	1.00	3101

Table 5.6: Condition number κ and number of bad eigenvectors $\#U_0$ versus several values of the threshold $\mathcal{K}_i$ for the configurations in Figure 5.7

increases. In all cases the estimate is satisfied. We also observe that for a fixed threshold more eigenmodes are used to build the coarse space in the Metis partition case. This is in agreement with the fact that this is a harder problem.

5.6 Conclusion

We have constructed a two-level BDD method and two two-level FETI methods for which the convergence rates depend only on a chosen parameter and the maximal number of neighbours of a subdomain. The choice of this parameter is key in dimensioning the coarse space. Optimizing the choice of the parameter with respect to efficiency and the size of the coarse space is crucial. Here it has been set heuristically. For FETI the result holds for the full preconditioner based on solving Dirichlet problems in the subdomains and also on the lumped version which is a lot less expensive to implement. Compared to the Schwarz-GenEO algorithm these methods have the advantage of being non overlapping methods which means that they do not carry the extra cost of computations in the overlap.

In this Chapter the fundamental ideas and proofs underlying the GenEO coarse space have been explained and the numerical efficiency has been illustrated on problems hard to solve with classical FETI approaches. Future research should investigate the computational cost incurred by the GenEO coarse space (computation of the GenEO modes per domain, building and solving the coarse space) in order to assess the overall computational efficiency of the FETI-GenEO when applied to realistic engineering problems.

Chapter 6

Application to elasticity in the incompressible limit

Almost incompressible elasticity is a known challenge for domain decomposition methods. Tackling it was one of the objectives of this thesis since Michelin tires, as all tires, are made of rubber and rubber is an almost incompressible material. In the introduction of this manuscript (Section 2.3.4) we have motivated, with a Fourier analysis, the reason why we chose to switch from Additive Schwarz methods to substructuring methods. Here we introduce more precisely the almost incompressible framework and then illustrate the behaviour of the GenEO algorithms: we will explain why the Schwarz-GenEO coarse space cannot handle the incompressible limit and show that on the other hand FETI-GenEO performs very well.

Contents

6.1 Almost incompressible elasticity

6.1.1 The need for a particular discretization scheme

Let Ω be an open subset of $\mathbb{R}^d$ for $d = 2$ or $d = 3$. Let $\partial\Omega$ be the boundary of Ω and $\partial\Omega_D \subset \partial\Omega$ be a part of the boundary where a homogeneous Dirichlet boundary condition is imposed. Next, introduce the space $\mathcal{V} := \{v \in H^1(\Omega)^d : v_{|\Omega_D} = 0\}$. For a given body force f, the variational formulation of the linear elasticity equations can be written: find

the set of displacements $v \in \mathcal{V}$ such that

$$2\int_\Omega \mu\, \epsilon(u) : \epsilon(v) dx + \int_\Omega \lambda(\nabla \cdot u)(\nabla \cdot v) dx = \int_\Omega \langle f, v\rangle dx \quad \forall v \in \mathcal{V}, \tag{6.1}$$

where the contribution of the linearized strain tensor is

$$\epsilon(u) : \epsilon(v) := \sum_{i=1}^{d}\sum_{j=1}^{d} \epsilon_{ij}(u)\epsilon_{ij}(v); \quad \epsilon_{ij}(u) := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

$$\langle f, v\rangle := \sum_{i=1}^{d} f_i v_i,$$

and μ and λ are two parameters called the Lamé parameters which describe the material and can be expressed in terms of Young's modulus E and Poisson's ratio ν as

$$\lambda := \frac{E\nu}{(1+\nu)(1-2\nu)},\ \mu := \frac{E}{2(1+\nu)}.$$

In this chapter we assume that the material properties are constant throughout the domain. The incompressible limit is the following limit on Poisson's ratio: $\nu \to \frac{1}{2}$. This in turns implies that $\lambda \to +\infty$. For classical choices of finite elements (such as the standard $\mathbb{P}_1$ elements we have used on previous test for elasticity) the solution becomes strongly mesh dependent. This is known as the locking phenomenon and it deteriorates the solution to the point where it becomes unreliable. We illustrate this rather naively in Figure 6.1. In the top plot we have solved (6.1) for $E = 2 \cdot 10^7$ and $\nu = 0.4999$ in $\Omega = [0,3] \times [0,1] \times [0,1]$ with zero Dirichlet boundary conditions at $x = 0$, free surfaces on the remainder of the boundary and a body force corresponding to a gravity term in the z direction: $f(v) = \langle v, g\rangle$ and $g = (0, 0, -10)^\top$. There is obviously a problem with the solution because even without taking into account the values of the magnitudes of the displacements we notice that the (x, z) median plane is not an axe of symmetry as it should be. In the bottom of the figure we show that with an almost incompressible formulation we recover the right symmetry properties. Next we introduce this formulation.

6.1.2 An almost incompressible formulation

Even when $\lambda \to +\infty$, the term $\int_\Omega \lambda(\nabla \cdot u)(\nabla \cdot v)dx$ must remain bounded for all v and so this implies that $\nabla \cdot u \to 0$ or, by the divergence theorem, that in the incompressible limit the volume remains constant. This is the explanation for the locking phenomenon: if each element does not have enough degrees of freedom it cannot move while satisfying the constant volume constraint. A well-known remedy [8, 62, 6] is to introduce the new variable $p = \lambda\nabla \cdot u$ referred to as the pressure variable and living in a space $\mathcal{P} \subset L^2(\Omega)$. With it the problem can be written: find $(u, p) \in \mathcal{V} \times \mathcal{P}$ such that

$$\int_\Omega \left[2\mu\epsilon(u) : \epsilon(v) + p\, \nabla \cdot v\right] dx = \int_\Omega f(v) dx \quad \forall v \in \mathcal{V}, \tag{6.2}$$

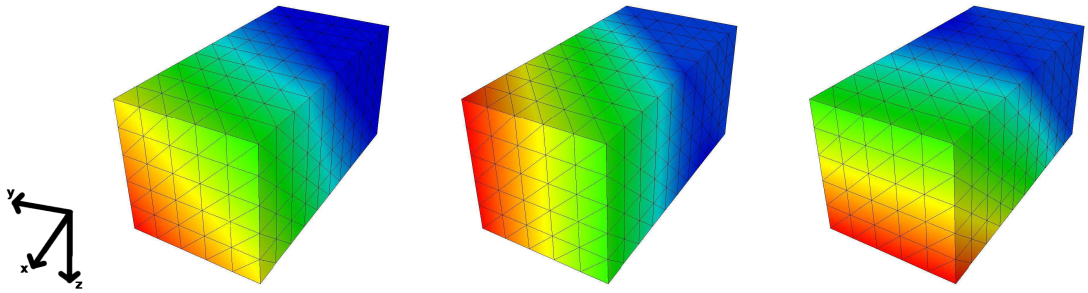
under the constraint that $p = \lambda\nabla \cdot u$.

Finally, the mixed (pressure - displacement) formulation of the problem is: find $(u, p) \in \mathcal{V} \times \mathcal{P}$ such that

$$\int_\Omega \left[2\mu\epsilon(u) : \epsilon(v) + p\, \nabla \cdot v\right] dx = \int_\Omega f(v) dx \quad \forall v \in \mathcal{V}.$$

$$\int_\Omega p\, q dx = \int_\Omega \lambda q \nabla \cdot u dx \quad \forall\, q \in \mathcal{P}. \tag{6.3}$$

Classical formulation (6.1), $\mathbb{P}_2$ elements:



Almost incompressible (penalized) formulation (6.12), $\mathbb{P}_2 - \mathbb{P}_0$ elements:

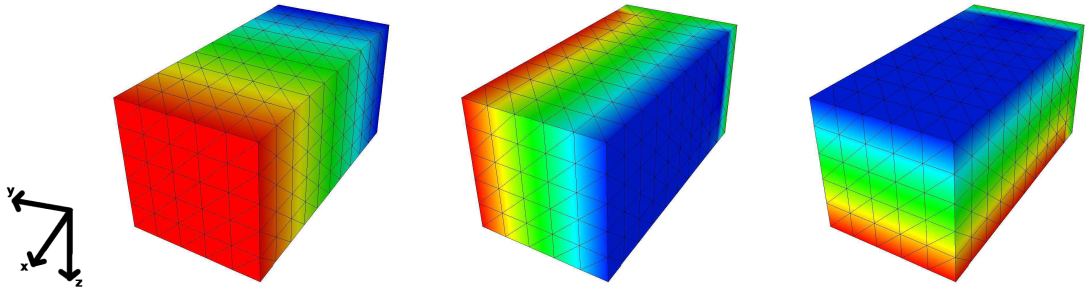


Figure 6.1: $x-$ (left), $y-$ (middle), and $z-$ (right) components of the displacement field. The domain is clamped at $x = 0$ and subject to gravity along z. The Lamé coefficients are $E = 2 \cdot 10^7$, $\nu = 0.4999$. With the classical formulation (top) the symmetry is not right which points to the fact that the formulation is flawed. The almost incompressible formulation (bottom) corrects this default of symmetry which points toward the fact that the solution is more reliable.

Next we discretize the problem. The choice of the finite element spaces is very important. In particular the spaces should satisfy the inf-sup condition. The choice $Q_2 - P_1$ of continuous tri-quadratics and discontinuous piecewise linears, for instance, is a good choice. Here we have made the more simple choice to use Lagrange finite elements: $\mathbb{P}_2 - \mathbb{P}_0$. More precisely we introduce the space

$$\mathcal{V}_h = \mathbb{P}_2^d \text{ for the field of displacements,}$$

and

$$\mathcal{P}_h = \mathbb{P}_0 \text{ for the field of pressures.}$$

Although this choice of elements does not satisfy the discrete inf-sup condition it is known to be stable. We write the discretized problem as: Find $(u^h, p^h) \in \mathcal{V}_h \times \mathcal{P}_h$ such that

$$\begin{aligned} \left[\int_\Omega 2\mu\epsilon(u_h) : \epsilon(v_h) + p_h \nabla \cdot v_h\right] dx &= \int_\Omega f(v_h)dx \quad \forall v_h \in \mathcal{V}_h, \\ \int_\Omega \frac{1}{\lambda} p_h \, q_h dx - \int_\Omega q_h \nabla \cdot u_h dx &= 0 \quad \forall \, q_h \in \mathcal{P}_h. \end{aligned} \tag{6.4}$$

It is obvious that we have made an approximation in writing the discretized system. Indeed we have replaced $\lambda \nabla \cdot u_h$ by p_h and imposed that p_h be a constant over each mesh element.

The equivalent matrix version of this system is the *perturbed* formulation: Find $(\mathbf{u}, \mathbf{p}) \in \mathbb{R}^m \times \mathbb{R}^n$ such that:

$$\begin{pmatrix} A & B^\top \\ B & -C \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix}, \tag{6.5}$$

where, given a basis $\{\phi_k\}$ of $\mathcal{V}_h$ and a basis $\{\psi_k\}$ of $\mathcal{P}_h$, the coefficients in the matrices are

$$a_{kl} = \int_\Omega 2\mu\epsilon(\phi_k) : \epsilon(\phi_l)dx, \tag{6.6}$$

$$b_{kl} = \int_\Omega \psi_k \nabla \cdot \phi_l dx, \tag{6.7}$$

$$c_{kl} = \int_\Omega \frac{1}{\lambda} \psi_l \cdot \psi_k dx, \tag{6.8}$$

and

$$f_k = \int_\Omega f(\phi_k)dx. \tag{6.9}$$

Because the ψ_j are the basis functions for the $\mathbb{P}_0$ finite elements, C is a diagonal with coefficients $c_{ii} = \int_{\tau_i} \frac{1}{\lambda} = \frac{1}{\lambda}\text{area}(\tau_i)$.

The perturbed formulation is equivalent to the following system on the vector valued unknown

$$\begin{aligned} &A\,\mathbf{u} + B^\top \mathbf{p} = \mathbf{f}, \\ &B\,\mathbf{u} - C\,\mathbf{p} = 0. \end{aligned} \tag{6.10}$$

From the second equation we get $B\,\mathbf{u} = C\,\mathbf{p}$. Since C is by definition non singular, we can write $\mathbf{p}$ as a function of $\mathbf{u}$ as

$$\mathbf{p} = C^{-1} B\,\mathbf{u}, \tag{6.11}$$

and inject this into the first equation in order to get the *penalized* formulation

$$\widetilde{A}\,\mathbf{u} := A\,\mathbf{u} + B^\top C^{-1} B\mathbf{u} = \mathbf{f}. \tag{6.12}$$

We have recovered a pure displacement formulation. As we have already noted, we have made an approximation so the penalized formulation (6.12) is less accurate than the original formulation. It is still a better choice in the incompressible limit because it is stable.

Remark 6.1. Although we refer to (6.12) as the penalized formulation we have not strictly speaking used a penalization procedure. This would have been necessary if the material was incompressible ($\nu = 1/2$) in which case the block matrix in (6.5) would have been $\begin{pmatrix} A & B^\top \\ B & 0 \end{pmatrix}$. The penalization technique replaces the zero block by ϵ times the identity where ϵ is very small and then eliminates the pressure variable using static condensation just like we did.

We have already assumed that λ is a constant. If furthermore the mesh is regular then C is a diagonal matrix with coefficients $c_{ii} = |\tau|/\lambda$, $|\tau|$ being the volume of a mesh element and the equation is

$$A\,\mathbf{u} + \frac{\lambda}{|\tau|} B^\top B\mathbf{u} = \mathbf{f} \tag{6.13}$$

which is particularly easy to implement.

We remark that another option would have been to solve directly the augmented formulation of the problem by summing both equations in system (6.4): Find $(u_h, p_h) \in \mathcal{V}_h \times \mathcal{P}_h$ such that for all $(v_h,\, q_h) \in (\mathcal{V}_h,\, \mathcal{P}_h)$

$$\int_\Omega \left[2\mu\epsilon(u_h) : \epsilon(v_h) + p_h\, \nabla \cdot v_h + q_h \nabla \cdot u_h - \frac{1}{\lambda} p_h\, q_h \right] dx = \int_\Omega f(v_h) dx. \tag{6.14}$$

The advantage which is not of particular interest to us here is that then the pressure field is known.

6.1.3 We may apply GenEO to the almost incompressible formulation

The main assumption in order to compute the GenEO coarse spaces for the penalized formulation (6.12) is that the problem matrix $\tilde{A}$ be a sum of element matrices $\tilde{A}_{\tau_k}$ over all the elements τ_k in the mesh $\mathcal{T}_h$. This way we can compute the restrictions of $\tilde{A}$ to each subdomain Ω_j and the overlap Ω_j°. Let's make sure that this is indeed the case for $\widetilde{A} = A + \frac{\lambda}{|\tau|} B^\top B$. Obviously the only part which may be problematic is $B^\top B$. By definition, for $i, j = 1, \ldots, \#\mathcal{P}_h$ the entries are

$$\begin{aligned}
\left(B^\top B\right)_{ij} &= \sum_{k=1}^{\#\mathcal{P}_h} \left(B^\top\right)_{ik} (B)_{kj} \\
&= \sum_{k=1}^{\#\mathcal{P}_h} (B)_{ki}\, (B)_{kj} \\
&= \sum_{k=1}^{\#\mathcal{P}_h} \left(\int_\Omega \psi_k \nabla \cdot \phi_i dx \right) \left(\int_\Omega \psi_k \nabla \cdot \phi_j dx \right) \\
&= \sum_{k=1}^{\#\mathcal{T}_h} \left(\int_{\tau_k} \psi_k \nabla \cdot \phi_i dx \right) \left(\int_{\tau_k} \psi_k \nabla \cdot \phi_j dx \right).
\end{aligned}$$

So $B^\top B$ wan also be written as a sum of element matrices $\left(B^\top B\right)_{\tau_k}$ with entries

$$\left(\left(B^\top B\right)_{\tau_k} \right)_{ij} = \left(\int_{\tau_k} \psi_k \nabla \cdot \phi_i dx \right) \left(\int_{\tau_k} \psi_k \nabla \cdot \phi_j dx \right).$$

Denoting by B_{τ_k} the matrix with entries $\int_{\tau_k} \psi_j \nabla \cdot \phi_i dx$ we have $B = \sum_{\tau_k \in \mathcal{T}_h} B_{\tau_k}$ and

$$\begin{aligned}\left(B_{\tau_k}^\top B_{\tau_k}\right)_{ij} &= \sum_{l=1}^{\#\mathcal{P}_h} \left({B_{\tau_k}}^\top\right)_{il} (B_{\tau_k})_{lj} \\ &= \sum_{l=1}^{\#\mathcal{P}_h} (B_{\tau_k})_{li} (B_{\tau_k})_{lj} \\ &= (B_{\tau_k})_{ki} (B_{\tau_k})_{kj}\end{aligned}$$

because all other terms in the sum are zero by definition of B_{τ_k} and $\mathcal{P}_h$. Finally $B_{\tau_k}^\top B_{\tau_k} = \left(B^\top B\right)_{\tau_k}$ and $\left(B^\top B\right)_{\tau_k}$ is positive semi definite:

$$\langle \left(B^\top B\right)_{\tau_k} u, u\rangle = \langle B_{\tau_k} u, B_{\tau_k} u\rangle \geq 0,$$

so we can compute the two local matrices needed for the GenEO eigenproblems.

6.2 Schwarz-GenEO and the incompressible limit

Next we consider solving the penalized formulation of the problem (6.12) preconditioned by the two level Schwarz preconditioner. Although in the introduction we concluded that Additive Schwarz is not the best preconditioner for this problem we study whether or not the GenEO coarse space can fix the slow convergence. First we introduce a good choice for the coarse space from the literature to show that such a choice exists.

6.2.1 Analytical coarse space

In [18] the authors propose a coarse space for the penalized formulation of the three dimensional linearized elasticity equations and a certain choice of the discretization spaces. They use a hybrid Schwarz preconditioner which means that the coarse correction in the two level preconditioner is a multiplicative contribution. With a coarse space consisting of 3 degrees of freedom per subdomain vertex, 5 degrees of freedom per subdomain edge and 1 degree of freedom per subdomain face they prove that the condition number of the preconditioned operator P_{hy} is bounded by

$$\kappa(P_{hy}) \lesssim \left(\frac{H}{\delta}\right)^3 \left(1 + \log\left(\frac{H}{h}\right)\right)^2,$$

where H is the subdomain size, δ is the width of the overlap, h is the mesh size and the constant hidden in $\lesssim$ does not depend on the number of subdomains, their diameters, the mesh size and the values of the Lamé parameters. It depends only on the shape regularity of the elements and the subdomains.

In particular this convergence result holds even in the incompressible limit. This tells us that there does exist a reasonably sized coarse space with which we can achieve robustness with respect to the almost incompressible behaviour using the Schwarz preconditioner. Unfortunately although the theoretical results will not be proved wrong GenEO does find this good coarse space for Additive Schwarz.

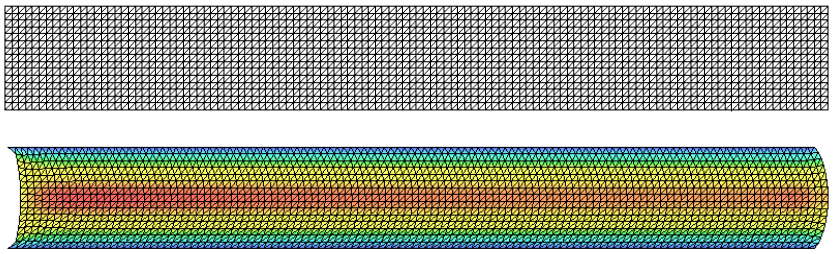


Figure 6.2: top: Original Mesh – bottom: Displaced mesh, the displacement on the left hand side boundary is prescribed and propagates through the domain, the color is a measure of the displacement in the x-direction: it is zero on the top and bottom boundaries and maximal on the median horizontal line.

6.2.2 Schwarz-GenEO does not find the 'best' coarse space

For this test case the domain is two-dimensional: $(x, y) \in [0, 8] \times [0, 1]$. The boundary conditions are: $u(0, y) = (1/2^2 - (y - 1/2)^2, 0)$, $u(x, 0) = (0, 0)$, $u(x, 1) = (0, 0)$ and the right hand side boundary is free. There is no body force. If the material were strictly incompressible we would be studying the Poiseuille flow and the theoretical solution would be $u(x, y) = (1/2^2 - (y - 1/2)^2, 0)$.

The domain, is split into 24×3 subdomains, each of size $1/3 \times 1/3$. Every subdomain is discretized using a regular mesh with 6×6 mesh nodes and then extended by 1 layer of elements over each of its neighbours. Finally, the choice for the Lamé coefficients is $E = 10^7$ and ν varies between $\nu = 0.4$ and $\nu = 0.4999$.

Figure 6.2 shows the numerical solution for $\nu = 0.4999$ computed with Freefem++ [50]: the prescribed displacement propagates throughout the domain, this is what is expected.

Now we build the GenEO coarse space (4.20) and study more particularly the solution of the GenEO generalized eigenproblem (4.18) for a floating subdomain with both formulations (a standard $\mathbb{P}^2$ discretization of (6.1) and the $\mathbb{P}_2 - \mathbb{P}_0$ penalized formulation (6.12)). Even though neither of these formulations is a good choice for the whole range of ν we make Poisson's ratio vary between $\nu = 0.4$ and $\nu = 0.4999$ with $E = 10^7$ in all cases. The values of the first 50 eigenvalues are shown in Figures 6.3 and 6.4. Because of the logarithmic scale we have not plotted the three first eigenvalues which are zero in all cases.

We notice that the spectrum does not vary very much depending on the formulation. The most important remark is that if we set the criterion to $\mathcal{K}_i = 0.5$ or even $\mathcal{K}_i = 0.1$ as we have in previous numerical examples, as soon as we approach the almost incompressible limit the coarse space becomes very large (recall that the number of degrees of freedom per subdomain is small in this example). What this means is that GenEO detects the challenge posed by the almost incompressible behaviour (it wants to enrich the coarse space) but it is not possible to conclude that just a few generalized eigenvectors slow down convergence and hence that a small coarse space will guarantee fast convergence. This becomes even more clear looking at the first ten eigenvectors plotted in Figure 6.5 (in increasing eigenvalue order): the first three are the rigid body modes and then there is no particular trend.

In the next section we will observe that the FETI-GenEO eigenproblem deals very well with the almost incompressible limit. First we try to explain the shortcoming of

Schwarz-GenEO.

6.2.3 A heuristic explanation

It is quite easy to understand heuristically what is happening. If we look at the GenEO eigenproblem in subdomain Ω_j and in matrix formulation it reads: find $(\lambda, \mathbf{v}) \in \mathbb{R}^+ \times \mathbb{R}^{n_j}$ such that

$$\mathcal{A}_{\Omega_j}\mathbf{v} = \lambda D_j \mathcal{A}_{\Omega_j^\circ} D_j \mathbf{v} \tag{6.15}$$

where n_j is the number of degrees of freedom in Ω_j including the boundary, $\mathcal{A}_{\Omega_j}$ and $\mathcal{A}_{\Omega_j^\circ}$ are the matrices of the problem assembled only over subdomain Ω_j and the overlap Ω_j° respectively:

$$\left(\mathcal{A}_{\Omega_j}\right)_{kl} = 2\int_{\Omega_j} \mu\, \epsilon(\phi_k) : \epsilon(\phi_l) dx + \int_{\Omega_j} \lambda(\nabla\cdot\phi_k)(\nabla\cdot\phi_l) dx,$$

$$\left(\mathcal{A}_{\Omega_j^\circ}\right)_{kl} = 2\int_{\Omega_j^\circ} \mu\, \epsilon(\phi_k) : \epsilon(\phi_l) dx + \int_{\Omega_j^\circ} \lambda(\nabla\cdot\phi_k)(\nabla\cdot\phi_l) dx,$$

and D_j is the partition of unity matrix for subdomain Ω_j. In particular it has zero values corresponding to degrees of freedom on the boundary $\partial\Omega_j$.

Because D_j has zero values corresponding to the degrees of freedom on the boundary, $D_j \mathcal{A}_{\Omega_j^\circ} D_j$ is the matrix of a Dirichlet problem (imposed displacements on the boundary). In the incompressible limit any displacement u_j which does not preserve the volume of the subdomain has very high elastic energy: the energy with respect to the matrix on the right hand side of the generalized eigenvalue problem blows up if $D_j\mathbf{v}$ does not preserve the volume. Since the matrix on the left hand side of the generalized eigenvalue problem is the matrix of a problem with Neumann boundary condition there is no reason for $\langle \mathcal{A}_{\Omega_j}\mathbf{v}, \mathbf{v}\rangle$ to blow up also. This explains why there are so many tiny eigenvalues as soon as $\nu \to 1/2$ in Figures 6.3 and 6.4 and why the corresponding eigenvectors are pretty much all the vectors (see Figure 6.5 for the first ten).

6.3 FETI-GenEO and the incompressible limit

Now we illustrate the fact that FETI-GenEO builds a very small coarse space to deal with the incompressible limit.

6.3.1 Analytical coarse space

Looking at coarse spaces for solving the linear elasticity problem in the incompressible limit with FETI, in [114] it is is explained that, with a discontinuous pressure field, one coarse vector per subdomain suffices to ensure that the volume is preserved or, equivalently, that the net flux over the subdomain boundary is zero. More recently [43] arrives at the same conclusion and gives a theoretical analysis. This is due to the non overlapping nature of FETI. Next we will check that GenEO finds that coarse vector.

In [70] (which proposes a coarse space quite similar to GenEO but where the generalized eigenvalue problems are posed on an interface) an almost incompressible elasticity problem is also solved. The authors report that the adaptive process make it possible to recover good convergence. The size of the coarse space is given without any more detail but it seems that it is larger than just one vector per subdomain. A further comparison between [70, 103] and GenEO would be very interesting.

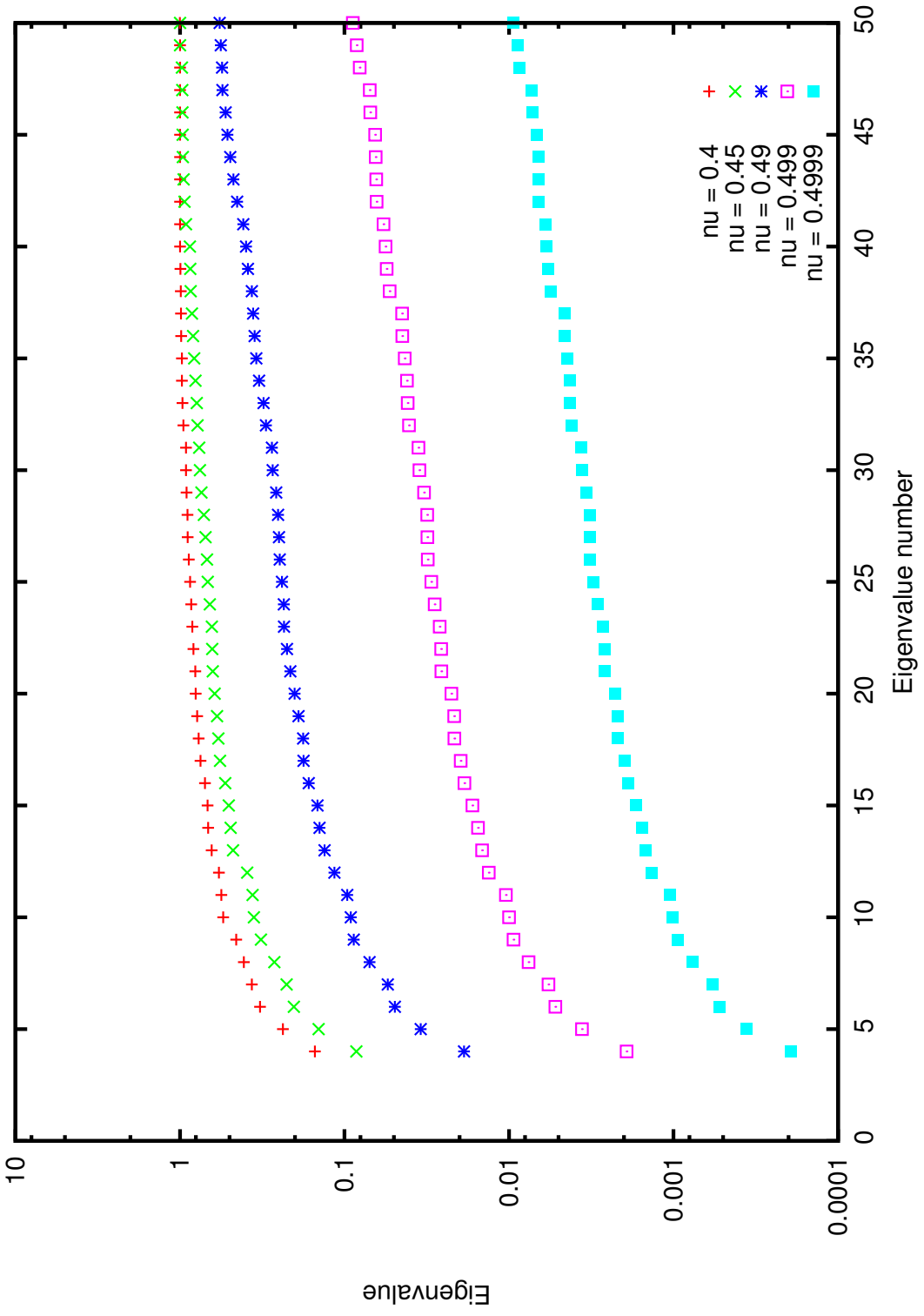


Figure 6.3: Schwarz-GenEO: Solution of the eigenproblem with the **classical** formulation (6.1) for a floating subdomain, ν varies between 0.4 and 0.4999. Eigenvalue (log scale) versus eigenvalue number.

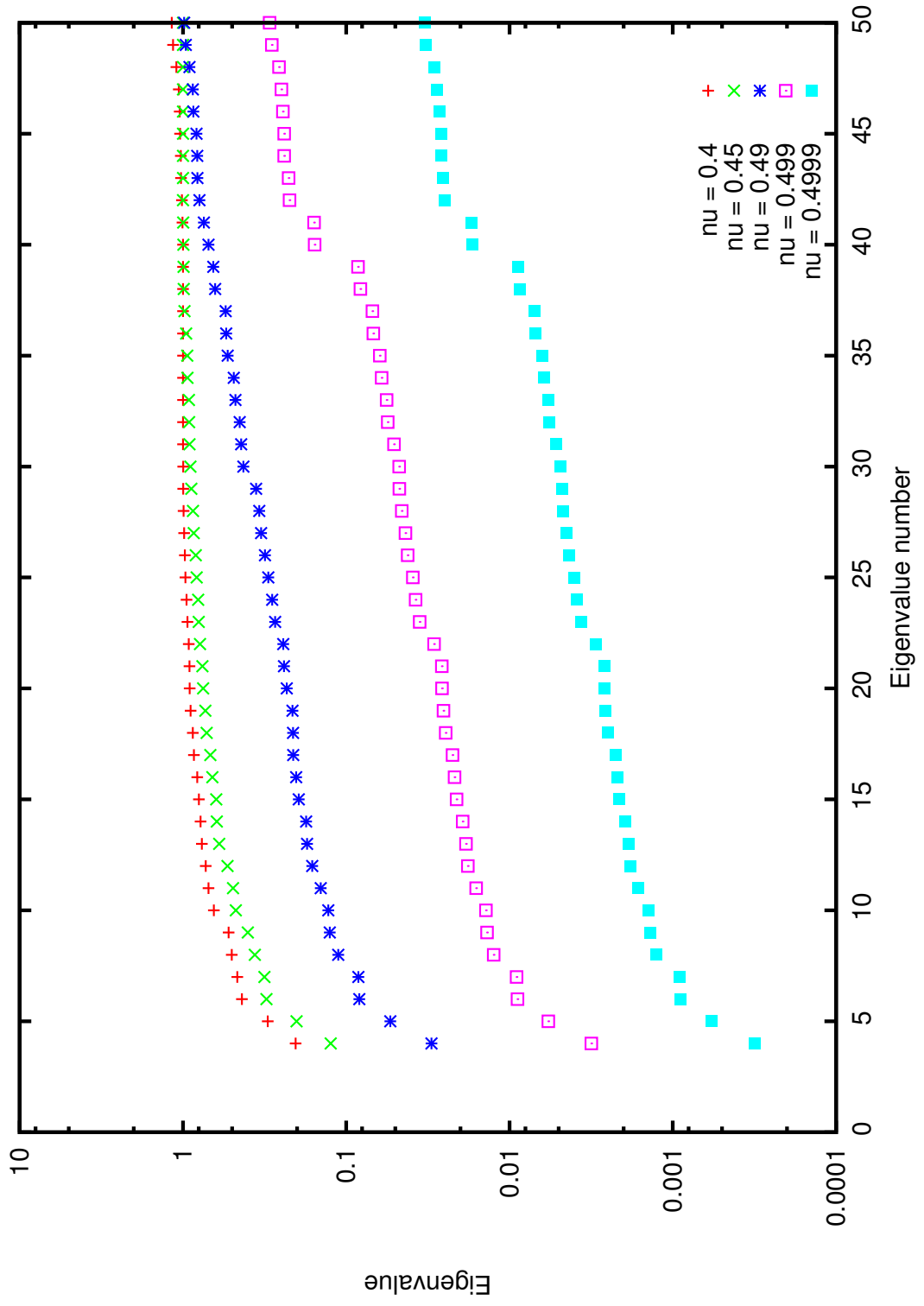


Figure 6.4: Schwarz-GenEO: Solution of the eigenproblem with the **penalized** formulation (6.12) for a floating subdomain, ν varies between 0.4 and 0.4999. Eigenvalue (log scale) versus eigenvalue number.

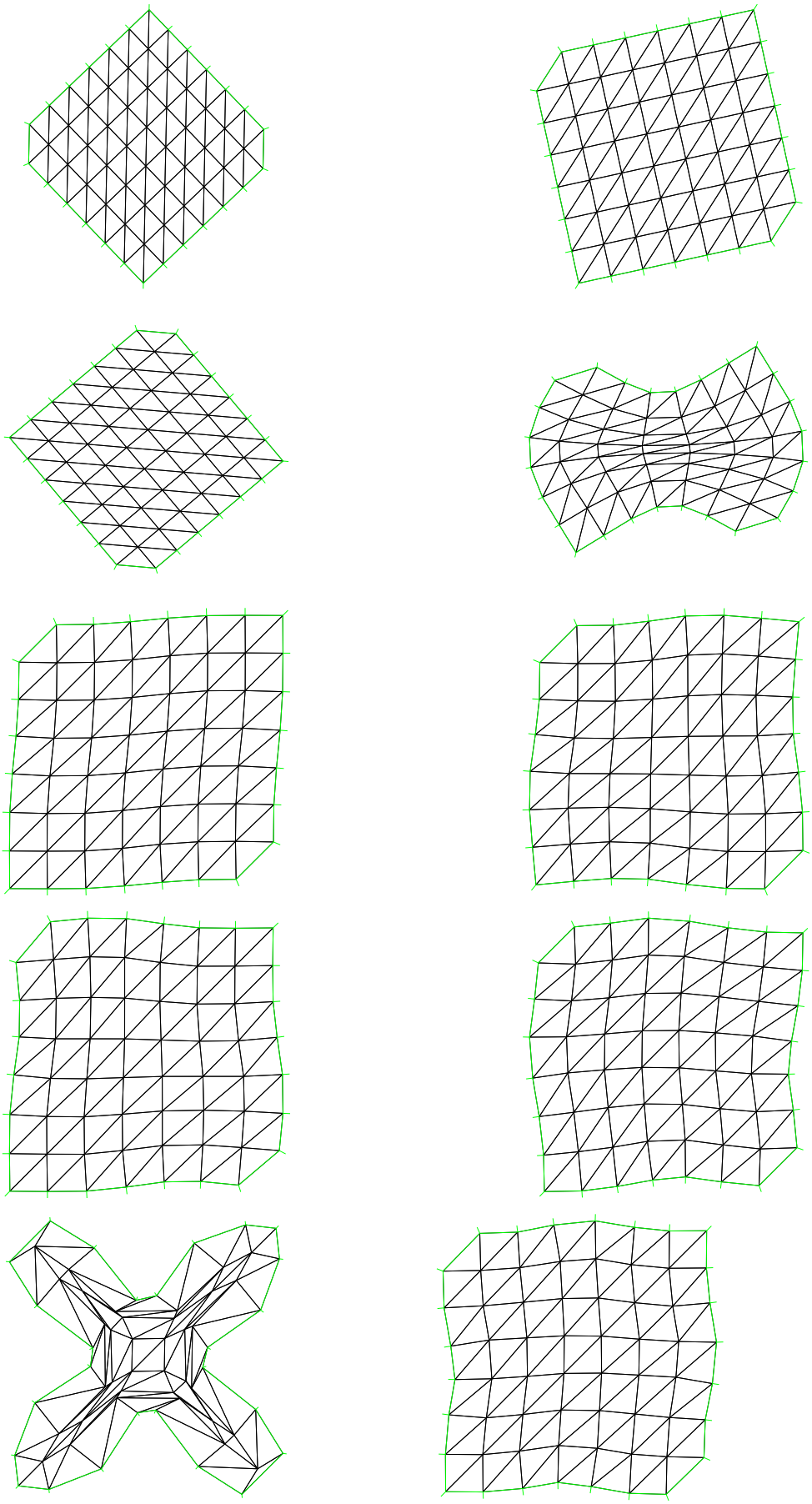


Figure 6.5: Schwarz-GenEO: $E = 10^7$ and $\nu = 0.499$ – Penalized formulation (6.12) – Eigenvectors 1 to 10 for a floating subdomain (the eigenvalues are 0; 0; 0; $3 \cdot 10^{-3}$; $6 \cdot 10^{-3}$; $9 \cdot 10^{-3}$; $1 \cdot 10^{-2}$; $1 \cdot 10^{-2}$; $1 \cdot 10^{-2}$).

6.3.2 GenEO finds the 'best' coarse space

We test the FETI-GenEO algorithm on the same test case as in the previous subsection. This time each subdomain is discretized with a regular mesh consisting of 11×11 nodes. Unless otherwise specified the FETI preconditioner is the Dirichlet preconditioner with K-scaling and the stopping criterion in the convergence tests is that the primal residual be reduced by 10^{-4}. In Figure 6.6 we plot the value of all eigenvalues smaller than 1 with the classical formulation (6.1) when ν varies between 0.4 and 0.4999. We notice that there are some small eigenvalues which appear as $\nu \to 1/2$ but there are far fewer than with Schwarz-GenEO and, maybe most importantly, there is a large gap in the spectrum which ensures that putting the smallest eigenvalues into the coarse space will actually help with convergence. Since we know that we cannot use the classical formulation in the incompressible limit we have not looked into this any further.

In Figure 6.7 we plot the eigenvalues smaller than 1 for the GenEO eigenproblem in a floating subdomain when the penalized formulation (6.12) is used and ν varies between 0.4 and 0.4999. It seems that GenEO is able to find exactly the one basis vector which is needed. Indeed, as usual, the first three eigenvalues are zero regardless of the value of ν and we also notice that the fifth eigenvalue is roughly 0.4 for all ν. The fourth eigenvalue is the interesting one: it varies very strongly with Poisson's ratio. In the incompressible limit it approaches zero whereas far away from the limit it approaches 0.4 (and hence the fifth eigenvalue). What this means is that the eigenvector corresponding to Λ^4 is directly related to the almost incompressible behaviour of the material.

We confirm this by showing in Figure 6.8 the eigenvectors corresponding to the first seven non zero eigenvalues. In the two first columns we compare the classical and penalized formulations at $\nu = 0.4$. The eigenvectors which arise are almost identical (although not necessarily in the same order), this is to show that the particular behaviour which we observe next is not just due to the new formulation but really to the incompressible limit. In the third column we plot the eigenvectors for $\nu = 0.4999$ and we notice that

- the first eigenvector in the column does not appear in any of the other families of eigenvectors. This is the eigenvector which corresponds to the very small fourth eigenvalue present only when $\nu \to 1/2$ in Figure 6.7.
- the next eigenvectors are very similar to eigenvectors which occur when $\nu = 0.4$. These would typically not be picked up by the coarse space since they are after the gap in the spectrum.

Finally the convergence results in table 6.1 confirm that with the FETI-GenEO coarse space we can ensure robustness with respect to ν even in the incompressible limit with a small coarse space. The size of the coarse space is not always exactly one vector per subdomain because we have used the automatic criterion $\tau = 0.15$ rather than impose its size.

As a last remark we have solved the FETI-GenEO eigenproblem for $\nu = 0.4999$ with the Lumped preconditioner. The spectrum is plotted in Figure 6.9. The result is really quite terrible with more than 100 very small eigenvalues. This is exactly what was to be expected: the Lumped preconditioner makes the assumption that the interior of each subdomain is infinitely hard and very many vectors are needed in the coarse space to make up for how wrong that is.

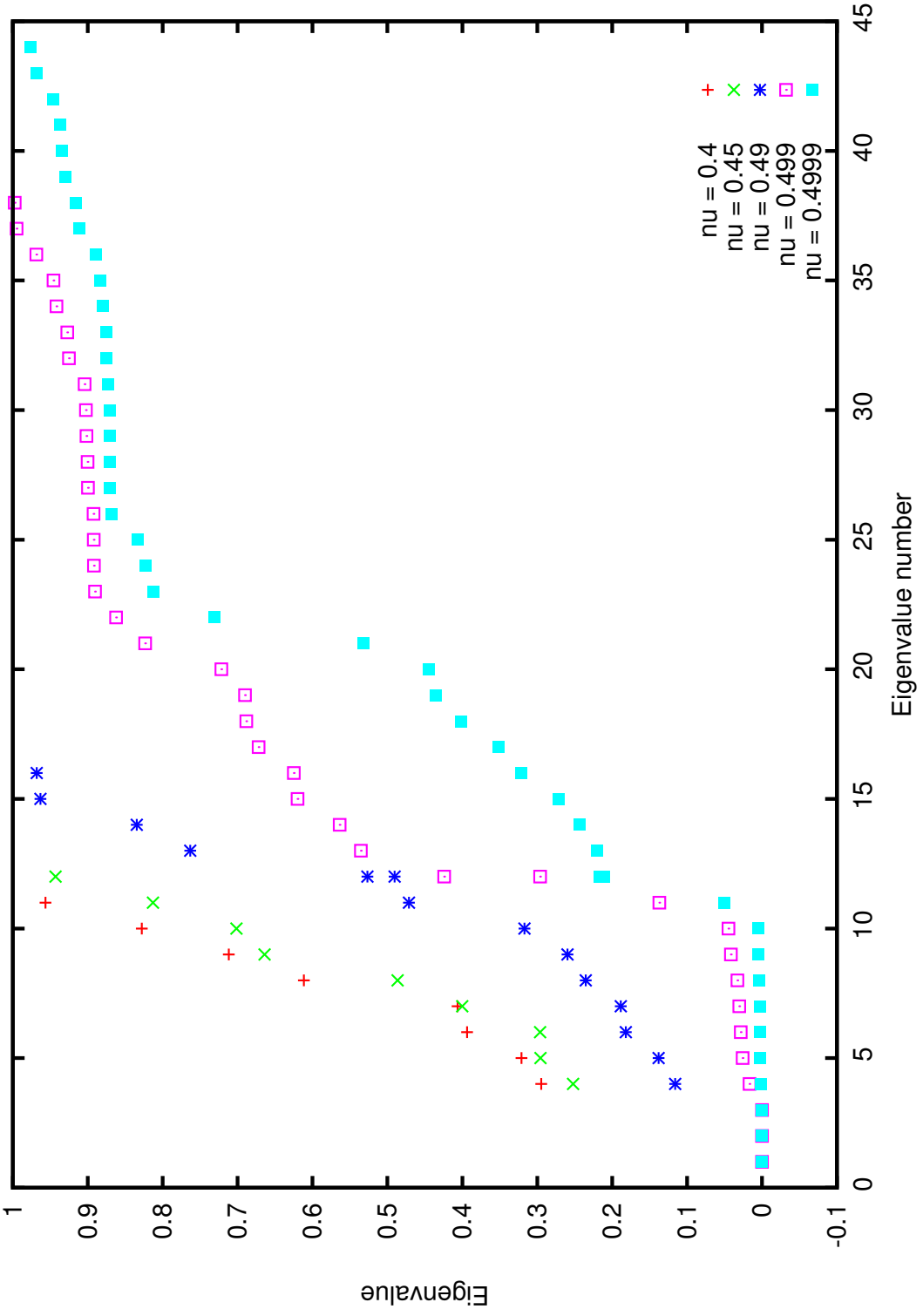


Figure 6.6: FETI-GenEO: **Classical** Formulation (6.1), Constant Coefficients: solution of the GenEO eigenproblem, for ν varying between 0.4 and 0.4999.

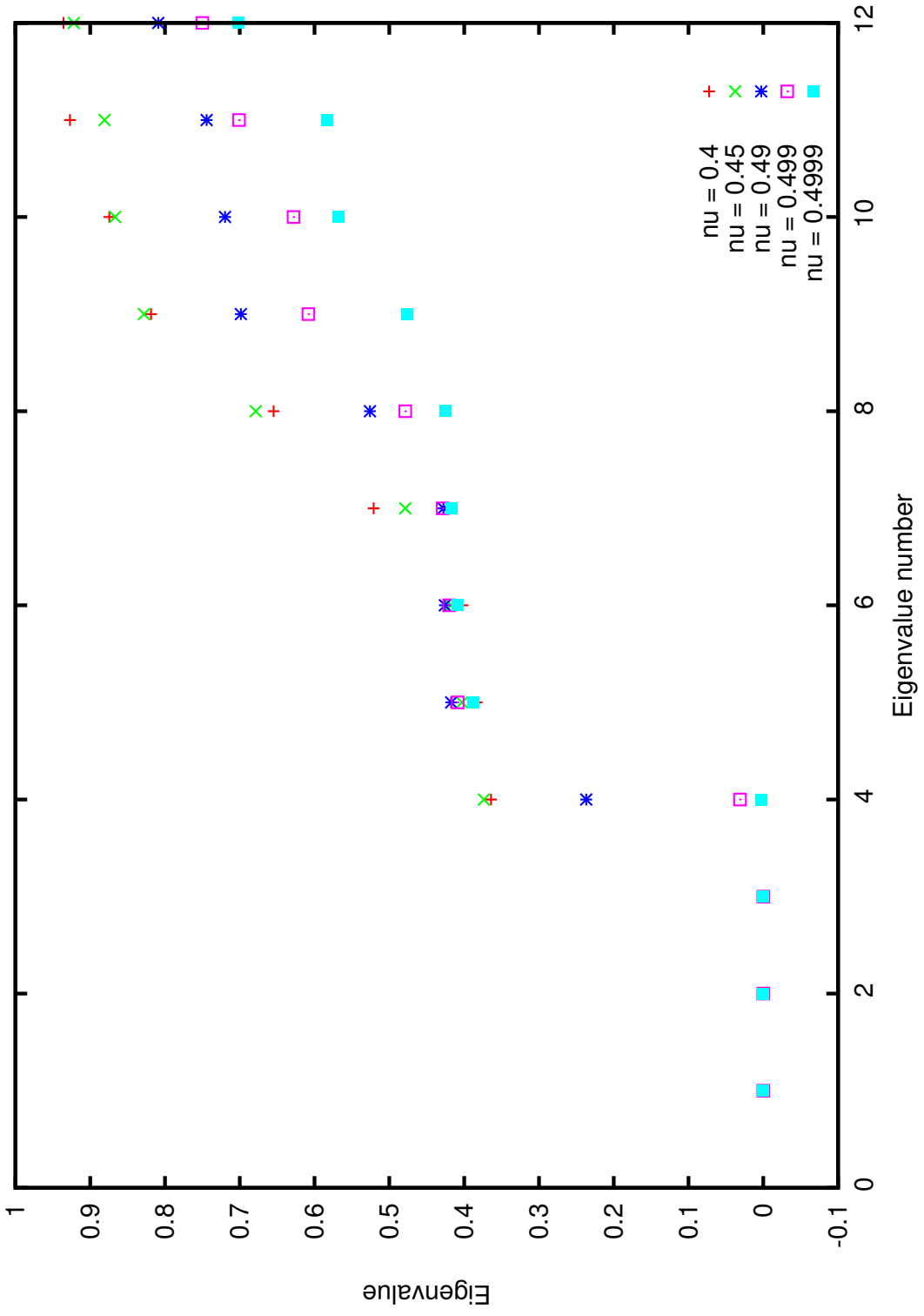


Figure 6.7: FETI-GenEO: **Penalized** Formulation (6.12), Constant Coefficients: solution of the GenEO eigenproblem, for ν varying between 0.4 and 0.4999. **In the incompressible limit just one bad eigenvalue appears!**

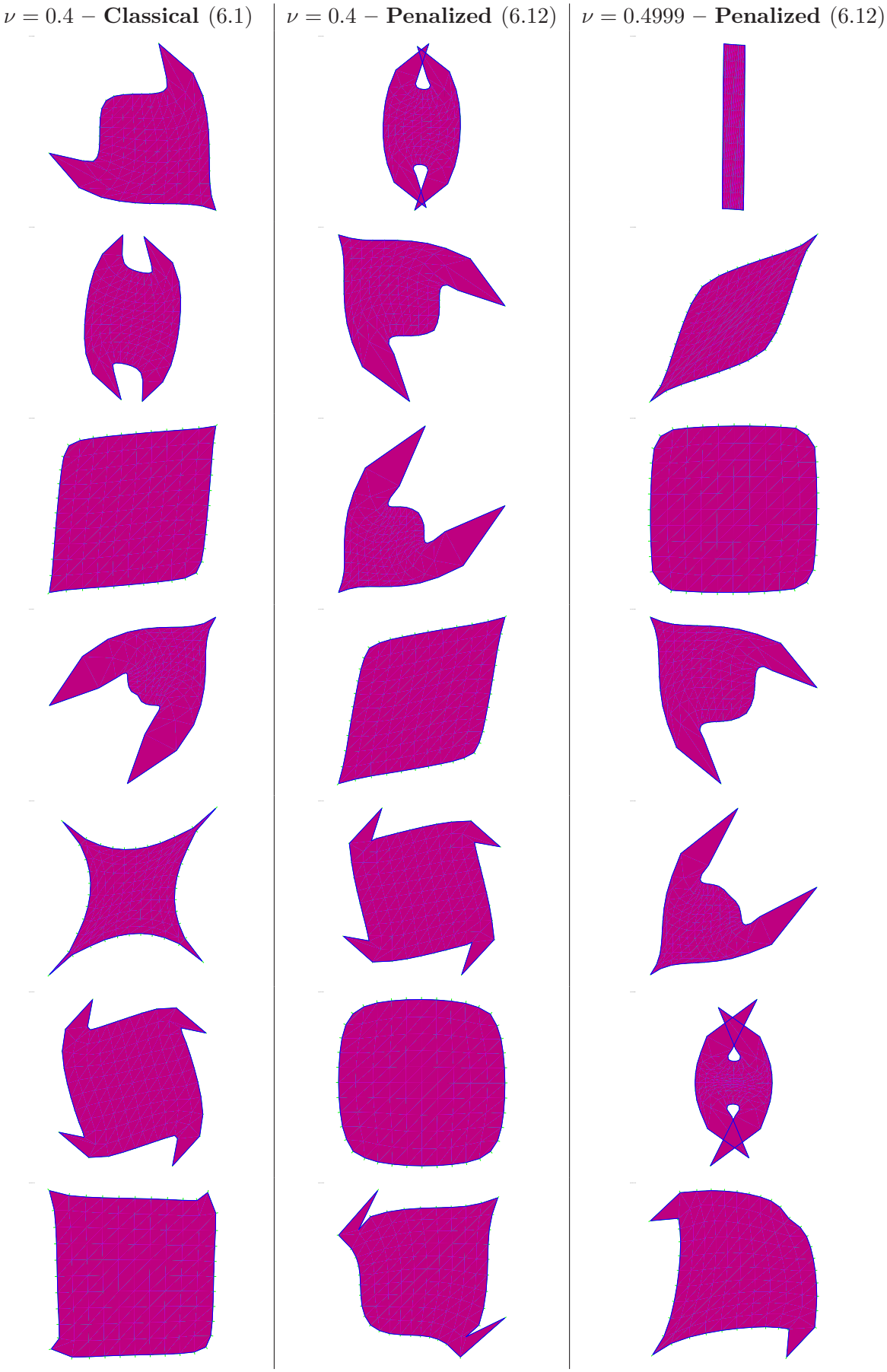


Figure 6.8: FETI-GenEO: Eigenvectors corresponding to the first seven non zero eigenvalues (in increasing order from top to bottom). For $\nu = 0.4$ the modes found with the classical and the penalized formulation are almost identical. At $\nu = 0.4999$ a new mode appears (top right), this is the almost incompressible mode.

Table 6.1: FETI convergence results, the criterion is always 0.15 – κ: condition number, it: number of iterations, $\#U_0$: size of the GenEO coarse space (does not include the rigid body modes). FETI-1 is the classical FETI.

	Penalized formulation (6.12)					Classical formulation (6.1)				
	FETI-GenEO			FETI-1		FETI-GenEO			FETI-1	
ν	κ	it	$\#U_0$	κ	it	κ	it	$\#U_0$	κ	it
0.4	5.9	12	16	13.4	19	7.1	14	16	15.9	21
0.45	5.7	12	16	14.5	19	8.2	16	16	18.4	23
0.49	11.9	15	16	34.8	23	8.3	15	57	49	33
0.499	5.2	11	37	281	28	3.1	9	156	380	52
0.4999	5.3	11	37	2749	30	6.6	15	158	3652	84

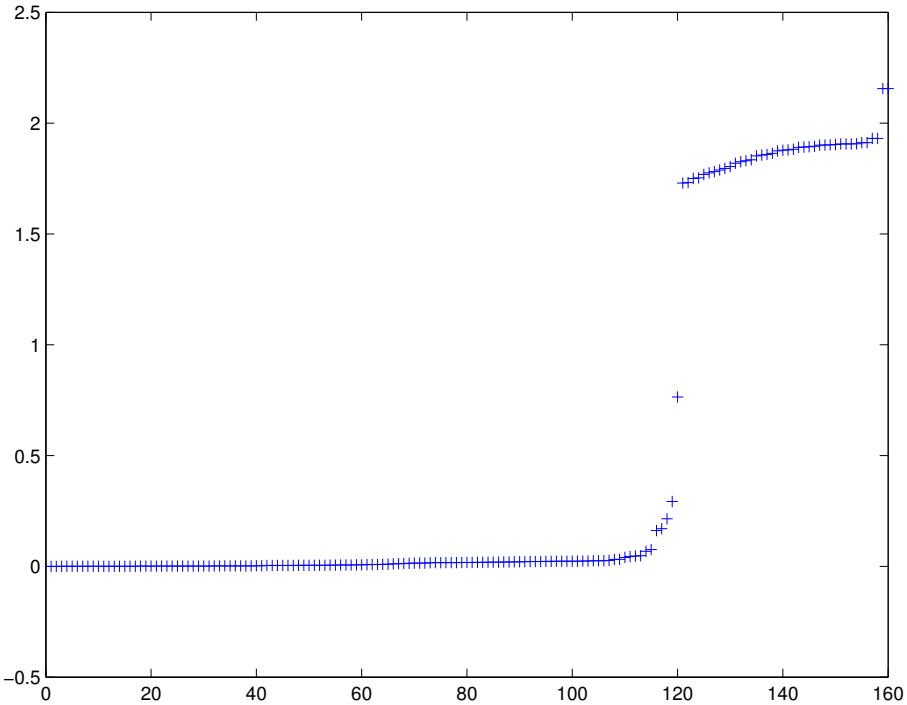


Figure 6.9: Solution (eigenvalue versus eigenvalue number) of the FETI-GenEO eigenproblem with the Penalized formulation (6.12) and the lumped preconditioner for FETI. Poisson's ratio is $\nu = 0.4999$.

6.4 Conclusion

In the introduction of this manuscript (Section 2.3.4), thanks to a Fourier analysis we explained why the choice to move on to FETI from Additive Schwarz stemmed from the need to solve elasticity problems in the almost incompressible limit. In this chapter we have illustrated the fact that our Schwarz-GenEO coarse space cannot provide a satisfying fix to compensate for the fact that Additive Schwarz performs rather weakly in the almost incompressible limit. On the other hand FETI-GenEO performs exactly as expected by finding the one eigenvector per subdomain which is responsible for slow convergence meaning that the automatic construction recovers the coarse space built analytically in [114, 43]. Making sure of this was absolutely crucial to Michelin since the tires which they make are in a large part rubber, the textbook example for an almost incompressible material. This final test as well as the others and the theoretical analysis in Chapter 5 are arguments toward the fact that FETI-GenEO is a Domain Decomposition method that is worth considering in an industrial code.

Chapter 7

Conclusions and Perspectives

Contents

7.1 Conclusions

Throughout this manuscript we have developed coarse spaces that are constructed automatically and lead to two level methods with guaranteed convergence rates. We have studied these coarse spaces both theoretically and numerically. Our main target, which was driven by the need to run industrial simulations, has been met: we are able to guarantee that even if the problem is very hard the solver will converge.

Of course there are still many ways to improve these methods. In the next three sections we describe three directions of research that seem promising. Each of them addresses one of the main concerns that is raised by GenEO:

- The coarse space may become very large. The solution we propose is to generalize GenEO to a multilevel method: if the coarse problem becomes too large for a direct solver then we apply GenEO to it. This would be the three level method and we may repeat the process recursively.
- The eigensolves, even though they can be performed in parallel, could we discouraging. We propose a FETI algorithm where the coarse space is built on the fly within the conjugate gradient iterations.
- Finally the evaluation of the method on real industrial cases is perhaps our highest priority.

What follows is still work in progress.

7.2 Multilevel Schwarz - GenEO

The ideas in this section are the result of many conversations with Frédéric Nataf, Clemens Pechstein and Robert Scheichl, most of the progress was made during my month long stay at the RICAM in Linz (Austria) for the Special Semester on Multiscale Simulation and Analysis in Energy and the Environment.

In the multigrid setting there are three main ingredients for each algorithm: the prolongation operator (how to navigate between levels), the smoother (what to do on each level) and an iterator which combines both of these contributions (additive, multiplicative, V-cycle, W-cycle, non linear AMLI iteration...).

In this section we describe the multilevel GenEO framework. More precisely we describe how to build the next coarse level. The following requires a graph partitioner (*e.g.* Metis [54]).

7.2.1 Multilevel setting

Notation

We use the subscript $\cdot_{[l]}$ to refer to the levels. The finest level (the mesh) is denoted by $\cdot_{[L]}$ while the coarsest level is denoted by $\cdot_{[0]}$. The indices corresponding to the subdomains within a level are the $^{(i)}$. Finally ^{k} denotes one of the basis functions for one level and one subdomain.

Required setting to build the coarser level

Because the GenEO algorithm was specifically defined to be as algebraic as possible all the information we need is the finite element information at level l:

- a basis $\{\phi^k_{[l]}\}_{1\leq k\leq n_{[l]}}$ which spans the solution space $V_{[l]}$,
- for each element $\tau \in \Omega_{[l]}$ the corresponding elementary matrix $A_{[l]_\tau}$ or the corresponding elementary bilinear forms $a_{[l]_\tau}$.

These can either be provided by the initial problem (on the finest grid) or be the results of a previous iteration during which level l was itself built from level $l+1$ using the process which is described below.

We denote by $A_{[l]} \in \mathbb{R}^{n_{[l]}\times n_{[l]}}$ the assembled global matrix problem: the entries of $A_{[l]}$ are $A_{[l]_{ij}} = \sum_\tau a_{[l]_\tau}(\phi^i_{[l]}, \phi^j_{[l]})$.

Three assumptions are required:

1. For any element τ, the elementary matrix $A_{[l]_\tau}$ is symmetric positive semi definite (spsd),
2. $A_{[l]}$ is symmetric positive definite (spd),
3. The basis $\{\phi^k_{[l]}\}_{1\leq k\leq n_{[l]}}$ verifies a unisolvence property on each element τ (the basis functions which are non zero on the element are linearly independent on the element).

Local setting

We use the graph partitioner to build a splitting of $\Omega_{[l]}$ into $N_{[l]}$ subdomains. Then we add a chosen number of layers to each of these subdomains returning an overlapping partition $\Omega_{[l]} = \bigcup_{i=1}^{N_{[l]}} \Omega^{(i)}_{[l]}$. We also define the partition of unity operators $\{\Xi^{(i)}_{[l]}\}_{1\leq i\leq N_{[l]}}$ following the algebraic definition. This is exactly as described in Chapter 4 for GenEO applied to additive Schwarz.

The local setting must verify three assumptions:

1. Each basis function $\phi^k_{[l]}$ is in the interior of at least one subdomain,
2. For each $i = 1, \ldots, N_{[l]}$, the bilinear form $a_{[l]_{\Omega^{(i),\circ}_{[l]}}}$ in the overlap is coercive on $\mathrm{Ker}(\Omega^{(i)}_{[l]})$.
3. For each $i = 1, \ldots, N_{[l]}$, the bilinear form $a_{\Omega^{(i)}_{[l]}}$ is coercive on $\mathrm{span}\{\phi^k_{[l]}; \mathrm{supp}(\phi^k_{[l]}) \subset \overline{\Omega_{[l]}}^{(i)}$ and $\mathrm{supp}(\phi^k_{[l]}) \not\subset (\Omega^{(i)}_{[l]} \setminus \Omega^{(i),\circ}_{[l]})\}$.

The first of these two assumptions is automatically satisfied if we add the layers of overlap following the procedure that is described in the GenEO chapter. Proving the last two would require more work. Another option is to replace the matrix on the right hand side of the generalized eigenvalue problem by the matrix of the problem restricted to the whole subdomain Ω_j reduced by one layer of elements. Then the assumptions are no longer required.

Coarse space construction and estimates

On each subdomain solve the generalized eigenvalue problem: find $\lambda^{(i),k}_{[l]} \in [0, +\infty]$ and $p^{(i),k}_{[l]} \in (V_{[l]})_{|\Omega_{[l]_i}}$ such that

$$a_{[l]_{\Omega^{(i)}_{[l]}}}(p^{(i),k}_{[l]}, v^{(i)}_{[l]}) = \lambda^{(i),k}_{[l]} a_{[l]_{\Omega^{(i),\circ}_{[l]}}}(\Xi^{(i)}_{[l]}(p^{(i),k}_{[l]}), \Xi^{(i)}_{[l]}(v^{(i)}_{[l]})), \text{ for all } v^{(i)}_{[l]} \in (V_{[l]})_{|\Omega_{[l]_i}}. \quad (7.1)$$

Then, given a threshold $K_{[l]}$ select the space of local contributions to V_H

$$V_H{}^{(i)}_{[l]} = \mathrm{span}(p^{(i),k}_{[l]}; \lambda^{(i),k}_{[l]} < K_{[l]}).$$

Let $\Pi^{(i)}_{[l]}$ be the projector onto this space of local contributions defined by

$$\Pi^{(i)}_{[l]} u := \sum_{\{k; \lambda^{(i),k}_{[l]} < K_{[l]}\}} a_{[l]_{\Omega^{(i),\circ}_{[l]}}}(\Xi^{(i)}_{[l]}(p^{(i),k}_{[l]}), \Xi^{(i)}_{[l]}(u)) p^{(i),k}_{[l]}.$$

The coarse space is the sum of the local contributions weighted by the partition of unity functions:

$$V^H_{[l]} = \sum_{i=1}^{N_{[l]}} R^{(i)\top}_{[l]} \Xi^{(i)}_{[l]}(V^H{}^{(i)}_{[l]}) = \{R^{(i)\top}_{[l]} \Xi^{(i)}_{[l]}(p^{(i),k}_{[l]}); \lambda^{(i),k}_{[l]} < K_{[l]}\}$$

and the prolongation operator is the rectangular matrix $R^{(0)\top}_{[l]}$ whose columns are the $R^{(i)\top}_{[l]} \Xi^{(i)}_{[l]}(p^{(i),k}_{[l]})$ that appear in the previous definition. It is then straightforward to define the coarse matrix $A^0_{[l]} = R^{(0)}_{[l]} A_{[l]} R^{(0)\top}_{[l]}$ as long as the vectors in $R^{(0)}$ are linearly independent.

Theorem 7.1. (GenEO stable splitting) The GenEO theory tells us that any $u_{[l]} \in V_{[l]}$ can be split into

$$u_{[l]} = \sum_{i=0}^{N_{[l]}} R^{(i)\top}_{[l]} u^{(i)}_{[l]},$$

where

$$u_{[l]}^{(0)} = \sum_{i=1}^{N_{[l]}} \Xi_{[l]}^{(i)}(\Pi_{[l]}^{(i)} u_{[l]}), \text{ and } u_{[l]}^{(i)} = \Xi_{[l]}^{(i)}(u_{[l]} - \Pi_{[l]}^{(i)} u_{[l]}). \tag{7.2}$$

The components satisfy the following estimates:

$$\sum_{i=1}^{N_{[l]}} \|u_{[l]}^{(i)}\|^2_{a_{[l]},\Omega_{[l]}^{(i)}} \leq k_{0[l]} \left(1 + \frac{1}{K_{[l]}}\right) \|u_{[l]}\|^2_{a_{[l]}}, \tag{7.3}$$

and

$$\|R_{[l]}^{(0)\top} u_{[l]}^{(0)}\|^2_{a_{[l]}} \leq \left[2 + 2\left(1 + \frac{1}{K_{[l]}}\right) k_{0[l]}^2\right] \|u_{[l]}\|^2_{a_{[l]}}. \tag{7.4}$$

where $k_{0[l]}$ is the coloring constant.

Initialize the next level

In order to be complete we must give the prolongation operator from level l to level $l+1$ and describe the initialization for adding the space for the next (coarser) level, making sure that the assumptions are verified.

- The basis functions are the low frequency eigenmodes:

$$\{\phi_{[l+1]}^{k'}\}_{k'=1,\ldots,\#V_{[l]}^H} = \{R_{[l]}^{(i)\top} \Xi_{[l]}^{(i)}(p_{[l]}^{(i),k}); \lambda_{[l]}^{(i),k} < K_{[l]};\, i = 1, \ldots, N_{[l]}\}.$$

  This means that we have chosen the next space to be $V_{[l+1]} = V_{[l]}^{(0)}$.
- The set of elements is the smallest (in the sense of inclusion) partition of $\Omega_{[l]}$ which resolves $\left\{\bigcap_{i\in\mathcal{I}} \Omega_{[l]}^{(i)}; \mathcal{I} \subset \{1, \ldots, n_{[l]}\}\right\}$.
- The elementary bilinear form for a coarse element $\tau \in \Omega_{[l+1]}$ is given by $a_{[l+1]_\tau} = \sum_{\tau'\in\tau} a_{[l]_{\tau'}} = a_{[l]_\tau}$.
- The projection operator used in the multilevel setting is:

$$P_{[l]} : V_{[l+1]} \to V_{[l]}, \quad P_{[l]} u_{[l+1]} = R_{[l]}^{0\top} u_{[l+1]}.$$

Assumptions that must be satisfied on each level

These assumptions were already stated in the previous paragraphs but I will sum them up here:

- Matrix $A_{[l+1]} = P_{[l]}^\top A_{[l]} P_{[l]}$ must be spd. A sufficient condition for this is that all the basis vectors for $V_{[l+1]}$ be linearly independent. The only problem which may occur is for basis functions which come from two different eigenvalue problems on two subdomains.
- Elementary matrices must be spsd.
- The basis functions must satisfy a unisolvence property on each element (all of the ones which are non zero on an element must be linearly independent on this element). This implies the first property.
- Having defined the subdomains, the coercivity
  - of bilinear form $a_{[l]_{\Omega_{[l]}^{(i),\circ}}}$ on $\mathrm{Ker}(\Omega_{[l]}^{(i)})$ for each subdomain,
  - and of bilinear form $a_{[l]_{\Omega_{[l]}^{(i)}}}$ on $\mathrm{span}\{\phi_{[l]}^k; \mathrm{supp}(\phi_{[l]}^k) \subset \bar{\Omega_{[l]}}^{(i)}$ and $\mathrm{supp}(\phi_{[l]}^k) \not\subset (\Omega_{[l]}^{(i)} \setminus \Omega_{[l]}^{(i),\circ})\}$ for each subdomain,

  must be ensured.

7.2.2 Fully additive multilevel preconditioner

We next define the fully additive preconditioner even though we are well aware that it is probably not the most optimal way to solve the problem. There is a small chance that the fact that all solves are done in parallel will be a sufficient argument to use this if there are only three levels.

Definition 7.2. Using the framework and notation introduced in the previous subsection, the L level additive multigrid preconditioner is:

$$\mathbf{B_{MG}^{add}}^{-1} = P_{[0]} \dots P_{[L-1]} A_{[L-1]}^{-1} P_{[L-1]}^{\top} \dots P_{[0]}^{\top} + \sum_{l=1}^{L-1} P_{[0]} \dots P_{[l-1]} \Lambda_{[l]}^{-1} P_{[l-1]}^{\top} \dots P_{[0]}^{\top} + \Lambda_{[0]}^{-1}, \tag{7.5}$$

where $A_{[L-1]}^{-1}$ is the exact coarse solve on the coarsest level and for $l = 0, \dots, L-1$, $\Lambda_{[l]}^{-1}$ are the smoothers. In our case we choose $\Lambda_{[l]}^{-1}$ to be the scaled one level Schwarz preconditioner for $A_{[l]}$,

$$\Lambda_{[l]}^{-1} = \frac{1}{\rho_{[l]}} \sum_{i=1}^{N_{[l]}} R_{[l]}^{(i)\top} A_{[l]}^{-1} R_{[l]}^{(i)}, \tag{7.6}$$

where $\rho_{[l]} > 0$ is the scaling constant.

In fact defining composite interpolation matrices

$$\bar{P}_{[l]} \, : \, V_{[l]} \mapsto V_{[0]}; \bar{P}_{[l]} := P_{[0]} \dots P_{[l-1]},$$

and

$$\bar{R}_{[l]}^{(i)} \, : \, V_{[0]} \mapsto V_{[l]}^{(i)}; \bar{R}_{[l]}^{(i)} := R_{[l]}^{(i)} \bar{P}_{[l]}^{\top},$$

the additive multigrid preconditioner can be rewritten as:

$$\mathbf{B_{MG}^{add}}^{-1} = \bar{P}_{[L]} A_{[L-1]}^{-1} \bar{P}_{[L]}^{\top} + \sum_{l=1}^{L-1} \bar{P}_{[l]} \Lambda_{[l]}^{-1} \bar{P}_{[l]}^{\top} + \Lambda_{[0]}^{-1}, \tag{7.7}$$

or

$$\mathbf{B_{MG}^{add}}^{-1} = \bar{P}_{[L]} A_{[L-1]}^{-1} \bar{P}_{[L]}^{\top} + \sum_{l=0}^{L-1} \sum_{i=1}^{N_{[k]}} \frac{1}{\rho_{[l]}} \left[\bar{R}_{[l]}^{(i)\top} {A_{[l]}^{(i)}}^{-1} \bar{R}_{[l]}^{(i)} \right]. \tag{7.8}$$

To be complete we define $\bar{P}_{[0]} u = u$ implying $\bar{R}_{[0]}^{(i)} = R_{[0]}^{(i)}$.

7.2.3 Convergence study when $\rho_{[l]} = 1$

If all the $\rho_{[l]} = 1$ this multilevel preconditioner is the classical One level Schwarz preconditioner applied to the original matrix $A_{[0]}$ and the set of subspaces

$$V = \bar{P}_{[L]} V_{[L]} + \sum_{l=1}^{L-1} \sum_{i=1}^{N_{[k]}} \bar{R}_{[l]}^{(i)\top} V_{[l]}^{(i)}.$$

This means that the abstract Schwarz theory for exact solvers applies and the cornerstone for a convergence proof is, once more, the existence of a stable splitting.

Lemma 7.3. For any $u \in V_{[0]}$, there exists a stable splitting onto

$$V = \bar{P}_{[L]} V_{[L]} + \sum_{l=1}^{L-1} \sum_{i=1}^{N_{[k]}} \bar{R}_{[l]}^{(i)\top} V_{[l]}^{(i)}.$$

The constant is

$$C_0^2 = \Pi_{m=0}^{L-1} \left[2 + 2 \left(1 + \frac{1}{K_{[m]}} \right) k_{0[m]}^2 \right] + \sum_{l=0}^{L-1} \left\{ k_{0[l]} \left(1 + \frac{1}{K_{[l]}} \right) \Pi_{m=0}^{l-1} \left[2 + 2 \left(1 + \frac{1}{K_{[m]}} \right) k_{0[m]}^2 \right] \right\} \tag{7.9}$$

Proof. For $u_{[0]} \in V_{[0]}$ we write

$$u = \bar{P}_{[L]} v_{[L]} + \sum_{l=1}^{L-1} \sum_{i=1}^{N_{[k]}} \bar{R}_{[l]}^{(i)\top} v_{[l]}^{(i)}.$$

On each level, the local component for level $l = 0, \ldots, L-1$ and subdomain $i = 1, \ldots, N_{[L]}$, is defined by

$$v_{[l]}^{(i)} = \Xi_{[l]}^{(i)} (u_{[l]} - \Pi_{[l]}^{(i)} u_{[l]}). \tag{7.10}$$

where, for $l = 1, \ldots, L-1$, these contributions are based on the coarse component $(u_{[l]})$ from the previous level:

$$u_{[l]} = v_{[l-1]}^{(0)} = \sum_{i=1}^{N_{[l-1]}} \Xi_{[l-1]}^{(i)} (\Pi_{[l-1]}^{(i)} u_{[l-1]}). \tag{7.11}$$

Finally, the component on the coarsest level is

$$v_{[L]} = \sum_{i=1}^{N_{[L-1]}} \Xi_{[L-1]}^{(i)} (\Pi_{[L-1]}^{(i)} u_{[L-1]}). \tag{7.12}$$

We recognize, with slightly modified notation, the splitting from (7.2) generalized to more levels. If all the assumptions for the GenEO framework are satisfied then so are estimates (7.4) and (7.3), namely

$$\|u_{[l+1]}\|_{a_{[l+1]}}^2 = \|P_l u_{[l+1]}\|_{a_{[l]}}^2 \leq \left[2 + 2 \left(1 + \frac{1}{K_{[l]}} \right) k_{0[l]}^2 \right] \|u_{[l]}\|_{a_{[l]}}^2, \tag{7.13}$$

and

$$\sum_{i=1}^{N_{[l]}} \|v_{[l]}^{(i)}\|_{a_{[l]}, \Omega_{[l]}^{(i)}}^2 \leq k_{0[l]} \left(1 + \frac{1}{K_{[l]}} \right) \|u_{[l]}\|_{a_{[l]}}^2. \tag{7.14}$$

Using the first of these estimates recursively from the finest level to level $l = 1, \ldots, L$ gives us

$$\|u_{[l]}\|_{a_{[l]}}^2 \leq \Pi_{m=0}^{l-1} \left[2 + 2 \left(1 + \frac{1}{K_{[m]}} \right) k_{0[m]}^2 \right] \|u_{[0]}\|_{a_{[0]}}^2, \tag{7.15}$$

We inject this into the other estimate in order to get a bound on the local components for $l = 0, \ldots, L-1$ with respect to the initial function

$$\sum_{i=1}^{N_{[l]}} \|v_{[l]}^{(i)}\|_{a_{[l]}, \Omega_{[l]}^{(i)}}^2 \leq k_{0[l]} \left(1 + \frac{1}{K_{[l]}} \right) \Pi_{m=0}^{l-1} \left[2 + 2 \left(1 + \frac{1}{K_{[m]}} \right) k_{0[m]}^2 \right] \|u_{[0]}\|_{a_{[0]}}^2. \tag{7.16}$$

Finally we add all the contributions from each level and get that the decomposition is stable for a constant

$$C_0^2 = \Pi_{m=0}^{L-1}\left[2+2\left(1+\frac{1}{K_{[m]}}\right)k_{0[m]}^2\right]+\sum_{l=0}^{L-1}\left\{k_{0[l]}\left(1+\frac{1}{K_{[l]}}\right)\Pi_{m=0}^{l-1}\left[2+2\left(1+\frac{1}{K_{[m]}}\right)k_{0[m]}^2\right]\right\} \tag{7.17}$$

which does not depend on any of the problem parameters or the number of subdomains ($k_{0[l]}$ is the coloring constant at level l and $K_{[l]}$ is the criterion which we've chosen to select eigenvectors). □

7.2.4 The constant in Lemma 7.3

Introducing the largest coloring constant $k_0 = \max_{0\le l\le L-1} k_{0[l]}$ and the largest cut off term $C = \max_{0\le l\le L-1}\left[\max_{1\le i\le N_{[l]}}\left(1+\frac{1}{K_{[l]}}\right)\right]$, a more simple stability constant $C_0'^2$ for the splitting is

$$\begin{aligned} C_0'^2 =& 2^L(1+k_0^2\,C)^L + \sum_{l=0}^{L-1}\left[k_0 C 2^l\,(1+k_0^2 C)^l\right] \\ =& 2^L(1+k_0^2\,C)^L + k_0 C\sum_{l=0}^{L-1}\left[2^l\,(1+k_0^2 C)^l\right] \\ =& 2^L(1+k_0^2\,C)^L + k_0\,C\left[\frac{1-\left(2(1+k_0^2 C)\right)^L}{1-\left(2(1+k_0^2 C)\right)}\right]. \end{aligned} \tag{7.18}$$

It seems pretty reasonable to replace the cut off constants $K_{[l]}$ by one common bound for all levels and all subspaces. In view of our experience, an acceptable choice would be $K_{[l]} = 0.5$ for all levels l , in which case $C = 3$.
As for k_0, if regular decompositions onto regular meshes are considered and subdomains are aggregated in a natural way at each level, it should be pretty much constant over the levels. Even using Metis for the initial partitioning into subdomains and the aggregating from one level to the next, it seems reasonable enough to use a common value k_0 in the approximation.

The next simplification is a little bit more of a stretch but it simplifies the expression some more. What we mean by this is that the decomposition is C_0-stable with all the C_0, C_0', C_0'' that we introduce but the bound is getting less sharp each time we simplify the expression. Using the fact that $k_0\,C > 1$ and starting from the second last line of the last calculation we get that a suitable constant is

$$\begin{aligned} C_0''^2 =& k_0\,C\sum_{l=0}^{L}\left[2(1+k_0^2 C)\right]^l \\ =& k_0\,C\left[\frac{1-\left(2(1+k_0^2 C)\right)^{L+1}}{1-\left(2(1+k_0^2 C)\right)}\right]. \end{aligned} \tag{7.19}$$

Suppose that $k_0 = 4$ and $C = 3$ which corresponds to $K = 0.5$ (these are pretty optimistic values) then we get the following values for $C_0'^2$ and $C_0''^2$:

level	1	2	3	4	5
$C_0'^2$	110	1.1×10^4	1.1×10^6	1.0×10^8	1.0×10^{10}
$C_0''^2$	1.2×10^3	1.2×10^5	1.1×10^7	1.1×10^9	1.1×10^{11}

For 'one level' these formula apply because it is one level as well as the fine level. In fact for $L = 1$ in the formula for C_0^2 we find the result that is in our GenEO paper $C_0^2 = k_0 \left(1 + \frac{1}{K}\right) [2k_0 + 1] + 2$.

Theorem 7.4 (Condition number for the fully additive multilevel GenEO)**.** The condition number for the l-level GenEO method with the coarse space based on generalized eigenvalue problems as defined in this paper can be bounded by

$$\kappa(\mathbf{B_{MG}^{add}}^{-1}\mathbf{A}) \ \leq \ (1 + \sum_{l=0}^{L-1} k_{0[l]})\, C_0^2,$$

where C_0^2 depends only on the number of levels $L + 1$, the coloring constant $k_{0[l]}$ at each level and the cut-off parameter $K_{[l]}$ for level l and is given by (7.9).

7.2.5 Fully additive multigrid with scaled smoother

Lets assume now that the smoother we use on each level is the one level Schwarz preconditioner with a damping by a factor $\rho_{[l]}$ from equation (7.6). That is

$$\Lambda_{[l]}^{-1} = \frac{1}{\rho_{[l]}} \sum_{i=1}^{N_{[l]}} R_{[l]}^{(i)\top} A_{[l]}^{-1} R_{[l]}^{(i)},$$

Then the fully additive multigrid preconditioner also fits into the abstract Schwarz theory but with inexact solvers. The stable splitting uses the norm induced by $\Lambda_{[l]}$. In particular if $\rho_{[0]} = 1$ and on each level l, for $1 \leq l \leq L$, the damping factor $\rho_{[l]}$ is chosen such that

$$\frac{1}{\rho_{[l]}} < \frac{1}{\rho_{[l-1]}} \left[2 + 2\left(1 + \frac{1}{K_{[l-1]}}\right) k_{0[l-1]}\right], \tag{7.20}$$

then the splitting which we have already introduced is stable with a constant

$$C_0^2 = 1 + \sum_{l=0}^{L-1} \left\{ k_{0[l]} \left(1 + \frac{1}{K_{[l]}}\right) \right\}. \tag{7.21}$$

This relies on the fact that (7.13) can be rewritten for the norm implied by the damped smoother $\Lambda_{[l]}$

$$\frac{1}{\rho_{[l]}} \|u_{[l+1]}\|_{\Lambda_{[l+1]}}^2 = \|P_l u_{[l+1]}\|_{\Lambda_{[l]}}^2 \leq \left[2 + 2\left(1 + \frac{1}{K_{[l]}}\right) k_{0[l]}^2\right] \|u_{[l]}\|_{\Lambda_{[l]}}^2, \tag{7.22}$$

so

$$\|u_{[l+1]}\|_{\Lambda_{[l]}}^2 < \|u_{[l]}\|_{\Lambda_{[l]}}^2.$$

This damping however has a dramatic bound on the upper bound which measures whether or not the inexact solver is a good approximation of the exact one. As expected damping really doesn't have much effect for the additive version of the multigrid preconditioner. However, for the multiplicative version it would be very efficient. The theoretical analysis of this, more complex, multilevel method is left for future work. It should rely on the results in Subsection 4.3.3 for the two level Hybrid Schwarz preconditioner. We expect that based on this we could recover better convergence estimates that the ones in other multilevel extensions of GenEO type methods [27, 116].

7.3 On the fly construction of the coarse space

7.3.1 Motivations

With FETI GenEO (Chapter 5), the overhead cost of solving the eigenproblems may be discouraging. For this reason we propose a two level FETI method where the basis vectors for the second level are selected on the fly within the conjugate gradient (CG) iterations. Thanks to this adaptive process we bypass the eigensolves meaning that the preprocessing step is expected to be a lot cheaper. The adaptive process we propose is different but inspired by the Adaptive Algebraic Multigrid [12].

The idea is to take full advantage of all the information which we compute or, in other words, to be frugal with our computational resources. In [91] it was pointed out that within each application of the FETI preconditioner a local problem is solved per subdomain but then all the different contributions are averaged and some valuable information may be lost. The FETI-S algorithm is proposed there where the next approximate solution is optimized over all the local search directions (before the averaging process). Here we build on the same initial statement since the crucial point is to operate the so-called τ-test (line 9 in Algorithm 7.2) on each local contribution to the preconditioner in order to evaluate whether it should be averaged or not. If it is deemed crucial then it is used as a basis vector for the coarse space.

We have not yet managed to write a full proof of convergence. One direction which we have been looking in is to use the Ritz vectors and values. Indeed it is well known that the Ritz values are approximations for the eigenvalues and in turn the eigenvalues govern the convergence of the conjugate gradient method. The tight relationship between the convergence of CG and Ritz values is a prolific direction of research (see [113, 46, 49]). Lemmas 7.6 and 7.5 provide us with ways to find bounds for the extremal Ritz values, more precisely the condition in Lemma 7.5 motivates the coarse space selection in our new algorithm.

7.3.2 The algorithm

The Frugal FETI algorithm for solving (2.15) preconditioned by M^{-1} is introduced in Algorithm 7.2 using two simple routines defined in Algorithm 7.1 and the following additional notation: G_F is the basis for the coarse space, P_F is the coarse projector, $\tau_{user} > 0$ is a threshold chosen by the user, k is an integer used to count iterations. For simplicity we have written it in the case where the S_i are non singular but the other case is not a problem either. Next we make two remarks meant to help understand Frugal FETI.

- If at each iteration and for each subdomain $\tau_s \geq \tau_{user}$, *i.e.* the test in line 13 of Algorithm 7.2 always succeeds, then Frugal FETI is the usual FETI algorithm.
- When the τ-test (line 13 of Algorithm 7.2) fails, we update the coarse space and restart the CG (line 16 in Algorithm 7.2). The intializeCG routine includes the computation of the new coarse projector. The coarse space is updated until the τ-test succeeds in all subdomains.

7.3.3 Some of our ideas for the proof

We are currently looking for a full proof of convergence for this algorithm. We next describe a few of the ideas that we have had. Let $\theta_j^{(m)}$ for $j = 1, \ldots, m$ be the Ritz values at iteration m of the conjugate gradient algorithm. A well known result is that the (sharp)

Algorithm 7.1 Two routines for Frugal FETI

Routine: `initializeCG`

Input λ_k, G_F
$P_F = I - G_F(G_F^\top F G_F)^\dagger G_F^\top F$
$\lambda_k = \lambda_k + G_F \left(G_F^\top F G_F\right)^\dagger G_F^\top (d - F\lambda_k)$
$r_k = d - F\lambda_k$
$z_k = M^{-1} r_k$
$p_k = z_k$
Return $\lambda_k, r_k, z_k, p_k, P_F$

Routine: `iterateCG`

Input $\lambda_k, r_k, z_k, p_k, P_F$
$p_k = P_F p_k$
$\alpha_k = \frac{\langle r_k, z_k \rangle}{\langle F p_k, p_k \rangle}$
$\lambda_{k+1} = \lambda_k + \alpha_k p_k$
$r_{k+1} = r_k - \alpha_k P_F^\top F p_k$
$z_{k+1} = M^{-1} r_{k+1}$
$\beta_k = \frac{\langle r_{k+1}, z_{k+1} \rangle}{\langle r_k, z_k \rangle}$
$p_{k+1} = z_{k+1} + \beta_k p_k$
Return $\lambda_{k+1}, r_{k+1}, z_{k+1}, p_{k+1}$

Algorithm 7.2 Frugal FETI algorithm for solving $F\lambda = d$ preconditioned by M^{-1}

Input λ_0
$G_F = [\ \ ]$ ▷ start with empty coarse space
$[\lambda_0, r_0, z_0, p_0, P_F] =$ `initializeCG`(λ_0, G_F)
$k = 0$
while convergence not achieved **do** ▷ CG outer loop
  while true **do** ▷ τ-test: loop until success
    for $s = 1, \ldots, N$ **do**

$$\tau_s = \frac{\sum_{\{t; B_t B_s^\top \neq 0\}} \langle (M^{-1})_t r_k, r_k \rangle}{\langle F_s z_k, z_k \rangle}$$

▷ where $\begin{cases} M^{-1} = \sum_{t=1}^{N} (M^{-1})_t \text{ and} \\ F = \sum_{s=1}^{N} \underbrace{B_s S_s^\dagger B_s^\top}_{:=F_s} \end{cases}$

      if $\tau_s < \tau_{user}$ **then**
        $G_F = [G_F \,|\, M^{-1} F_s z_k]$ ▷ add column to coarse basis by concatenation
      end if
    end for
    if $\min_{s \in [1;N]} \tau_s \geq \tau_{user}$ **then**
      go back to outer loop line 19 ▷ success: carry on with CG
    else
      $[\lambda_k, r_k, z_k, p_k, P_F] =$ `initializeCG`(λ_k, G_F) ▷ failure: need to reinitialize
    end if
  end while
  $[\lambda_{k+1}, r_{k+1}, z_{k+1}, p_{k+1}] =$ `iterateCG`$(\lambda_k, r_k, z_k, p_k, P_F)$
  test for convergence
  $k = k + 1$
end while
Return λ_k

lower bound for the spectrum of the preconditioned FETI operator is $\lambda_{min} = 1$ [58] and the, by now classical, result in Lemma 7.5 ensures that $\theta_{min}^{(m)} \geq 1$ also.

Lemma 7.5. At any given iteration m, the Ritz values are the eigenvalues of the section of the original matrix A to the m-th Krylov subspace $\mathcal{K}_m$ for the l_2 projection so $\theta_{min}^{(m)} \geq \lambda_{min}$ where λ_{min} is the smallest eigenvalue of A.

Finding a bound for the largest of the Ritz values is a lot trickier because there is no simple upper bound for the spectrum of the preconditioned FETI operator.

Lemma 7.6. Let $m \in \mathbb{N}$. If there exist m vectors $u_1, \ldots, u_m \in \mathcal{K}_m$ which are orthogonal in the l_2 (Euclidean) inner product and which satisfy $\langle Au_j, u_j \rangle \leq C \langle u_j, u_j \rangle$, $\forall\, j = 1, \ldots, m$, for some constant $C > 0$ then $\theta_{max}^{(m)} \leq mC$.

The successive conjugate gradient residuals are l_2 orthogonal so these are good test vectors for Lemma 7.6. In fact the whole purpose of the τ-test is to ensure that the lemma applies with constant $\mathcal{N}/\tau_{user}$.

It remains to prove that the convergence of the conjugate gradient algorithm is driven by the Ritz values. In fact we are not sure that this is true. For GMRES applied to diagonalizable matrices we have derived such a result so maybe we should build the adaptive algorithm within GMRES. The theorem uses the following proposition where the Krylov subspace is $\mathcal{K}_m := \text{span}(r_0, Ar_0, \ldots, A^{m-1}r_0)$.

Lemma 7.7 (Part of Proposition 6.3 in Saad's book)**.** Let Q_m be any projector onto $\mathcal{K}_m$ and let A_m be the section of A to $\mathcal{K}_m$; that is, $A_m = Q_m A_{|\mathcal{K}_m}$. Then, for any polynomial q of degree not exceeding $m - 1$

$$q(A)v = q(A_m)v.$$

Proof. See the proof of Proposition 6.3 in [95]. □

Next we recall the definition of the Ritz values and Ritz vectors of A following the presentation given in [113]. A particular choice of projection operator onto $\mathcal{K}_m$ is the l_2 (Euclidean)-orthogonal projection upon $\mathcal{K}_m$ which we denote by Π_m. For any $m \leq n$ the Ritz values $\theta_1^{(m)}, \ldots, \theta_m^{(m)}$ of A with respect to $\mathcal{K}_m$ are defined as the eigenvalues of the mapping $A_m = \Pi_m A_{|\mathcal{K}_m}$. The eigenvectors corresponding to the $\theta_i^{(m)}$ are the Ritz vectors and we introduce the notation $y_i^{(m)}$, $i = 1, \ldots, m$, for them. The whole set of Ritz vectors and Ritz values changes at each iteration.

Theorem 7.8 (Convergence rate of GMRES based on the Arnoldi matrix)**.** Let r_m be the residual associated with the approximate solution x_m obtained at the m-th step of the GMRES algorithm and r_0 be the residual associated with the initial guess x_0. In the case where A is diagonalizable the following convergence bound holds at iteration m:

$$\frac{\|r_m\|}{\|r_0\|} \leq \kappa_2(X^{(m+1)}) \frac{C_m\left(\frac{a^{(m+1)}}{d^{(m+1)}}\right)}{\left|C_m\left(\frac{c^{(m+1)}}{d^{(m+1)}}\right)\right|}, \tag{7.23}$$

in which C_m is the Chebyshev polynomial of degree m of the first kind, $A^{(m+1)} = X^{(m+1)}\Lambda^{(m+1)}{X^{(m+1)}}^{-1}$, $\Lambda^{(m+1)}$ is a diagonal matrix whose entries are the Ritz values $\theta_i^{(m+1)}$, these eigenvalues are in the Ellipse$(c^{(m+1)}, d^{(m+1)}, a^{(m+1)})$ which excludes the origin and $\kappa_2(X^{(m+1)}) = \|X^{(m+1)}\|\|{X^{(m+1)}}^{-1}\|$.

Proof. This proof is inspired by the proof of convergence of GMRES in [95].

First we recall some features of the GMRES algorithm. Let $r_0 := b - Ax_0$ be the initial residual. The standard GMRES method for solving $Ax_* = b$ generates a sequence $x_1, x_2, ...$ with the following characterizing property:

$$x_m \in x_0 + \mathcal{K}_m, \text{ and } \|r_m\| := \|Ax_m - b\| = \min_{\{x \in x_0 + \mathcal{K}_m\}} \|Ax - b\|. \tag{7.24}$$

If for any $M \in \mathbb{N}$ the set of polynomials of degree at most M is denoted by $\mathbb{P}^M$ then

$$\begin{aligned} x \in x_0 + \mathcal{K}_m &\Leftrightarrow x = x_0 + q(A)r_0 \quad \text{with } q \in \mathbb{P}^{m-1} \\ &\Leftrightarrow x - x_0 = r(A)(x_* - x_0) \quad \text{with } r \in \mathbb{P}^m \text{ and } r(0) = 1. \end{aligned}$$

With this (7.24) implies

$$\|r_m\| = \min_{\{p \in \mathbb{P}^m;\, p(0)=1\}} \|p(A)(r_0)\|. \tag{7.25}$$

By definition of the Krylov subspace, $r_0 \in \mathcal{K}^{m+1}$. Moreover p is a polynomial of degree at most m so we may apply the result from Lemma 7.7 for $Q_{m+1} = \Pi_{m+1}$ (the l_2-orthogonal projection onto $\mathcal{K}^{m+1}$):

$$p(A)r_0 = p(A_{m+1})r_0.$$

Now the norm of the residual at iteration m can be rewritten without any occurrence of A:

$$\|r_m\| = \min_{\{p \in \mathbb{P}^m;\, p(0)=1\}} \|p(A_{m+1})(r_0)\| \tag{7.26}$$

and we can prove both convergence estimates.

Lets assume that A is diagonalizable then for any m, A_{m+1} is diagonalizable because $A = V_{m+1}A_{m+1}V_{m+1}^\top$ where V_{m+1} is the matrix whose columns are the orthonormal basis vectors for $\mathcal{K}_{m+1}$ generated by GMRES. We may write

$$A_{m+1} = X^{(m+1)}\Lambda^{(m+1)}\left(X^{(m+1)}\right)^{-1}$$

where $\Lambda^{(m+1)}$ is a diagonal matrix which entries are the eigenvalues of A_{m+1}. These are also the Ritz values of A at iteration $m+1$ and which we again denote with $\theta_i^{(m+1)}$ for $i = 1, \ldots, m+1$.

Now (7.26) can be rewritten as

$$\begin{aligned} \|r_m\| &= \min_{\{p \in \mathbb{P}^m;\, p(0)=1\}} \|p(A_{m+1})(r_0)\| \\ &= \min_{\{p \in \mathbb{P}^m;\, p(0)=1\}} \|X^{(m+1)}p(\Lambda_{m+1}){X^{(m+1)}}^{-1}(r_0)\| \\ &\leq \min_{\{p \in \mathbb{P}^m;\, p(0)=1\}} \|X^{(m+1)}\|\|{X^{(m+1)}}^{-1}\|\|p(\Lambda^{(m+1)})\|\|r_0\| \\ &\leq \underbrace{\|X^{(m+1)}\|\|{X^{(m+1)}}^{-1}\|}_{\kappa_2\left(X^{(m+1)}\right)} \|r_0\| \min_{\{p \in \mathbb{P}^m;\, p(0)=1\}} \max_{i=1,\ldots,m+1} |p(\theta_i^{(m+1)})| \end{aligned}$$

In the last line we used the fact that Λ is a diagonal matrix. The proof of (7.23) is ended by applying a result from approximation theory (see for instance [95] Corollary 6.33). □

7.4 An industrial problem

As part of the collaboration with Michelin, the work for this thesis included spending several months at the technology center in Clermont Ferrand (Auvergne, France) to implement some of the ideas that were developed to make domain decomposition more robust. Since the FETI method was already implemented, and is more suitable for incompressible problems, it was natural to choose to implement either FETI GenEO (Chapter 5) or the adaptive Frugal FETI algorithm from the previous section. For reasons related to implementation techniques it was decided to test the Frugal FETI algorithm. It requires no eigensolver. Another nice feature of the Frugal FETI algorithm is that if no vectors are selected for the coarse space then the additional cost compared to the usual FETI is very low.

Of course the simulations which are run at Michelin are a lot more complex than any of the examples that we considered in the previous chapters. One of the main differences is that Michelin solves non linear systems. The only part of the code which we modified is the linear solver which is applied at each step of Newton's method. A very natural and probably very efficient idea would be to reuse information from one iteration of Newton's method to the next. In particular the coarse space, or parts of it, could be reused in a way inspired by [46, 49].

The numerical results which we present correspond to the simulation of a whole tire, the geometry of which is illustrated in Figure 7.1. The displacement is caused by an obstacle and the mesh is most refined around this indent. Away from this zone of interest the mesh is rather coarse which explains the polygonal shape of the model. The domain is partitioned into four subdomains and we compare the results for Frugal FETI (see previous subsection) with a threshold $\tau = 0.1$ and the 'usual' FETI. The Frugal FETI algorithm was implemented during this thesis and when we mention FETI we mean that no modification was made to the FETI code already available at Michelin.

In Table 7.1 we compare the performance of the two methods in terms of number of Newton steps, total number of iterations (adding together the number of iterations at each Newton step) and computation time. We notice that with Frugal FETI, seven Newton steps are necessary instead of four with FETI. In our opinion there is no particular reason for this behaviour and it can be explained by the fact that the stopping criterion for each linear solve is quite complicated and has been slightly modified in Frugal FETI. There is a regularization step in the Michelin FETI code and another possible explanation is that it has not been well incorporated into Frugal FETI. Once more we remind the reader that this is still work in progress and that we do not consider out current implementation to be at a stage advanced enough to draw any final conclusions.

	Time (seconds)	Nb. of Newton steps	Total nb. of iterations
FETI	198	4	856
Frugal FETI	452	7	388

Table 7.1: Comparison of Frugal FETI and FETI on the test case illustrated in Figure 7.1.

In Table 7.2 we compare the behaviour of the linear solvers more precisely by giving for each Newton step the number of iterations needed to reach convergence, the final primal residual and, for Frugal FETI, the number of times that the coarse space was updated. On average the coarse space was updated 61 times. This is very small compared to the size of the FETI operator (14577).

FETI

Newton step #	Nb. iterations	Final Residual
1	202	$8.3 \cdot 10^{-8}$
2	207	$9.7 \cdot 10^{-7}$
3	223	$1.6 \cdot 10^{-7}$
4	224	$7.3 \cdot 10^{-11}$

Frugal FETI

Newton step #	Nb. iterations	Final Residual	Nb. of updates
1	36	$1.4 \cdot 10^{-7}$	61
2	110	$1.6 \cdot 10^{-6}$	45
3	54	$4.9 \cdot 10^{-7}$	59
4	40	$4.7 \cdot 10^{-8}$	64
5	55	$1.7 \cdot 10^{-7}$	59
6	32	$2.4 \cdot 10^{-6}$	75
7	61	$2.7 \cdot 10^{-6}$	61

Table 7.2: Comparison of the Newton steps for Frugal FETI and FETI on the test case illustrated in Figure 7.1.

It appears that the number of iterations needed to solve one linear system is indeed significantly reduced with Frugal FETI. In fact, even if seven Newton steps are needed instead of four, the total number of iterations is still smaller with Frugal FETI (388 versus 856). For this reason we are optimistic about future results and writing a bug free and tuned code is one of our top priorities.

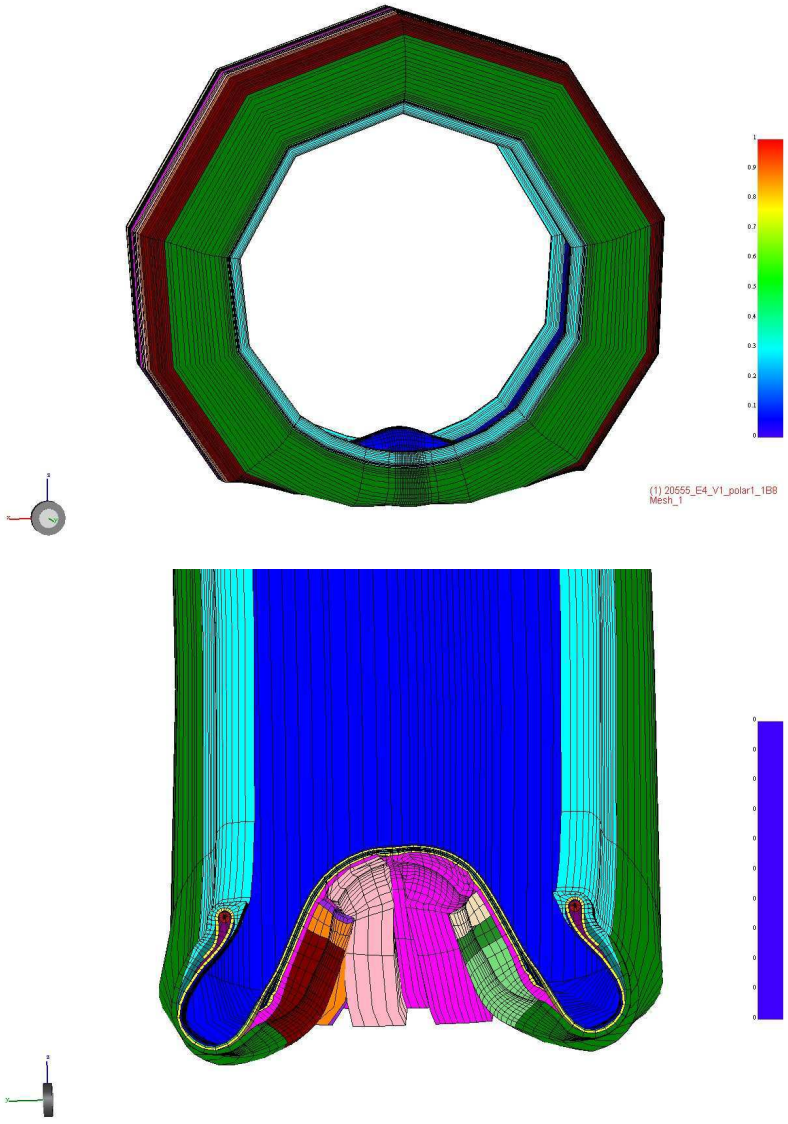


Figure 7.1: Michelin test case. The different colors correspond to different materials. The rank of the FETI operator is 14577.

Chapter 8

Appendix: two posters

On the next two pages are two posters that were presented at international conferences:
- *Special Semester on Multiscale Simulation and Analysis in Energy and the Environment* in Linz, Austria (November 2011),
- *22nd International Conference on Domain Decomposition Methods* in Lugano, Switzerland (September 2013).

A Robust Coarse Space for Systems of PDEs with high heterogeneities: GenEO

Nicole Spillane

Université Pierre et Marie Curie, Paris, FRANCE
Joint work with: V. Dolean, P. Hauret, F. Nataf, C. Pechstein, R. Scheichl

Abstract

Coarse spaces are instrumental in obtaining scalability for domain decomposition methods for partial differential equations (PDEs). However, it is known that most popular choices of coarse spaces perform rather weakly in the presence of heterogeneities in the PDE coefficients, especially for systems of PDEs. Here, we introduce in a variational setting a new coarse space that is robust even when there are such heterogeneities. We achieve this by solving local generalized eigenvalue problems in the overlaps of subdomains that isolate the terms responsible for slow convergence. We have proved a general theoretical result that rigorously establishes the robustness of the new coarse space and we give some numerical examples on two and three dimensional heterogeneous PDEs and systems of PDEs that confirm this.

Problems we solve

Let V^h be a finite element space of functions in Ω based on a mesh $\mathcal{T}^h = \{\tau\}$ of domain Ω.
Given $f \in (V^h)^*$ find $u \in V^h$

$$a(u,\, v) \;=\; \langle f,\, v\rangle \qquad \forall v \in V^h$$
$$\iff \qquad \mathbf{A}\,\mathbf{u} \;=\; \mathbf{f}$$

Assumptions:

1. $\mathbf{A}$ *symmetric positive definite*
2. $\mathbf{A}$ is given as a set of **element stiffness matrices** + connectivity (list of DOF per element)

   and verifies the assembling property:

   $$a(v,\, w) \;=\; \sum_{\tau} a_\tau(v_{|\tau},\, w_{|\tau})$$

   where $a_\tau(\cdot,\cdot)$ *symmetric positive semi-definite*
3. The finite element basis $\{\phi_k\}_{k=1}^n$ of V^h verifies a *unisolvence* property on each element τ.
4. Two more technical assumptions on $a(\cdot,\cdot)$ later!

Examples:

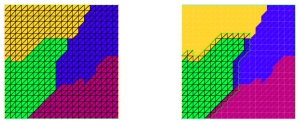
- Darcy $\quad a(u,\, v) = \int_\Omega \boldsymbol{\kappa} \nabla u \cdot \nabla v \, dx$
- Elasticity $\quad a(\boldsymbol{u},\, \boldsymbol{v}) = \int_\Omega \boldsymbol{C}\, \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx$
- Eddy current
  $a(\boldsymbol{u},\, \boldsymbol{v}) = \int_\Omega \boldsymbol{\nu}\, \mathrm{curl}\, \boldsymbol{u} \cdot \mathrm{curl}\, \boldsymbol{v} + \boldsymbol{\sigma}\, \boldsymbol{u} \cdot \boldsymbol{v}\, dx$

with heterogeneities, high contrast in parameters

General Setting: Additive Schwarz

The following is done using only the connectivity information and a graph partitioner such as Metis.

1. Build a non overlapping partition of Ω.
2. Add one layer of elements to each subdomain $j = 1, \ldots, N$ to get a partition into overlapping subdomains Ω_j.



Adding one layer of overlap to the green subdomain.

3. Define the local finite element spaces: $V_j := \mathrm{span}\{\phi_k : \mathrm{supp}(\phi_k) \subset \Omega_j\}$. Then denote by $\mathbf{R}_j^\top : V_j \to V^h$ the natural local/global embedding and by $a_{\Omega_j}(u, v) := \sum_{\tau \subset \Omega_j} a_\tau(u_{|\tau}, v_{|\tau})$ the local bilinear form.
4. Define a coarse space V_H and denote by $\mathbf{R}_H^\top : V_H \to V^h$ the natural coarse/global embedding.

Two level additive Schwarz

$$\mathbf{M}_{AS,2}^{-1} \;:=\; \mathbf{R}_H^\top \mathbf{A}_H^{-1} \mathbf{R}_H + \sum_{j=1}^N \mathbf{R}_j^\top \mathbf{A}_j^{-1} \mathbf{R}_j$$

where $\quad \mathbf{A}_j = \mathbf{R}_j^\top \mathbf{A} \mathbf{R}_j \quad$ and $\quad \mathbf{A}_H = \mathbf{R}_H^\top \mathbf{A} \mathbf{R}_H$.

If we prove the existence of a C_0-stable decomposition (as defined next) for each $v \in V_h$ then the general Schwarz theory tells us that the condition number of the preconditioned operator is bounded by

$$\kappa(\mathbf{M}_{AS,2}^{-1}\mathbf{A}) \le C_0^2 (k_0 + 1),$$

where each point belongs to at most k_0 subdomains.

Definition (C_0-Stable decomposition)

Given a coarse space $V_H \subset V_h$, local subspaces $\{V_j\}_{1 \le j \le N}$ and a constant C_0, a C_0-stable decomposition of $v \in V_h$ is a family of functions,

$$(v_H, v_1, \ldots, v_N) \in V_H \times V_1 \times \ldots \times V_N,$$

which satisfies

$$v = v_H + \sum_{j=1}^N v_j,$$

and

$$a(v_H, v_H) + \sum_{j=1}^N a_{\Omega_j}(v_j, v_j) \le C_0^2 a(v, v).$$

A sufficient condition for this last inequality is: there exists a constant C_1 such that

$$a_{\Omega_j}(v_j, v_j) \le C_1\, a_{\Omega_j}(v_{|\Omega_j}, v_{|\Omega_j}) \text{ for all } j = 1, \ldots, N. \quad (1)$$

Then the decomposition is C_0-stable with

$$C_0^2 = 2 + C_1 k_0 (2k_0 + 1).$$

Objective: define the coarse space in such a way that there exists a decomposition of any $v \in V^h$ which fulfills (1) for a C_1 which is independant of the heterogeneities and the decomposition. Then the bound on the condition number and hence on the convergence rate will also be independant of these quantities leading to a robust method.
In order to do this we need to introduce partition of unity operators which will allow us to define the coarse space and the local components.

Definition ('Discrete' partition of unity)

For any $j = 1, \ldots N$, let

$$dof(\Omega_j) \;:=\; \{k : \mathrm{supp}(\phi_k) \cap \Omega_j \neq \emptyset\}$$

denote the space of all degrees of freedom in Ω_j, and

$$idof(\Omega_j) \;:=\; \{k : \mathrm{supp}(\phi_k) \subset \overline{\Omega_j}\}$$

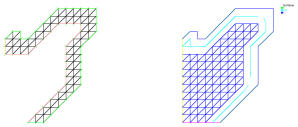
denote the space of internal degrees of freedom in Ω_j.
Notice that: $(V^h)_{|\Omega_j} = \mathrm{span}\{\phi_k\}_{k \in dof(\Omega_j)} \not\subset V^h$.
and $V_j = \mathrm{span}\{\phi_k\}_{k \in idof(\Omega_j)} \subset V^h$.
Then for any $v = \sum_{k=1}^n v_k \phi_k \in V^h$ define the partition of unity operator as:

$$\Xi_j(v) \;:=\; \sum_{k \in idof(\Omega_j)} \frac{1}{\#\{j : k \in idof(\Omega_j)\}} v_k\, \phi_k \in V_j.$$

It is indeed a partition of unity: $\sum_{j=1}^N \Xi_j v = v$.

Definition (Ω_j°)

Let Ω_j° denote the part of Ω_j that is overlapped (left), then $(\Xi_j v)_{|\Omega_j \setminus \Omega_j^\circ} = v_{\Omega_j \setminus \Omega_j^\circ}$ (right).



Finally define, $a_{\Omega_j^\circ}(v, v) = \sum_{\tau \subset \Omega_j^\circ} a_\tau(v_{|\tau}, v_{|\tau})$.

Theorem: GenEO Coarse Space and convergence result

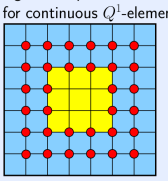
On each subdomain Ω_j, $j = 1 \ldots N$, find $p_{j,k} \in V_{h|\Omega_j}$ and $\lambda_{j,k} \ge 0$:

$$a_{\Omega_j}(p_{j,k},\, v) \;=\; \lambda_{j,k}\, a_{\Omega_j^\circ}(\Xi_j p_{j,k},\, \Xi_j v) \qquad \forall v \in V_{h|\Omega_j}$$
$$\Leftrightarrow \mathbf{A}_j \mathbf{p}_{j,k} \;=\; \lambda_{j,k}\, \mathbf{X}_j \mathbf{A}_j^\circ \mathbf{X}_j\, \mathbf{p}_{j,k} \qquad (\mathbf{X}_j \ldots \text{diagonal})$$

Select the first $m_j := \min\left\{m : \lambda_{m+1}^j > \frac{\delta_j}{H_j}\right\}$ ($H_j \ldots$ subdomain diameter, δ_j overlap width), eigenvectors per subdomain and define the coarse space as $V_H = \mathrm{span}\{\Xi_j p_{j,k}\}_{k=1,\ldots,m_j}^{j=1,\ldots,N}$. Then the condition number of the preconditioned operator is bounded by:

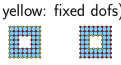
$$\kappa(\mathbf{M}_{AS,2}^{-1}\mathbf{A}) \;\le\; (1 + k_0)\left[2 + k_0\,(2k_0 + 1)\, \max_{j=1}^N\left(1 + \frac{H_j}{\delta_j}\right)\right]$$

DOFs that are free in the eigenvalue problem for continuous Q^1-elements



Both matrices typically singular $\implies \lambda_{j,k} \in [0, \infty]$
The proof requires two technical assumptions.
Assumption 1: $a_{\Omega_j}(\cdot,\cdot)$ SPD on $\mathrm{span}\{\phi_{k|\Omega_j}\}_{k \in dof(\Omega_j) \setminus idof(\Omega_j)}$
Assumption 2: $a_{\Omega_j^\circ}(\cdot,\cdot)$ SPD on $\mathrm{span}\{\phi_{k|\Omega_j^\circ}\}_{k \in idof(\Omega_j) \setminus idof(\Omega_j \setminus \Omega_j^\circ)}$

Assumptions 1 and 2 hold if certain mixed "boundary" value problems are solvable:
(red: free dofs, yellow: fixed dofs)



Stable decomposition

Coarse component: $v_H = \sum_{j=1}^N \Xi_j \Pi_j v_{|\Omega_j} \in V_H$, and **Local components:** $v_j = \Xi_j (v - \Pi_j v) \in V_j$,
where Π_j is the local projector onto $\mathrm{span}\{\Xi_j p_{j,k}\}_{k=1,\ldots,m_j}$.

$$(1) \Leftrightarrow \underbrace{\Xi_j(v - \Pi_j v)}_{v_j}|_{a,\Omega_j}^2 \le C_1 |v|_{a,\Omega_j}^2, \Leftrightarrow \Xi_j(v - \Pi_j v)|_{a,\Omega_j^\circ}^2 + \underbrace{|\Xi_j(v - \Pi_j v)|_{a,\Omega_j \setminus \Omega_j^\circ}^2}_{= |v - \Pi_j v|_{a,\Omega_j \setminus \Omega_j^\circ}^2 \le |v - \Pi_j v|_{a,\Omega_j}^2} \le C_1 |v|_{a,\Omega_j}^2.$$

So the only term that we are left to work on is: $\Xi_j(v - \Pi_j v)|_{a,\Omega_j^\circ}^2 \overset{\text{HOW?}}{\le} C_1 |v|_{a,\Omega_j}^2$, and the generalized eigenvalue problem bounds just that.

Numerical results

Coefficients

Decompositions

	AS		ZEM			GenEO		
κ_2	it	*cond*	it	*cond*	dim	it	*cond*	dim
1	16	*229*	11	*6.3*	8	11	*8.4*	7
10^2	27	*230*	19	*22*	8	13	*8.4*	14
10^4	29	*230*	23	*210*	8	15	*8.4*	14
10^6	26	*230*	22	*230*	8	11	*8.4*	14

Table: 3D Darcy: number of PCG iterations (it), condition number (*cond*) and coarse space dimension (dim) vs. jump in κ for $\kappa_1 = 1$, $\ell = 1$ added layers, $L = 8$ regular subdomains

		AS		ZEM			GenEO		
L	glob DOF	it	*cond*	it	*cond*	dim	it	*cond*	dim
4	14520	79	$2.4 \cdot 10^3$	54	$2.9 \cdot 10^3$	24	16	*10*	46
8	29040	177	$1.3 \cdot 10^4$	87	$1.0 \cdot 10^3$	48	16	*10*	102
16	58080	378	$1.5 \cdot 10^5$	145	$1.4 \cdot 10^3$	96	16	*10*	214

Table: 3D Elasticity: number of PCG iterations (it), condition number (*cond*), and coarse space dimension (dim) vs. number of regular subdomains, for $\ell = 1$ added layers, $g = 10$, $(E_1, \nu_1) = (2 \cdot 10^{11}, 0.3)$ and $(E_2, \nu_2) = (2 \cdot 10^7, 0.45)$.

Achieving robustness through coarse space enrichment

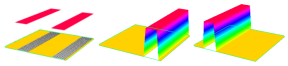
Nicole Spillane (Laboratoire Jacques Louis Lions, UPMC, Paris, France)

Joint work with: V. Dolean, P. Hauret, P. Jolivet, F. Nataf, C. Pechstein, D.J. Rixen, R. Scheichl

I) Some cases where the 'best' coarse space is known

IA) Darcy in Heterogeneous media: $-\nabla \cdot (\alpha \nabla u) = 1$

One coarse function per subdomain per high contrast layer.



C Vuik, A Segal, and J.A. Meijerink, JCP (1999).

One coarse function per subdomain.

I. G. Graham, P. O. Lechner, and R. Scheichl, Numerische Mathematik (2007).

Hou, Wu, and Cai, Mathematics of Computation (1999).

IB) Elasticity in the incompressible limit: $\nu \to 1/2$ with discontinuous pressure

Non-overlapping DD: one coarse vector per subdomain (preserves the volume).

B. Vereecke, H. Bavestrello, D. Dureisseix, CMAME (2003).

S. Gippert, A. Klawonn, M. Lanser, P. Radtke, and O. Rheinbach, DD21 Proceedings (2013).

Overlapping DD One coarse vector per face.

C. R. Dohrmann and O. B. Widlund, IJNME (2010).

Others: P. Le Tallec, J. Mandel and M. Vidrascu, SIAM J. Numer. Anal. (1998). (plates & shells)

C. Farhat, P.S. Chen and J. Mandel, INJME (1995). (time dependent)

II) Automatic construction of coarse spaces based on Generalized Eigenvalue Problems in the overlaps (GenEO)

Abstract Schwarz framework for the problem: Find $\mathbf{x}_* \in V$ such that $A\mathbf{x}_* = \mathbf{b}$.

Assume that $A = \Sigma_\tau (A_\tau)$ is spd, A_τ are spsd.

Choose

1. local subspaces $V_j \subset V, \quad j = 1, \dots, N$
2. interpolators $\tilde{R}_j^\top : V_j \to V$
3. local solvers $\tilde{A}_j : V_j \to V_j$
4. the coarse space $V_H \subset V$, $R_H^\top : V_H \to V$

Requirement: $V = \Sigma_{j=1}^N \tilde{R}_j^\top V_j + R_H^\top V_H$.

Two level preconditioner

$$M_{2level}^{-1} = P_0 + (I - P_0)\Sigma_{j=1}^N \tilde{R}_j^\top \tilde{A}_j^{-1} \tilde{R}_j (I - P_0)^\top,$$

where P_0 is the A-orthogonal projection onto V_H.

Two Crucial assumptions Stable splitting: $\forall u \in \text{range}(I - P_0)$ there exists $(z_1, \dots, z_N) \in V_1 \times \dots V_N$ such that

(1) $\quad u = \Sigma_{j=1}^N \tilde{R}_j^\top z_j$ and $\Sigma_{j=1}^N \langle \tilde{A}_j z_j, z_j \rangle \leq C_0^2 \langle Au, u \rangle$.

Stability of the local solver w.r.t. the exact solver: for all $u_j \in \text{range}(\tilde{A}_j^{-1} \tilde{R}_j (I - P_0))$,

(2) $\quad \langle A\tilde{R}_j^\top u_j, \tilde{R}_j^\top u_j \rangle \leq \omega \langle \tilde{A}_j u_j, u_j \rangle$.

Then, $\begin{vmatrix} eig(M_{2level}^{-1} A) \geq C_0^2 \\ eig(M_{2level}^{-1} A) \leq \mathcal{N}\omega \end{vmatrix}$ so $cond(M_{2level}^{-1} A) \leq \mathcal{N}\, \omega\, C_0^{-2}$, where $\mathcal{N}$ measures the maximal number of neighbours of a subdomain.

Additive Schwarz

N. S., V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl, Numerische Mathematik (2013).

N. S., V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl, C.R. Mathématiques (2011).

N. S., PhD thesis, (coming very soon!).

Definition in the Abstract framework:

1. V_j: N Overlapping subdomains
2. $R_j^\top$: Assembly operators
3. $\tilde{A}_j = R_j A R_j^\top$ (Exact local solvers)

(2) holds with $\omega = 1$ on the whole of V:

$$\langle \tilde{A}_j u_j, u_j \rangle = \langle A R_j^\top u_j, R_j^\top u_j \rangle.$$

(1) can be rewritten as N local conditions for any given C_0:

- Choose partition of unity operators: $\Xi_j : V_h(\Omega_j) \to V_{h,0}(\Omega_j)$.
- Define the components in the stable splitting: $z_j = \Xi_j(u_{|\Omega_j})$.
- Now (1)$\Leftarrow$ $(\Sigma_{j=1}^N \langle A R_j^\top \Xi_j(u_{|\Omega_j}), R_j^\top \Xi_j(u_{|\Omega_j}) \rangle \leq C_0^2 \langle Au, u \rangle)$.
- Or, locally (since $\mathcal{N}\langle Au, u \rangle \leq \Sigma_{j=1}^N \langle A_{|\Omega_j} u_{|\Omega_j}, u_{|\Omega_j} \rangle$)

$$(1) \Leftarrow \left(\langle \tilde{A}_j \Xi_j(u_{|\Omega_j}), \Xi_j(u_{|\Omega_j}) \rangle \leq \frac{C_0^2}{\mathcal{N}} \langle A_{|\Omega_j} u_{|\Omega_j}, u_{|\Omega_j} \rangle, \quad \forall j \right).$$

Definition: Schwarz GenEO coarse space

Find generalized eigenpairs $(\Lambda_j^k, p_j^k) \in (\mathbb{R}^+ \cup \{+\infty\}) \times V_h(\Omega_j)$ of

$$A_{|\Omega_j} p_j^k = \Lambda_j^k \Xi_j \tilde{A}_j \Xi_j p_j^k.$$

Choose a threshold τ and define

$$V_H = \text{span}\{R_j^\top \Xi_j(p_j^k); \Lambda_j^k < \tau,\ j = 1, \dots, N\}.$$

Then

$cond(M_{2level}^{-1} A) \leq \mathcal{N}\left(1 + \frac{\mathcal{N}}{\tau}\right)$ $\quad \begin{vmatrix} \mathcal{N} : \text{'number of neighbours'} \\ \tau : \text{threshold chosen by user} \end{vmatrix}$

Notation for Schwarz & FETI

Space of FE functions in Ω_j:
$V_{h,0}(\Omega_j) = \{u_{|\Omega_j}; u \in V_h \text{ and } \text{supp}(u) \subset \bar{\Omega}_j\}$.
Restrictions to Ω_j: $V_h(|\Omega_j) = \{u_{|\Omega_j}; u \in V_h\}$.
Problem restricted to Ω_j: $A_{|\Omega_j} = \Sigma_{\tau \subset \bar{\Omega}_j} A_\tau$.
Schur complement: $S_j = A^{\Gamma\Gamma}_{|\Omega_j} + A^{\Gamma I}_{|\Omega_j} A^{II\,-1}_{|\Omega_j} A^{I\Gamma}_{|\Omega_j}$.

- $R_1^\top \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} + R_2^\top \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} x_1^1 + x_2^1 \\ x_1^2 + x_2^2 \\ x_1^3 + x_2^3 \end{pmatrix}$ **assembly**
- $B_1 \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{pmatrix} + B_2 \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} x_1^1 - x_2^1 \\ x_1^2 - x_2^2 \\ x_1^3 - x_2^3 \end{pmatrix}$ **jump**

Bibliography: Related methods

Schwarz:

J. Galvis and Y. Efendiev, Multiscale Modeling & Simulation (2010).

Y. Efendiev, J. Galvis, R. Lazarov, and J. Willems, ESAIM (2012).

V. Dolean, F. Nataf, R. Scheichl, and N. Spillane, CMAM (2012).

BDD and FETI:

J. Mandel and B. Sousedík, CMAME (2007).

J. Šístek, , B. Sousedík, and J. Mandel (2013).

Spectral AMG:

M. Brezina, C. Heberton, J. Mandel, and P. Vaněk (2001).

T. Chartier, R. D. Falgout, V. E. Henson, J. Jones, T. Manteuffel, S. McCormick, J. Ruge, and P. S. Vassilevski, SIAM J. Sci. Comput (2003).

J. Xu, L. Zikatanov, Computing and Vizualization in Schience (2003).

FETI

N. S., D.J. Rixen, Automatic Spectral coarse spaces for robust FETI and BDD methods, IJNME (2013).

Find λ in the set of admissible constraints:

$$PM_{FETI}^{-1} P^\top \underbrace{(\Sigma_{i=1}^N B_j S_j^\dagger B_j^\top)}_{:=F} = PM_{FETI}^{-1} P^\top d$$

where (with diagonal scaling matrices $D = \text{diag}(D_1, \dots, D_N)$)
$M_{FETI}^{-1} = \Sigma_{i=1}^N (BD^{-1}B^\top)^{-1} B_j D_j^{-1} S_j D_j^{-1} B_j (BD^{-1}B^\top)^{-1}$.

Spectrum of $M_{FETI}^{-1} F$ = spectrum $F M_{FETI}^{-1}$.

We write $\mathbf{FM}_{\mathbf{FETI}}^{-1}$ in the Abstract Framework: $A = M_{FETI}^{-1}$

1. $V_j = \{\text{dofs on the boundary of } \Omega_j\}$
2. $\tilde{R}_j^\top = B_j$ jump operator
3. $\tilde{A}_j = S_j$

Indeed, $\Sigma_{j=1}^N \tilde{R}_j^\top \tilde{A}_j^\dagger \tilde{R}_j := \Sigma_{j=1}^N B_j S_j^\dagger B_j^\top := F$.

(1) holds with $C_0^2 = 1$ on the whole of V: Given $u \in V$, let $z_j := D_j^{-1} B_j (BD^{-1}B^\top)^{-1} u$ then

$$\Sigma_{j=1}^N B_j z_j = \Sigma_{j=1}^N B_j D_j^{-1} B_j (BD^{-1}B^\top)^{-1} u = u,$$

and
$\Sigma_{j=1}^N \langle \tilde{A}_j z_j, z_j \rangle := \Sigma_{j=1}^N \langle S_j z_j, z_j \rangle = \langle M_{FETI}^{-1} u, u \rangle := \langle Au, u \rangle$.

(2) rewrites: $\langle M_{FETI}^{-1} B_j u_j, B_j u_j \rangle \leq \omega \langle S_j u_j, u_j \rangle$.

Definition: FETI GenEO coarse space

Find generalized eigenpairs $(\Lambda_j^k, p_j^k) \in \mathbb{R}^+ \times V_j$ of

$$S_j p_j^k = \Lambda_j^k B_j^\top M_{FETI}^{-1} B_j p_j^k.$$

Choose a threshold τ and define

$$V_H = \text{span}\{M_{FETI}^{-1} B_j p_j^k; 0 < \Lambda_j^k < \tau,\ j = 1, \dots, N\}.$$

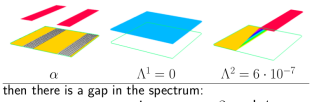
Then

$cond(M_{2level}^{-1} P^\top F_{|\text{Im}(P)}) \leq \frac{\mathcal{N}}{\tau}$ $\quad \begin{vmatrix} \mathcal{N} : \text{'number of neighbours'} \\ \tau : \text{threshold chosen by user} \end{vmatrix}$

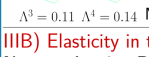
III) Does the Automatic coarse space find the 'best' coarse space?

IIIA) Darcy in Heterogeneous media: $-\nabla \cdot (\alpha \nabla u) = 1$ with Schwarz-GenEO

Layers: Number of basis functions = Number of high coefficient layers.



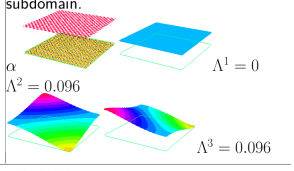
then there is a gap in the spectrum:



eigenvectors 3 and 4 are NOT related to α and NOT in the coarse space.

$\Lambda^3 = 0.11$ $\Lambda^4 = 0.14$

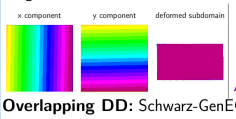
Islands: One coarse function per subdomain.



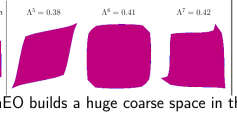
IIIB) Elasticity in the incompressible limit: $\nu = 0.4999$ ($\mathbb{P}_2 - \mathbb{P}_0$ pressure eliminated)

Non-overlapping DD:

Regular Subdomain: $\Lambda^4 = 0.003$



Next 3 eigenvectors



Metis Subdomain: $\Lambda^4 = 0.04$

Overlapping DD: Schwarz-GenEO builds a huge coarse space in the near incompressible limit. There is no easy fix without going back to the PDEs and building a new "mass" matrix.

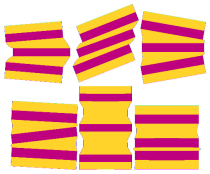
IIIC) 2D Elasticity (Pink: $E = 10^6$; Yellow: $E = 1$) with FETI-GenEO

Layers: Number of basis functions = (Number of hard layers - 1) × Number of rigid body modes

- Rigid body modes $\Lambda^{1 \text{ to } 3} = 0$.

(the rigid body modes are already taken care of by FETI.)

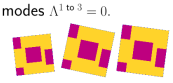
- Coarse space: $\Lambda^{4 \text{ to } 9} = \{8 \cdot 10^{-4}; 9 \cdot 10^{-4}; 3 \cdot 10^{-3}; \cdot 10^{-2}; 4 \cdot 10^{-2}; 0.11\}$.



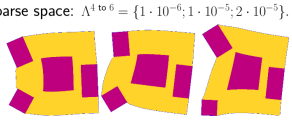
- Next 2 eigenvectors: $\Lambda^{10 \text{ and } 11} = 0.99$

Islands: Only corner islands are picked up

- Rigid body modes $\Lambda^{1 \text{ to } 3} = 0$.



- Coarse space: $\Lambda^{4 \text{ to } 6} = \{1 \cdot 10^{-6}; 1 \cdot 10^{-5}; 2 \cdot 10^{-5}\}$.



- Next 3 eigenvectors: $\Lambda^{7 \text{ to } 10} = \{0.1; 0.3; 0.4\}$

Remark: this is in agreement with a heuristic generalization to elasticity of

C. Pechstein, R. Scheichl, Numerische Mathematik (2011).

Bibliography

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, third edition, 1999.

[2] M. Bhardwaj, D. Day, C. Farhat, M. Lesoinne, K. Pierson, and D. Rixen. Application of the FETI method to ASCI problems: Scalability results on a thousand processors and discussion of highly heterogeneous problems. *Internat. J. Numer. Methods Engrg.*, 47(1–3):513–535, 2000.

[3] P. E. Bjørstad and O. B. Widlund. Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM J. Numer. Anal.*, 23(6):1097–1120, 1986.

[4] J. H. Bramble, J. E. Pasciak, and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring. I. *Math. Comp.*, 47(175):103–134, 1986.

[5] A. Brandt, S. McCormick, and J. Ruge. Algebraic multigrid (AMG) for sparse matrix equations. In *Sparsity and its applications (Loughborough, 1983)*, pages 257–284. Cambridge Univ. Press, Cambridge, 1985.

[6] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[7] M. Brezina, C. Heberton, J. Mandel, and P. Vaněk. An iterative method with convergence rate chosen a priori. Technical Report 140, University of Colorado Denver, April 1999.

[8] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[9] X.-C. Cai and M. Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM J. Sci. Comput.*, 21(2):792–797 (electronic), 1999.

[10] X.-C. Cai and O. B. Widlund. Multiplicative Schwarz algorithms for some nonsymmetric and indefinite problems. *SIAM J. Numer. Anal.*, 30(4):936–952, 1993.

[11] T. F. Chan and T. P. Mathew. Domain decomposition algorithms. In *Acta numerica, 1994*, Acta Numer., pages 61–143. Cambridge Univ. Press, Cambridge, 1994.

[12] T. Chartier, R. D. Falgout, V. E. Henson, J. Jones, T. Manteuffel, S. McCormick, J. Ruge, and P. S. Vassilevski. Spectral AMGe (ρAMGe). *SIAM J. Sci. Comput.*, 25(1):1–26, 2003.

[13] C. Chevalier and F. Pellegrini. PT-Scotch: a tool for efficient parallel graph ordering. *Parallel Comput.*, 34(6-8):318–331, 2008.

[14] Y.-H. De Roeck and P. Le Tallec. Analysis and test of a local domain-decomposition preconditioner. In *Fourth International Symposium on Domain Decomposition Meth-*

ods for Partial Differential Equations (Moscow, 1990), pages 112–128, Philadelphia, PA, 1991. SIAM.

[15] J. W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[16] B. Després. *Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régime harmonique. Le théorème de Borg pour l'équation de Hill vectorielle*. Institut National de Recherche en Informatique et en Automatique (INRIA), Rocquencourt, 1991. Thèse, Université de Paris IX (Dauphine), Paris, 1991.

[17] C. R. Dohrmann and O. B. Widlund. An overlapping Schwarz algorithm for almost incompressible elasticity. *SIAM J. Numer. Anal.*, 47(4):2897–2923, 2009.

[18] C. R. Dohrmann and O. B. Widlund. Hybrid domain decomposition algorithms for compressible and almost incompressible elasticity. *Internat. J. Numer. Methods Engrg.*, 82(2):157–183, 2010.

[19] V. Dolean, P. Jolivet, and F. Nataf. An Introduction to Domain Decomposition Methods: algorithms, theory and parallel implementation. Lecture notes (in preparation).

[20] V. Dolean, F. Nataf, and G. Rapin. Deriving a new domain decomposition method for the Stokes equations using the Smith factorization. *Math. Comp.*, 78(266):789–814, 2009.

[21] V. Dolean, F. Nataf, R. Scheichl, and N. Spillane. Analysis of a two-level Schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. *Comput. Methods Appl. Math.*, 12(4):391–414, 2012.

[22] M. Dryja, M. V. Sarkis, and O. B. Widlund. Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. *Numer. Math.*, 72(3):313–348, 1996.

[23] M. Dryja, B. F. Smith, and O. B. Widlund. Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions. *SIAM J. Numer. Anal.*, 31(6):1662–1694, 1994.

[24] M. Dryja and O. B. Widlund. Some domain decomposition algorithms for elliptic problems. In *Iterative methods for large linear systems (Austin, TX, 1988)*, pages 273–291. Academic Press, Boston, MA, 1990.

[25] M. Dryja and O. B. Widlund. Domain decomposition algorithms with small overlap. *SIAM J. Sci. Comput.*, 15(3):604–620, 1994. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992).

[26] Y. Efendiev, J. Galvis, R. Lazarov, and J. Willems. Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms. *ESAIM Math. Model. Numer. Anal.*, 46(5):1175–1199, 2012.

[27] Y. Efendiev, J. Galvis, and P. Vassilevski. Multiscale spectral AMGe solvers for high-contrast flow problems. ISC-Preprint 2012-02, Inst. Scientific Computation, Texas A&M University, 2012.

[28] Y. Efendiev, J. Galvis, and P. S. Vassilevski. Spectral element agglomerate algebraic multigrid methods for elliptic problems with high-contrast coefficients. In *Domain decomposition methods in science and engineering XIX*, volume 78 of *Lect. Notes Comput. Sci. Eng.*, pages 407–414, Heidelberg, 2011. Springer.

[29] E. Efstathiou and M. J. Gander. Why restricted additive Schwarz converges faster than additive Schwarz. *BIT*, 43(suppl.):945–959, 2003.

[30] Y. A. Erlangga and R. Nabben. Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetric matrices. *SIAM J. Matrix Anal. Appl.*, 30(2):684–699, 2008.

[31] J. F. Escobar. The geometry of the first non-zero Stekloff eigenvalue. *J. Funct. Anal.*, 150(2):544–556, 1997.

[32] C. Farhat, P. Chen, and J. Mandel. A scalable Lagrange multiplier based domain decomposition method for time-dependent problems. *Internat. J. Numer. Methods Engrg.*, 38(22):3831–3853, 1995.

[33] C. Farhat, P.-S. Chen, J. Mandel, and F. X. Roux. The two-level FETI method. II. Extension to shell problems, parallel implementation and performance results. *Comput. Methods Appl. Mech. Engrg.*, 155(1-2):153–179, 1998.

[34] C. Farhat, N. Maman, and G. W. Brown. Mesh partitioning for implicit computations via iterative domain decomposition: impact and optimization of the subdomain aspect ratio. *Internat. J. Numer. Methods Engrg.*, 38(6):989–1000, 1995.

[35] C. Farhat and J. Mandel. The two-level FETI method for static and dynamic plate problems. I. An optimal iterative solver for biharmonic systems. *Comput. Methods Appl. Mech. Engrg.*, 155(1-2):129–151, 1998.

[36] C. Farhat and F.-X. Roux. A method of finite element tearing and interconnecting and its parallel solution algorithm. *Internat. J. Numer. Methods Engrg.*, 32:1205–1227, 1991.

[37] C. Farhat and F.-X. Roux. Implicit parallel processing in structural mechanics. *Comput. Mech. Adv.*, 2(1):124, 1994.

[38] J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Model. Simul.*, 8(4):1461–1483, 2010.

[39] J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high contrast media: reduced dimension coarse spaces. *Multiscale Model. Simul.*, 8(5):1621–1644, 2010.

[40] M. J. Gander. Optimized Schwarz methods. *SIAM J. Numer. Anal.*, 44(2):699–731 (electronic), 2006.

[41] M. J. Gander. Schwarz methods over the course of time. *Electron. Trans. Numer. Anal.*, 31:228–255, 2008.

[42] M. J. Gander, F. Magoulès, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60 (electronic), 2002.

[43] S. Gippert, A. Klawonn, M. Lanser, P. Radtke, and O. Rheinbach. Nonlinear domain decomposition, adaptive coarse spaces, and a new coarse space for almost incompressible linear elasticity. In *Domain decomposition methods in science and engineering XXI*, Lect. Notes Comput. Sci. Eng., Heidelberg, 2013. Springer.

[44] P. Gosselet, V. Chiaruttini, C. Rey, and F. Feyel. A monolithic strategy based on an hybrid domain decomposition method for multiphysic problems: Application to poroelasticity. *Revue Européenne des Éléments*, 13(5-7):523–534, 2004.

[45] P. Gosselet and C. Rey. Non-overlapping domain decomposition methods in structural mechanics. *Arch. Comput. Methods Engrg.*, 13(4):515–572, 2006.

[46] P. Gosselet, C. Rey, and J. Pebrel. Total and selective reuse of Krylov subspaces for the resolution of sequences of nonlinear structural problems. *Internat. J. Numer. Methods Engrg.*, 94(1):60–83, 2013.

[47] I. G. Graham, P. O. Lechner, and R. Scheichl. Domain decomposition for multiscale PDEs. *Numer. Math.*, 106(4):589–626, 2007.

[48] W. Hackbusch. Multi-grid methods and applications, vol. 4 of springer series in computational mathematics, 1985.

[49] P. Havé, R. Masson, F. Nataf, M. Szydlarski, H. Xiang, and T. Zhao. Algebraic domain decomposition methods for highly heterogeneous problems. *SIAM J. Sci. Comput.*, 35(3):C284–C302, 2013.

[50] F. Hecht. *FreeFem++*. Numerical Mathematics and Scientific Computation. Laboratoire J.L. Lions, Université Pierre et Marie Curie, http://www.freefem.org/ff++/, 3.23 edition, 2013.

[51] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

[52] P. Jolivet, V. Dolean, F. Hecht, F. Nataf, C. Prud'Homme, and N. Spillane. High performance domain decomposition methods on massively parallel architectures with freefem++. *J. Numer. Math.*, 20(3-4):287–302, 2012.

[53] S. Kaniel. Estimates for some computational techniques in linear algebra. *Mathematics of Computation*, 20(95):369–378, 1966.

[54] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392 (electronic), 1998.

[55] A. Klawonn and O. Rheinbach. Robust FETI-DP methods for heterogeneous three dimensional elasticity problems. *Comput. Methods Appl. Mech. Engrg.*, 196(8):1400–1414, 2007.

[56] A. Klawonn and O. Rheinbach. Deflation, projector preconditioning, and balancing in iterative substructuring methods: connections and new results. *SIAM J. Sci. Comput.*, 34(1):A459–A484, 2012.

[57] A. Klawonn, O. Rheinbach, and O. B. Widlund. An analysis of a FETI-DP algorithm on irregular subdomains in the plane. *SIAM J. Numer. Anal.*, 46(5):2484–2504, 2008.

[58] A. Klawonn and O. B. Widlund. FETI and Neumann-Neumann iterative substructuring methods: connections and new results. *Comm. Pure Appl. Math.*, 54(1):57–90, 2001.

[59] A. V. Knyazev. Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comput.*, 23(2):517–541 (electronic), 2001. Copper Mountain Conference (2000).

[60] C. Lanczos. Solution of systems of linear equations by minimized-iterations. *J. Research Nat. Bur. Standards*, 49:33–53, 1952.

[61] P. Le Tallec. Domain decomposition methods in computational mechanics. *Comput. Mech. Adv.*, 1(2):121–220, 1994.

[62] P. Le Tallec. Numerical methods for nonlinear three-dimensional elasticity. In *Handbook of numerical analysis, Vol. III*, Handb. Numer. Anal., III, pages 465–622. North-Holland, Amsterdam, 1994.

[63] P. Le Tallec, J. Mandel, and M. Vidrascu. A Neumann-Neumann domain decomposition algorithm for solving plate and shell problems. *SIAM J. Numer. Anal.*, 35(2):836–867 (electronic), 1998.

[64] G. Leborgne. Diagonalisation : valeurs propres, valeurs propres généralisées, 2008. ISIMA Lecture Notes in French.

[65] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide*, volume 6 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods.

[66] P.-L. Lions. On the Schwarz alternating method. III. A variant for nonoverlapping subdomains. In *Third International Symposium on Domain Decomposition Methods for Partial Differential Equations (Houston, TX, 1989)*, pages 202–223, Philadelphia, PA, 1990. SIAM.

[67] J. Mandel. Balancing domain decomposition. *Comm. Numer. Methods Engrg.*, 9(3):233–241, 1993.

[68] J. Mandel and M. Brezina. Balancing domain decomposition for problems with large jumps in coefficients. *Math. Comp.*, 65(216):1387–1401, 1996.

[69] J. Mandel, C. R. Dohrmann, and R. Tezaur. An algebraic theory for primal and dual substructuring methods by constraints. *Appl. Numer. Math.*, 54(2):167–193, 2005.

[70] J. Mandel and B. Sousedík. Adaptive selection of face coarse degrees of freedom in the BDDC and the FETI-DP iterative substructuring methods. *Comput. Methods Appl. Mech. Engrg.*, 196(8):1389–1399, 2007.

[71] J. Mandel and R. Tezaur. Convergence of a substructuring method with Lagrange multipliers. *Numer. Math.*, 73(4):473–487, 1996.

[72] A. M. Matsokin and S. V. Nepomnyashchikh. The Schwarz alternation method in a subspace. *Izv. Vyssh. Uchebn. Zaved. Mat.*, (10):61–66, 85, 1985.

[73] G. Meinardus. *Approximation of functions: Theory and numerical methods*. Expanded translation of the German edition. Translated by Larry L. Schumaker. Springer Tracts in Natural Philosophy, Vol. 13. Springer-Verlag New York, Inc., New York, 1967.

[74] A. Napov and Y. Notay. Algebraic analysis of aggregation-based multigrid. *Numer. Linear Algebra Appl.*, 18(3):539–564, 2011.

[75] A. Napov and Y. Notay. An algebraic multigrid method with guaranteed convergence rate. *SIAM J. Sci. Comput.*, 34(2):A1079–A1109, 2012.

[76] F. Nataf and G. Rapin. Construction of a new domain decomposition method for the Stokes equations. In *Domain decomposition methods in science and engineering XVI*, volume 55 of *Lect. Notes Comput. Sci. Eng.*, pages 247–254. Springer, Berlin, 2007.

[77] F. Nataf, F. Rogier, and E. de Sturler. Optimal interface conditions for domain decomposition methods. *CMAP (Ecole Polytechnique)*, 1994. Tech. Rep 301.

[78] F. Nataf, H. Xiang, and V. Dolean. A two level domain decomposition preconditioner based on local Dirichlet-to-Neumann maps. *C. R. Math. Acad. Sci. Paris*, 348(21-22):1163–1167, 2010.

[79] F. Nataf, H. Xiang, V. Dolean, and N. Spillane. A coarse space construction based on local Dirichlet-to-Neumann maps. *SIAM J. Sci. Comput.*, 33(4):1623–1642, 2011.

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

[81] B. N. Parlett. *The symmetric eigenvalue problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[82] A. Pechstein and J. Schöberl. Tangential-displacement and normal-normal-stress continuous mixed finite elements for elasticity. *Math. Models Methods Appl. Sci.*, 21(8):1761–1782, 2011.

[83] C. Pechstein. *Finite and boundary element tearing and interconnecting solvers for multiscale problems*, volume 90 of *Lecture Notes in Computational Science and Engineering*. Springer, Heidelberg, 2013.

[84] C. Pechstein, M. Sarkis, and R. Scheichl. New theoretical coefficient robustness results for FETI-DP. In *Domain Decomposition Methods in Science and Engineering XX*, Lecture Notes in Computational Science and Engineering. Springer-Verlag, 2011.

[85] C. Pechstein and R. Scheichl. Analysis of FETI methods for multiscale PDEs. *Numer. Math.*, 111(2):293–333, 2008.

[86] C. Pechstein and R. Scheichl. Scaling up through domain decomposition. *Appl. Anal.*, 88(10-11):1589–1608, 2009.

[87] C. Pechstein and R. Scheichl. Analysis of FETI methods for multiscale PDEs. Part II: interface variation. *Numer. Math.*, 118(3):485–529, 2011.

[88] C. Pechstein and R. Scheichl. Weighted Poincaré inequalities and applications in domain decomposition. In *Domain decomposition methods in science and engineering XIX*, volume 78 of *Lect. Notes Comput. Sci. Eng.*, pages 197–204, Heidelberg, 2011. Springer.

[89] C. Pechstein and R. Scheichl. Weighted Poincaré inequalities. *IMA J. Numer. Anal.*, 33(2):652–686, 2013.

[90] D. Rixen. Extended preconditioners for FETI method applied to constrained problems. *Internat. J. Numer. Methods Engrg.*, 54(1):1–26, 2002.

[91] D. J. Rixen. *Substructuring and dual methods in structural analysis*. PhD thesis, University of Liège, Belgique, 1997.

[92] D. J. Rixen. Dual Schur complement method for semi-definite problems. In *Domain decomposition methods, 10 (Boulder, CO, 1997)*, volume 218 of *Contemp. Math.*, pages 341–348. Amer. Math. Soc., Providence, RI, 1998.

[93] D. J. Rixen and C. Farhat. A simple and efficient extension of a class of substructure based preconditioners to heterogeneous structural mechanics problems. *Internat. J. Numer. Methods Engrg.*, 44(4):489–516, 1999.

[94] J. W. Ruge and K. Stüben. Algebraic multigrid. In *Multigrid methods*, volume 3 of *Frontiers Appl. Math.*, pages 73–130. SIAM, Philadelphia, PA, 1987.

[95] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2003.

[96] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[97] M. Sarkis. Partition of unity coarse spaces: enhanced versions, discontinuous coefficients and applications to elasticity. In *Domain decomposition methods in science and engineering*, pages 149–158 (electronic). Natl. Auton. Univ. Mex., México, 2003.

[98] R. Scheichl and E. Vainikko. Additive Schwarz with aggregation-based coarsening for elliptic problems with highly variable coefficients. *Computing*, 80(4):319–343, 2007.

[99] R. Scheichl, P. S. Vassilevski, and L. T. Zikatanov. Weak approximation properties of elliptic projections with functional constraints. *Multiscale Model. Simul.*, 9(4):1677–1699, 2011.

[100] R. Scheichl, P. S. Vassilevski, and L. T. Zikatanov. Multilevel methods for elliptic problems with highly varying coefficients on nonaligned coarse grids. *SIAM J. Numer. Anal.*, 50(3):1675–1694, 2012.

[101] H. A. Schwarz. Über einen Grenzübergang durch alternirendes Verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, 15:272–286, 1870.

[102] B. F. Smith, P. E. Bjørstad, and W. D. Gropp. *Domain decomposition*. Cambridge University Press, Cambridge, 1996. Parallel multilevel methods for elliptic partial differential equations.

[103] B. Sousedík, J. Šístek, and J. Mandel. Adaptive-Multilevel BDDC and its parallel implementation. *Computing*, 95(12):1087–1119, 2013.

[104] N. Spillane. How to make a domain decomposition method more robust. In *LSSC Proceedings – Springer Lecture Notes in Computer Science (accepted)*, 2013.

[105] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. A robust two-level domain decomposition preconditioner for systems of PDEs. *C. R. Math. Acad. Sci. Paris*, 349(23-24):1255–1259, 2011.

[106] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numerische Mathematik*, pages 1–30 (currently published online), 2013.

[107] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. Achieving robustness through coarse space enrichment in the two level Schwarz framework. In *Domain decomposition methods in science and engineering XXI*, Lect. Notes Comput. Sci. Eng., Heidelberg, 2013. Springer.

[108] N. Spillane, V. Dolean, P. Hauret, F. Nataf, and D. J. Rixen. Solving generalized eigenvalue problems on the interfaces to build a robust two-level FETI method. *C. R. Math. Acad. Sci. Paris*, 351(5-6):197–201, 2013.

[109] N. Spillane and D. J. Rixen. Automatic spectral coarse spaces for robust finite element tearing and interconnecting and balanced domain decomposition algorithms. *Internat. J. Numer. Methods Engrg.*, 95(11):953–990, 2013.

[110] J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga. Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods. *J. Sci. Comput.*, 39(3):340–370, 2009.

[111] R. Tezaur. *Analysis of Lagrange multiplier based domain decomposition*. ProQuest LLC, Ann Arbor, MI, 1998. Thesis (Ph.D.)–University of Colorado at Denver.

[112] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.

[113] A. van der Sluis and H. A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48(5):543–560, 1986.

[114] B. Vereecke, H. Bavestrello, and D. Dureisseix. An extension of the FETI domain decomposition method for incompressible and nearly incompressible problems. *Computer methods in applied mechanics and engineering*, 192(31):3409–3429, 2003.

[115] C. Vuik, A. Segal, and J. A. Meijerink. An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. *Journal of Computational Physics*, 152(1):385–403, 1999.

[116] J. Willems. Robust multilevel solvers for high-contrast anisotropic multiscale problems. *J. Comput. Appl. Math.*, 251:47–60, 2013.

[117] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34(4):581–613, 1992.

Résumé – Abstract

UPMC
SORBONNE UNIVERSITÉS

Nicole SPILLANE

Méthodes de décomposition de domaine robustes pour les problèmes symétriques définis positifs

Résumé : L'objectif de cette thèse est de concevoir des méthodes de décomposition de domaine qui sont robustes même pour les problèmes difficiles auxquels on est confronté lorsqu'on simule des objets industriels ou qui existent dans la nature. Par exemple une difficulté à laquelle est confronté Michelin est que les pneus sont constitués de matériaux avec des lois de comportement très différentes (caoutchouc et acier). Ceci induit un ralentissement de la convergence des méthodes de décomposition de domaine classiques dès que la partition en sous domaines ne tient pas compte des hétérogénéités. Pour trois méthodes de décomposition de domaine (Schwarz Additif, BDD et FETI) nous avons prouvé qu'en résolvant des problèmes aux valeurs propres généralisés dans chacun des sous domaines on peut identifier automatiquement quels sont les modes responsables de la convergence lente. En d'autres termes on divise le problème de départ en deux : une partie où on peut montrer que la méthode de décomposition de domaine va converger et une seconde où on ne peut pas. L'idée finale est d'appliquer des projections pour résoudre ces deux problèmes indépendamment (c'est la déflation) : au premier on applique la méthode de décomposition de domaine et sur le second (qu'on appelle le problème grossier) on utilise un solveur direct qu'on sait être robuste. Nous garantissons théorétiquement que le solveur à deux niveaux qui résulte de ces choix est robuste. Un autre atout de nos algorithmes est qu'ils peuvent être implémentés en boîte noire ce qui veut dire que les matériaux hétérogènes ne sont qu'un exemple des difficultés qu'ils peuvent contourner.

Abstract : The objective of this thesis is to design domain decomposition methods which are robust even for hard problems that arise when simulating industrial or real life objects. For instance one particular challenge which the company Michelin is faced with is the fact that tires are made of rubber and steel which are two materials with very different behavior laws. With classical domain decomposition methods, as soon as the partition into subdomains does not accommodate the discontinuities between the different materials convergence deteriorates. For three popular domain decomposition methods (Additive Schwarz, FETI and BDD) we have proved that by solving a generalized eigenvalue problem in each of the subdomains we can identify automatically which are the modes responsible for slow convergence. In other words we can divide the original problem into two problems : the first one where we can guarantee that the domain decomposition method will converge quickly and the second where we cannot. The final idea is to apply projections to solve these two problems independently (this is also known as deflation) : on the first we apply the domain decomposition method and on the second (we call it the coarse space) we use a direct solver which we know will be robust. We guarantee theoretically that the resulting two level solver is robust. The other main feature of our algorithms is that they can be implemented as black box solvers meaning that heterogeneous materials is only one type of difficulty that they can identify and circumvent.