Domain decomposition methods for multiphysics problems

Vanzan, Tommaso

UNIVERSITÉ DE GENÈVE
Section de Mathématiques

FACULTÉ DES SCIENCES
Professeur Martin J. Gander

Domain decomposition methods for multiphysics problems

THÈSE

Présentée à la Faculté des Sciences de l'Université de Genève
pour obtenir le grade de Docteur ès Sciences, mention Mathématiques

par

Tommaso VANZAN
de
Biella (Italie)

Thèse N 5485

GENÈVE
Atelier d'impression ReproMail
2020

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION MATHEMATIQUES

Thèse de Monsieur Tommaso VANZAN

intitulée :

«Domain Decomposition Methods for Multiphysics Problems»

La Faculté des sciences, sur le préavis de Monsieur M. GANDER, professeur ordinaire et directeur de thèse (Section de mathématiques), Monsieur B. VANDEREYCKEN, professeur associé (Section de mathématiques), Monsieur G. CIARAMELLA, professeur (Fachbereich Mathematik und Statistik, Universität Konstanz, Deutschland), Monsieur R. MASSON, professeur (Laboratoire de mathématiques J.A. Dieudonné, Université de Nice Sophia Antipolis, France) et Monsieur A. QUARTERONI, professeur (MOX, Politecnico di Milano, Italia), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 8 septembre 2020

Thèse - 5485 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

“O frati”, dissi,“che per cento milia
perigli siete giunti a l’occidente,
a questa tanto picciola vigilia
d’i nostri sensi ch’è del rimanente
non vogliate negar l’esperïenza,
di retro al sol, del mondo sanza gente.
Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza.”

Dante Alighieri, La Divina
Commedia- Inferno: C. XXVI,
vv:112-120

Abstract

The aim of this thesis is to contribute to the development of domain decomposition methods which are techniques to solve efficiently large linear or nonlinear systems arising from the discretization of PDEs. Throughout the thesis, a special attention is dedicated to heterogeneous problems, that is problems where multiphysics phenomena are present. This thesis is divided in six chapters.

In the first part of Chapter 1, we introduce several one-level domain decomposition methods namely the parallel Schwarz method, the Dirichlet-Neumann method, the Neumann-Neumann method and the optimized Schwarz method. In the second part, we address the concept of scalability, that is how the convergence of a one-level domain decomposition method is affected when the number of subdomains grows. We present a theoretical analysis for the scalability of one-level domain decomposition methods for a strip decomposition. We then introduce discrete fracture networks, which are advanced mathematical models to simulate flows in a fractured medium, and we provide a theoretical analysis of the scalability of the optimized Schwarz method in some specific geometries.

In Chapter 2 we start our analysis of multiphysics PDEs. We remark that in this context, the use of a domain decomposition method does not necessarily involve a decomposition into thousands of subdomains in order to take advantage of a parallel architecture. On the contrary, the goal is to use a domain decomposition method to design efficient and robust partitioned strategies which permit to solve the different phenomena separately. Thus, the number of subdomains usually coincides with the number of different physical phenomena present in the domain. In this thesis we decided to study the optimized Schwarz methods, as the presence of transmission conditions to optimize allows one to tune them according to the physical parameters of the problem, which in turn makes the method more robust compared to other decoupling strategies. In this chapter we present a theoretical analysis for the coupling of heterogeneous second order PDEs and for the coupling of the Helmholtz equation with a Laplace equation. Guided by the theoretical analysis, we further define a numerical algorithm which permits to find optimized trans-

mission conditions even in those situations where the theoretical analysis fails.

One-level domain decomposition methods are rarely used as stand-alone solvers or preconditioners since they are generally not scalable. Scalability and a better convergence can be achieved introducing a coarse correction. In Chapter 3 we define and study a multilevel domain decomposition method where an optimized Schwarz method is used as a smoother on each level. Our analysis is based on Fourier techniques and provides good estimates for the optimized transmission conditions. This multilevel method is highly attractive for heterogeneous problems as it inherits robustness and efficiency from the one-level optimized Schwarz smoother.

Chapter 4 introduces a new computational framework for two-level and multilevel methods where both the smoother and the coarse correction are defined exclusively on the interfaces between the subdomains. An extensive theoretical analysis is provided for both spectral and geometric coarse spaces. Numerical results are presented which show the effectiveness of this approach and, in particular, the case of highly jumping diffusion coefficients is investigated.

We then consider the Stokes-Darcy coupling which is a mathematical model to describe the flow of a Newtonian fluid which interacts with a porous medium. The goal of Chapter 5 is to apply the techniques studied in the previous chapters, namely the one-level optimized Schwarz methods, the multilevel optimized Schwarz methods and the substructured two-level methods to solve efficiently the Stokes-Darcy system.

Finally Chapter 6 is dedicated to the exciting new field of nonlinear preconditioning. Starting from a recent nonlinear algorithm, we introduce a substructured version which relies on the framework discussed in Chapter 4. We study its convergence behaviour and we analyse extensively the advantages and disadvantages of the substructured algorithm compared to the volume one.

Résumé

Le but de cette thèse est de contribuer au développement des méthodes de décomposition de domaine, qui sont des techniques numériques pour résoudre efficacement des grands systèmes linéaires ou non-linéaires obtenus par la discrétisation des EDPs. Nous prêtons une attention particulière aux problèmes hétérogènes, c'est-à-dire des problèmes où des différents phénomènes physiques interagissent entre eux. Cette thèse contient six chapitres.

Dans la premiere partie du chapitre 1, nous présentons plusieurs méthodes classiques de décomposition de domaine à un niveau: la méthode de Schwarz, la méthode de Dirichlet-Neumann, la méthode de Neumann-Neumann et la méthode de Schwarz optimisée. Dans la deuxième partie, nous étudions la scalabilité, c'est-à-dire comment la convergence des méthodes est influencée quand le nombre des sous-domaines augmente. Nous présentons une analyse théorique de la scalabilité des différentes méthodes à un niveau pour une décomposition en bande. Après, nous introduisons les «réseaux des fractures discrétisées», qui sont des modèles mathématiques avancées pour décrire le mouvement des fluides dans un milieu fracturé, et nous présentons une analyse théorique de la scalabilité pour la méthode de Schwarz optimisée.

Dans le chapitre 2, nous commençons notre analyse pour les EDPs hétérogènes. Nous remarquons que dans ce contexte, nous n'utilisons pas nécessairement une méthode de décomposition de domaine avec des milliers des sous-domaines. En effet, le but n'est pas de profiter d'une architecture parallèle, mais d'utiliser une méthode de décomposition de domaine pour établir des stratégies efficaces et robustes qui permettent de résoudre les phénomènes différents séparément. Nous avons décidé d'étudier la méthode de Schwarz optimisée car elle a des conditions de transmission avec des paramètres à optimiser. Un bon choix de ces paramètres permet d'obtenir une méthode plus robuste par rapport aux autres. Dans ce chapitre, nous présentons une analyse théorique pour la solution des EDPs hétérogène elliptiques du second degré et pour le couplage d'une équation de Helmholtz avec une équation de Laplace. En s'appuyant sur les résultats théoriques, nous

définissons ainsi un algorithme numérique pour trouver des paramètres optimisés pour les situations où l'analyse théorique n'est plus valable.

Les méthodes à un niveau sont rarement utilisées seules comme solveurs or préconditionneurs car, en général, elles ne sont pas scalables. Une meilleure convergence et la scalabilité peuvent être obtenues en ajoutant un deuxième niveau grossier. Dans le chapitre 3, nous étudions une méthode de décomposition de domaine à plusieurs niveaux dans laquelle, à chaque niveau, une méthode de Schwarz optimisée est utilisée comme lisseur. Notre analyse utilise des techniques de Fourier et elle fournit de bonnes approximations pour les paramètres optimisés sur les plusieurs niveaux. Cette méthode à plusieurs niveaux est particulièrement intéressante pour les problèmes hétérogènes car elle hérite la robustesse et l'efficacité de le méthode de Schwarz optimisée à un niveau.

Le chapitre 4 introduit un nouveau cadre computationnel pour des méthodes à deux ou plusieurs niveaux où le lisseur et la correction grossière sont tous le deux définies directement sur les interfaces entre les sous-domaines. Une analyses théorique complète est présentée pour un espace grossier soit spectral soit géométrique. Des résultats numériques sont discutés et ils montrent le bon fonctionnement de cette approche. Une attention particulière est dédiée au cas où les coefficients de diffusion sont fortement discontinus.

Le chapitre 5 considère le système de Stokes-Darcy qui est un modèle mathématique pour décrire un fluide newtonien qui interagit avec un milieu poreux. Le but de ce chapitre est d'appliquer les différents algorithmes présentés dans les chapitres précédents, à savoir, la méthode de Schwarz optimisée à un niveau et à plusieurs niveaux ainsi que les méthodes sous-structurés, pour résoudre le système de Stokes-Darcy.

Pour finir, le chapitre 6 est consacré au nouveau domaine de préconditionnement nonlinéaire. Nous commençons par analyser un algorithme récent et nous définissons ensuite une version sous-structurée en utilisant les idées présentées dans le chapitre 4. Nous étudions la convergence de cette nouvelle méthode puis nous analysons les avantages et les inconvénients de cette dernière par rapport à la méthode en volume.

Acknowledgements

These last four years have represented an exciting chapter of my life. Not only I had the opportunity to do what I so far like the most, that is studying mathematics, but further I met some great people and made amazing friends. As both aspects cannot be taken for granted, it is a must to acknowledge those with whom I shared this journey.

First of all, I want to express my sincere gratitude to Prof. Martin J. Gander for giving me the opportunity to join his research group and for countless insightful research discussions which shaped me as a researcher. But besides research aspects, I am grateful to Prof. Martin J. Gander for his constant support, for creating a positive work environment and for numerous extra-academic advices. A special thank goes to Prof. Gabriele Ciaramella with whom I have established a long collaboration and friendship which resulted in parts of this thesis. I am also grateful to Prof. Roland Masson, Prof. Alfio Quarteroni and Prof. Bart Vandereycken who honour me by accepting to be part of the jury for my Ph.D. defense. I further acknowledge interesting discussions during the M.Sc. and the Ph.D. with Prof. Lamberto Rondoni, Prof. Giovanni Monegato and Prof. Stefano Berrone.

If the journey has mattered more than the destination is largely thanks to the "Paccari's group". I would like to thank my two academic brothers, Faycal and Pratik, for endless discussions on numerical analysis, for the great support they provided in several aspects and for being excellent travel companions. If I rarely felt home sick, it is largely thanks to Marco on whom I knew I could always count for a drink, a joke, an excursion or a heated political debate. A special thank goes to Jhih-Huang; although we crossed in Geneva for less than two years, I will always remember full of joy our bike trips in Biella, my visit in Taiwan and not to forget your tasty Taiwanese polenta! Thanks also to Eiichi and Ibrahim for all the jokes we shared and for have never kicked me out of their office during my daily visits. During these years I shared the "full-rank office" with Bo, Ding, Guillaume and Giancarlo, who always created the perfect work environment and transformed our small office in a pleasant place where to work daily. I am grateful to all the other members of the Section: Aitor, Renaud, Adrien, Gilles, Yaroslav, Elias, Parisa, Justine, Julian, Pascaline,

Thibaut, Michal, Conor, Louis-Hadrien, Fathi and Pablo. I also thank Mrs. Joselle Besson for her precious help with all the administrative work. Living daily in such an international environment allowed me to learn a lot about different cultures and traditions as well as it helped me to grow as a person; again, thanks to all of you!

Moving away from one's hometown poses a serious challenge on existing relationships. Only the best and strongest ones manage to survive. Hence, I am very grateful to my "Biellesi" friends, the karate set formed by Alessandro S., Alessandro V., Stefano, Camilla, and the singletons Alberto and Mirko.

Looking backwards to the path leading to this thesis, I cannot forget the friends I met in Turin and in Sweden. Even though now we are all scattered around the world, I want to express my gratitude to Luca, Giuseppe, Sirio, Stefano, Jaka and Fabio.

My deepest gratitude goes to my family. To the cutest dog in the world Tobia. To my grandmother Francesca and my uncle Gianni; To my super athletic uncle Marzio who still defeats me at every physical challenge.
To my sister Ludovica and Hadrien. We had the fortune to spend these years living next to each other, and I thank both of you for creating a family shelter in Geneva. I am extremely grateful to my girlfriend Elisa for her love, immense support and patience. Thank you so much for everything we shared and will share. Last but not least I thank my parents who taught me the importance of education and provided me with everything I needed to accomplish this goal.

To conclude, I would like to dedicate this thesis to my grandmother. Nonna, I deeply wished you could have been here today, but I know you are watching over me from the sky.

> The idea of the resurrection of the dead, the idea that life does not end in the few years we have, the idea that the beloved ones who died are still living somehow, is a fundamental aspect of my life and of my research activity.
> If I can still study and imagine now, as I am towards the end of my academic career, it is because life is a journey in which I have to love Wisdom completely, hoping that this love will somehow continue in the after life.
>
> ENNIO DE GIORGI, ITALIAN MATHEMATICIAN

Contents

CHAPTER 1

Introduction to Domain Decomposition Methods

Divide et impera

— Latin locution.

1.1 Iterative methods and preconditioners

Let us consider a domain $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, with Lipschitz boundary, the boundary value problem in the strong form

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{1.1.1}$$

and its corresponding variational formulation

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = (f,v)_\Omega, \quad \forall v \in H_0^1(\Omega), \tag{1.1.2}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \quad \text{and} \quad (f,v)_\Omega = \int_\Omega f v. \tag{1.1.3}$$

A discretization of the strong form (1.1.1) with the finite difference method or finite volume method, or of the weak formulation (1.1.2) with the finite element methods, leads to the discrete linear system

$$A\mathbf{u} = \mathbf{f}. \tag{1.1.4}$$

In the applications we will study, the large size of A usually prevents the use of direct solvers, and thus we are interested in iterative methods. Among the iterative methods, we can distinguish two major classes: stationary iterative methods and Krylov methods. Given a matrix A, a stationary iterative method in its correction form starts from an initial guess $\mathbf{u}^0$ and computes for $n = 1,2,\ldots$

$$\mathbf{u}^n = \mathbf{u}^{n-1} + M^{-1}(\mathbf{f} - A\mathbf{u}^{n-1}), \tag{1.1.5}$$

where M is invertible and comes from the splitting $A = M - N$. The convergence of stationary iterative methods is well understood. A stationary iterative method converges if and only $\rho(M^{-1}N) < 1$, where $\rho(\cdot)$ denotes the spectral radius [153]. Due to the positive-definiteness property of A, inherited by the bilinear form $a(\cdot,\cdot)$, another standard technique to solve (1.1.4) is the conjugate gradient (CG) method [114], which is a Krylov method generating a sequence of approximations $\{\mathbf{u}^n\}_{n\geq 1}$ in the affine Krylov space $\mathbf{u}^0 + \mathscr{K}_n(A,\mathbf{r}_0)$, where $\mathscr{K}_n(A,\mathbf{r}_0) = \{\mathbf{r}_0, A\mathbf{r}_0, \dots A^{n-1}\mathbf{r}_0\}$, and $\mathbf{r}_0 := \mathbf{f} - A\mathbf{u}^0$ is the residual of the initial guess. It is known that

$$\|\mathbf{u} - \mathbf{u}^n\|_A \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^n \|\mathbf{u} - \mathbf{u}^0\|_A, \quad \text{with} \quad \kappa(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)},$$

where $\|\mathbf{u}\|_A := \mathbf{u}^\top A\mathbf{u}$ and $\kappa(A)$ is the condition number of A. As $\kappa(A)$ grows, the estimate deteriorates and CG may require more and more iterations to achieve an error less than some specific tolerance 'Tol', i.e. $\|\mathbf{u} - \mathbf{u}^n\|_A \leq$ Tol. Moreover for problems such as (1.1.1), we have $\kappa(A) \sim h^{-2}$, where h is a measure of the mesh size. Thus, the CG method is generally used in combination with a preconditioner, which means applying the CG method to the preconditioned system

$$BA\mathbf{u} = B\mathbf{f}, \tag{1.1.6}$$

where B is the preconditioner. We remark that every stationary method defines a preconditioner. Indeed, taking the limit for $n \to \infty$ in (1.1.5), we get

$$M^{-1}A\mathbf{u} = M^{-1}\mathbf{f}, \tag{1.1.7}$$

that is, M^{-1} is the preconditioner associated to the iterative method.

The aim of this chapter is to provide an introduction to stationary iterative methods and preconditioners based on domain decomposition methods. First we present a class of overlapping domain decomposition methods called Schwarz methods. The origins of Schwarz methods date back to 1870 and to Schwarz's proof of the Dirichlet Principle conjectured by Riemann [80, Section 2.1]. Around the 80s-90s, these methods received a tremendous attention and they can be cast into the subspace correction framework [159], for which an extensive theory has been developed. Then, we discuss nonoverlapping domain decomposition methods, which are sometimes called "subtructuring methods", and they can be traced back to the work of Przemieniecki in the context of structural engineering [138]. It is curious to remark that, historically, these two classes of methods developed separately, although they share the same ideas.
The first part of the chapter is standard and can be found in several textbooks, for instance [151, 139, 143, 16]. In the second part we address the concept of scalability of domain decomposition methods and we discuss the results presented in [28] as well as new calculations for domain decomposition methods applied to discrete fracture networks.

1.2 Overlapping Domain Decomposition Methods

To introduce the oldest domain decomposition method, i.e. the alternating Schwarz method (ASM), we decompose Ω into two nonoverlapping subdomains Ω_1 and Ω_2, such that

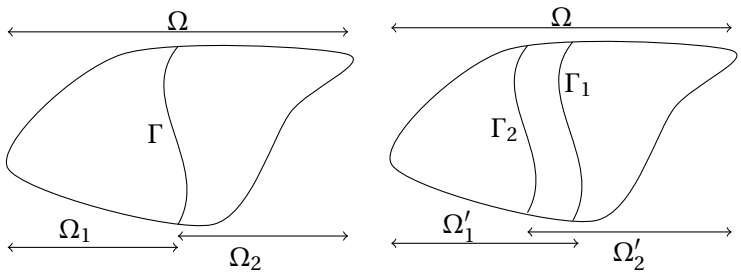


Figure 1.1: Example of a decomposition of a domain Ω into nonoverlapping subdomains (left) and into overlapping subdomains (right).

$\overline{\Omega} = \overline{\Omega_1 \cup \Omega_2}$, $\Omega_1 \cap \Omega_2 = \emptyset$, and $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$. These two nonoverlapping subdomains are enlarged to create overlapping subdomains Ω'_1, Ω'_2, such that $\Omega = \Omega'_1 \cup \Omega'_2$, $\Omega'_1 \cap \Omega'_2 \neq \emptyset$ and $\Gamma_j = \partial\Omega'_j \cap \Omega'_{3-j}$, $j = 1,2$, see Figure 1.1 for a graphical representation. Starting from an initial guess u^0 which vanishes on $\partial\Omega$, the ASM computes

$$\begin{aligned} -\Delta u^{n+1/2} &= f && \text{in } \Omega'_1, & -\Delta u^{n+1} &= f && \text{in } \Omega'_2, \\ u^{n+1/2} &= u^n && \text{on } \partial\Omega'_1, & u^{n+1} &= u^{n+1/2} && \text{on } \partial\Omega'_2, \\ u^{n+1/2} &= u^n && \text{in } \Omega'_2 \setminus \overline{\Omega'}_1, & u^{n+1} &= u^{n+1/2} && \text{in } \Omega'_1 \setminus \overline{\Omega'}_2. \end{aligned} \tag{1.2.1}$$

The first proof of convergence of the ASM has been given by Schwarz in [142] while proving the Dirichlet principle. To take advantage of the first parallel computers, Lions proposed a parallel versions, called parallel Schwarz method (PSM) [125], which computes the approximations u_1^n and u_2^n such that

$$\begin{aligned} -\Delta u_1^n &= f \quad \text{in } \Omega'_1, \qquad u_1^n = u_2^{n-1} \text{ on } \Gamma_1, \\ -\Delta u_2^n &= f \quad \text{in } \Omega'_2, \qquad u_2^n = u_1^{n-1} \text{ on } \Gamma_2. \end{aligned} \tag{1.2.2}$$

Both the ASM and the PSM are defined at the continuous level. To introduce equivalent discrete methods, we partition the unknowns in $\mathbf{u}$ into those interior to Ω'_1, $\mathbf{u} = [\mathbf{u}_1, \times]$, and those interior to Ω'_2, $\mathbf{u} = [\times, \mathbf{u}_2]$. We define the restriction matrices R_j such that $R_j\mathbf{u} = \mathbf{u}_j$, $j = 1,2$. Then it can be shown, [75, Theorem 3.3], that the correct discretization of (1.2.1) is

$$\begin{aligned} \mathbf{u}^{n+1/2} &= \mathbf{u}^n + R_1^\top A_1^{-1} R_1(\mathbf{f} - A\mathbf{u}^{n-1}), \\ \mathbf{u}^{n+1} &= \mathbf{u}^{n+1/2} + R_2^\top A_2^{-1} R_2(\mathbf{f} - A\mathbf{u}^{n+1/2}), \end{aligned} \tag{1.2.3}$$

where $A_j = R_j A R_j^\top$. At the discrete level, a very popular method which is tightly linked with the PSM is the Additive Schwarz method (AS) whose iteration reads

$$\mathbf{u}^n = \mathbf{u}^{n-1} + \sum_{j=1}^{2} R_j^\top A_j^{-1} R_j(\mathbf{f} - A\mathbf{u}). \tag{1.2.4}$$

The AS method does not convergence as iterative method since it counts twice the contributions in the overlap. However, if used as preconditioner, it permits to preserve the

symmetry of the matrix A, thus allowing one to use the conjugate gradient algorithm. It is curious to note that the correct discretization of the PSM corresponds to the Restricted Additive Schwarz (RAS) method, which has been found while trying to reduce the communication time in an implementation of the AS method on a computer [21], for a proof of equivalence see [69]. The RAS method reads

$$\mathbf{u}^n = \mathbf{u}^{n-1} + \sum_{j=1}^{2} \widetilde{R}_j^\top A_j^{-1} R_j (\mathbf{f} - A\mathbf{u}), \tag{1.2.5}$$

where the matrices $\widetilde{R}_j$ are associated to a nonoverlapping decomposition of the unknowns, $\mathbf{u} = [\widetilde{\mathbf{u}}_1, \widetilde{\mathbf{u}}_2]$, with $\widetilde{R}_j \mathbf{u} = \widetilde{\mathbf{u}}_j$. The only difference between RAS and AS methods lies in the use of the extension operators $\widetilde{R}_j$ in the RAS method. The RAS method, being the correct discretization of the PSM, converges as an iterative method but does not preserve the symmetry of A if used as preconditioner. However, if used in combination with GMRES it usually outperforms conjugate gradient preconditioned by the AS method [21]. To both stationary iterative methods we can assign their preconditioner. Considering now a decomposition into N overlapping subdomains, the AS and RAS preconditioners are

$$M_{AS}^{-1} = \sum_{j=1}^{N} R_j^\top A_j^{-1} R_j, \quad M_{RAS}^{-1} = \sum_{j=1}^{N} \widetilde{R}_j^\top A_j^{-1} R_j. \tag{1.2.6}$$

There is no general theory for the convergence of a Krylov method preconditioned by the RAS method. Meanwhile there is a very extensive theory for the AS method and we discuss it further in the next section.

1.2.1 Abstract Schwarz framework

In this section we compress in a nutshell, the extensive literature on the so called Additive Schwarz preconditioners. For more details, we refer the interested reader to the monographs [151, Chapters 2,3], [16, Chapter 7], [143, Chapter 5], [61, Chapter 5] and the survey article [159].

Let V be a finite dimensional vector space, V' its dual, and $A : V \to V'$ a symmetric positive definite operator. We can think of A as the operator which satisfies $\langle Au, v\rangle = a(u,v), \forall\, u, v \in V$, where $\langle\cdot,\cdot\rangle$ denotes the duality between V and V' and $a(\cdot,\cdot)$ is the symmetric and coercive bilinear form defined in (1.1.2). We introduce a set of auxiliary spaces V_j, $0 \le j \le N$, the operators $B_j : V_j \to V_j'$ and the interpolation operators $R_j^\top : V_j \to V$. We suppose that $V = \sum_{j=0}^{N} R_j^\top V_j$, i.e. every element $v \in V$ can be written as sum of terms $v_j \in V_j$, and this decomposition does not need to be unique. We should think of B_j as approximations of A on the smaller space V_j. Then, an abstract additive Schwarz preconditioner for A can be defined as

$$B = \sum_{j=0}^{N} R_j^\top B_j^{-1} R_j, \quad B : V' \to V. \tag{1.2.7}$$

It can be proven that B is a symmetric positive definite operator, see [16, Theorem 7.1.11], and BA is symmetric positive definite with respect the scalar product $\langle B^{-1}\cdot,\cdot\rangle$. There is a great freedom in the choice of the auxiliary spaces V_j and of the operators B_j. For instance, the auxiliary spaces can be defined through a domain decomposition, or from several discretizations of Ω with different mesh sizes. The operators B_j can be defined as the restriction of A onto the subspaces V_j, or they can be constructed through different scalar products. Such freedom results in a wide variety of different preconditioners. We briefly introduce three of them and we focus more on the two-level additive Schwarz preconditioner.

1.2.1.1 Hierarchical Basis preconditioner

In the Hierarchical basis preconditioner [161] one considers a set of mesh $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N$, such that $\mathcal{T}_j$ is obtained by regular subdivision of $\mathcal{T}_{j-1}$, with $h_j = 2^{N-j} h_N$, where $h_k := \max_{T\in\mathcal{T}_k} \mathrm{diam} T$ for $k = 1, ..., N$ is the maximum diameter of a triangle on the mesh $\mathcal{T}_k$. The unknowns on each mesh $\mathcal{T}_j$ form a space W_j. The auxiliary spaces V_j are then defined as subspaces of W_j,

$$V_j := \{\, v \in W_j : v(p) = 0 \text{ for all vertices p of } \mathcal{T}_{j-1}\} \,.$$

If we assume $W_1 = V_1$, then it holds $W_j = W_{j-1} + V_j$, and thus $W_N = V_1 \oplus \cdots \oplus V_N$. On each auxiliary space V_j, the operators B_j^{HB} are defined as

$$\langle B_j^{HB} v_1, v_2\rangle = \sum_{p\in\mathcal{V}_j\setminus\mathcal{V}_{j-1}} v_1(p) v_2(p), \quad \forall\, v_1, v_2 \in V_j,$$

where $\mathcal{V}_J$ denotes the set of vertices of $\mathcal{T}_j$. The extension operators $R_j^\top$ are straightforward injections operators between V_j and V_N. The Hierarchical basis preconditioner is then equal to $B^{\mathrm{HB}} := \sum_{j=0}^N R_j^\top \left(B_j^{\mathrm{HB}}\right)^{-1} R_j$.

1.2.1.2 BPX preconditioner

The BPX preconditioner has been introduced by Bramble, Pasciak and Xu in [14]. The auxiliary spaces V_j, $0 \le j \le N$, are defined as the finite element spaces on the meshes $\mathcal{T}_j$ and the operators B_j^{BPX} are defined as

$$\langle B_j^{\mathrm{BPX}} v_1, v_2\rangle = \sum_{p\in\mathcal{V}_j} v_1(p) v_2(p), \quad \forall\, v_1, v_2 \in V_j.$$

The extension operators $R_j^\top$ are straightforward injections operators between V_j and V_N. The global BPX preconditioner is then $B^{\mathrm{BPX}} := \sum_{j=0}^N R_j^\top \left(B_j^{\mathrm{BPX}}\right)^{-1} R_j$.

1.2.1.3 Two-Level additive Schwarz preconditioner

In the two-level additive Schwarz preconditioner, we consider a domain Ω discretized with a mesh $\mathcal{T}_h$ and a corresponding finite element space V. We divide Ω into a collection

of open subsets $\Omega'_1, \Omega'_2, ..., \Omega'_N$, whose diameters have order H. The boundaries of Ω'_j are aligned with the fine mesh $\mathcal{T}_h$ for every j. We assume there exist nonnegative $C^\infty(\Omega)$ functions θ_j, $j = 1, ..., N$ (their existence can be proven under suitable assumptions on the subdomains Ω'_j, see [151, Lemma 3.4]) which satisfy

$$\begin{aligned} &\theta_j = 0 \quad \text{on } \Omega \setminus \Omega'_j, \\ &\sum_{j=1}^{N} \theta_j = 1 \quad \text{on } \Omega, \\ &\text{There exists a positive constant } \delta \text{ such that } \|\nabla \theta_j\|_{L^\infty(\Omega)} \leq C/\delta. \end{aligned} \tag{1.2.8}$$

These hypotheses imply that $\left\{\Omega'_j\right\}_j$ form an overlapping decomposition of Ω, and δ is a measure of the overlap. On each subdomain, we introduce the space $V_j = \{\, v \in V : v = 0 \text{ on } \Omega \setminus \Omega'_j\}$ and an operator $B_j^{\mathrm{AD}} : V_j \to V'_j$ such that $\langle B_j^{\mathrm{AD}} u_j, v_j \rangle = b_j(u_j, v_j)$, where $b_j : V_j \times V_j \to \mathcal{R}$ is a coercive and symmetric bilinear form which approximates $a(\cdot,\cdot)$ on the subspace V_j. One can choose the bilinear forms $b_j(\cdot,\cdot)$ to be 'exact local solvers', which means defining them through the 'exact' bilinear form $a(\cdot,\cdot)$

$$b_j(u_j, v_j) = a(R_j^\top u_j, R_j^\top v_j), \quad \forall u_j, v_j \in V_j. \tag{1.2.9}$$

In this case we would have $B_j^{\mathrm{AD}} = R_j A R_j^\top$.
Besides the subdomain spaces V_j, we further consider a coarse discretization of Ω, resulting in the coarse mesh $\mathcal{T}_0$ and the coarse finite element space V_0. Then the two-level additive Schwarz preconditioner is

$$P_{ad} = B^{\mathrm{AD}} A, \quad \text{where} \quad B^{\mathrm{AD}} := \sum_{j=0}^{N} R_j^\top \left(B_j^{\mathrm{AD}}\right)^{-1} R_j. \tag{1.2.10}$$

We remark that B^{AD} corresponds exactly to the AS preconditioner M_{AS}^{-1} defined in (1.2.6), if one uses exact solvers on each subspace V_j and neglects the coarse solver indexed by $j = 0$.

It is interesting to observe that the preconditioned operator P_{ad} can be written as a sum of projection operators. To see this, for $j = 0, ..., N$, we introduce the operators $P_j : V \to R_j^\top V_j \subset V$, $P_j := R_j^\top \widetilde{P}_j$, where $\widetilde{P}_j : V \to V_j$ are defined by

$$b_j(\widetilde{P}_j u, v_i) = a(u, R_j^\top v_j), \quad \forall v_j \in V_j. \tag{1.2.11}$$

If we use exact solvers, it follows immediately that

$$a(P_j u, R_j^\top v_j) = a(R_j^\top \widetilde{P}_j u, R_j^\top v_j) = b_j(\widetilde{P}_j u, R_j^\top v_j) = a(u, R_j^\top v_j), \quad \forall v_j \in V_j.$$

The following Lemma holds.

Lemma 1.2.1 (Lemma 2.1 in [151])**.** *The operators P_j are self adjoint with respect to the scalar product defined by $a(\cdot,\cdot)$ and positive semi-definite. Moreover it holds*

$$P_j = R_j^\top (B_j^{AD})^{-1} R_j A, \quad 0 \le j \le N,$$

and if $b_j(\cdot,\cdot)$ satisfy (1.2.9)*, then P_j is a projection and thus $P_j^2 = P_j$.*

Proof. We first show the equality $P_j = R_j^\top (B_j^{\mathrm{AD}})^{-1} R_j A$. To do so, we consider the matrix form of (1.2.11),

$$v_i^\top B_j^{\mathrm{AD}} \widetilde{P}_j u_j = v_j^\top R_j A u_j, \quad \forall u_j, v_j \in V_j.$$

Since it holds for every $u_j, v_j \in V_j$ we can write $B_j^{\mathrm{AD}} \widetilde{P}_j = R_j A$. We assume that $b_j(\cdot,\cdot)$ is coercive, thus B_j^{AD} is invertible and using the definition of P_j the result follows. To prove that P_j is self adjoint, it is sufficient to show that

$$a(P_j u, v) = v^\top A P_j u = v^\top A (R_j^\top (B_j^{\mathrm{AD}})^{-1} R_j A u) = (R_j^\top (B_j^{\mathrm{AD}})^{-1} R_j A v)^\top A u = a(u, P_j v).$$

The positive definiteness of P_i follows from the positive definiteness of B_j^{AD}, indeed,

$$a(P_j u, u) = u^\top A P_j u = u^\top A R_j^\top (B_j^{\mathrm{AD}})^{-1} R_j A u = w_j^\top (B_j^{\mathrm{AD}})^{-1} w_j \ge 0,$$

where $w_j := R_j A u$. Finally, in case we are using exact solvers, we have $B_j^{\mathrm{AD}} = R_j A R_j^\top$, and thus

$$P_j^2 = R_j^\top (R_j A R_j^\top)^{-1} R_j A R_j^\top (R_j A R_j^\top)^{-1} A = R_j^\top (R_j A R_j^\top)^{-1} A = P_j.$$

□

We close this subsection observing that the preconditioned operator P_{ad} can be written as

$$P_{ad} = B^{\mathrm{AD}} A = \sum_{j=0}^{N} R_j^\top (R_j A R_j^\top)^{-1} R_j A = \sum_{j=0}^{N} P_j.$$

1.2.1.4 Convergence theory

As discussed in the introduction, the convergence of the conjugate gradient method depends on the condition number of the matrix A. As the preconditioned system BA is self-adjoint with respect to the scalar product defined by $a(\cdot,\cdot)$, see Lemma (1.2.1), the largest and smallest eigenvalue of BA are given by the Rayleigh quotient formula as

$$\lambda_{\max}(BA) := \sup_{u \in V} \frac{a(BAu, u)}{a(u,u)}, \quad \lambda_{\min}(BA) := \inf_{u \in V} \frac{a(BAu, u)}{a(u,u)}.$$

There is a well-developed theory to find estimates for the maximum and minimum eigenvalue for additive Schwarz preconditioners and it is called abstract Schwarz framework. This theory relies on three assumptions. For every new preconditioner, if one verifies these three assumptions, then a general result permits to find estimates for the extreme eigenvalues and thus for the condition number of the preconditioned system. These three assumptions are:

Assumption 1.2.2 (Stable decomposition). There exists a constant C_0 such that every $u \in V$ admits a decomposition $u = \sum_{j=0}^{N} R_j^\top u_j$, for some $u_j \in V_j$, such that

$$\sum_{j=0}^{N} \langle B_j u_j, u_j \rangle \leq C_0^2 \langle Au, u \rangle. \tag{1.2.12}$$

Assumption 1.2.2 permits to find a lower bound for $\lambda_{\min}(BA)$. To have a robust preconditioner, it is essential that the constant C_0 does not depend strongly on some parameters of the problem such as the mesh size on the finest grid or the size/number of the subdomains. In this perspective, the choice of the coarse space V_0 plays the key role.

Assumption 1.2.3 (Strengthened Cauchy-Schwarz Inequality). There exist constants $0 \leq \epsilon_{ij} \leq 1$, $1 \leq i, j \leq N$, such that

$$|a(R_i^\top u_i, R_j^\top u_j)| \leq \epsilon_{ij}\, a(R_i^\top u_i, R_i^\top u_i)^{\frac{1}{2}}\, a(R_j^\top u_j, R_j^\top u_j)^{\frac{1}{2}}, \tag{1.2.13}$$

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

The quantity $\rho(\mathscr{E})$ will be involved in the upper bound of $\lambda_{\max}(BA)$.

Assumption 1.2.4 (Local stability). There exists a constant $\omega > 0$, such that,

$$a(R_i^\top u_i, R_i^\top u_i) \leq \omega \langle B_i u_i, u_i \rangle, \quad u_i \in V_i,\ 0 \leq i \leq N. \tag{1.2.14}$$

Assumption 1.2.4 guarantees that the bilinear forms induced by B_i are coercive. If all these assumptions are verified it is possible to prove the following theorem.

Theorem 1.2.5 (Theorem 2.7 in [151])**.** *Let Assumptions* (1.2.2)-(1.2.3)-(1.2.4) *be satisfied. Then the condition number of the additive Schwarz operator satisfies*

$$\kappa(BA) \leq C_0^2 \omega(\rho(\mathscr{E}) + 1). \tag{1.2.15}$$

Of course, different methods will lead to different values of these parameters. We conclude this section reporting some classical bounds for the condition numbers of the HB, BPX and two-level AS preconditioners,

$$\kappa(B^{\mathrm{HB}}A) \leq C_1(1 + |\ln h_N|^2), \quad \kappa(B^{\mathrm{BPX}}A) \leq C_2, \quad \kappa(B^{\mathrm{AD}}A) \leq C_3\left(1 + \frac{H}{\delta}\right),$$

where h_N is the finest mesh size. For more details, we refer the reader to [16, Chapter 7] and [159] for BPX and HB preconditioners, and Chapter 2 and 3 in the dedicated monograph [151] for the two-level AS preconditioner.

1.3 Nonoverlapping Domain Decomposition Methods

Let us suppose that Ω is decomposed into two nonoverlapping subdomains $\overline{\Omega} = \overline{\Omega_1 \cup \Omega_2}$, with $\Omega_1 \cap \Omega = \emptyset$ and $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$. Problem (1.1.1) can be reformulated into the following system

$$\begin{aligned} -\Delta u_1 &= f && \text{in } \Omega_1, \quad u_1 = 0 \quad \text{on } \partial\Omega_1 \setminus \Gamma, \\ -\Delta u_2 &= f && \text{in } \Omega_2, \quad u_2 = 0 \quad \text{on } \partial\Omega_2 \setminus \Gamma, \\ u_1 &= u_2 && \text{on } \Gamma, \\ \frac{\partial u_1}{\partial \mathbf{n}_1} &= -\frac{\partial u_2}{\partial \mathbf{n}_2} && \text{on } \Gamma. \end{aligned} \tag{1.3.1}$$

We define the standard Sobolev spaces $V_j := H^1(\Omega_j)$, $V_j^0 := \left\{ v \in V_j : v = 0 \text{ on } \partial\Omega_j \setminus \Gamma \right\}$, $\Lambda := H_{00}^{\frac{1}{2}}(\Gamma)$ as the space of traces of functions which lie in V_j^0 [139] and the bilinear forms $a_j(u_j, v_j) := \int_{\Omega_j} \nabla u_j \nabla v_j, \forall u_j, v_j \in V_j^0$, $j = 1,2$. The weak formulation of (1.3.1) corresponds to

$$\begin{aligned} & a_1(u_1, v_1) = (f, v_1)_{\Omega_1} \quad \forall v_1 \in V_1^0, \\ & a_2(u_2, v_2) = (f, v_2)_{\Omega_2} \quad \forall v_2 \in V_2^0, \\ & u_1 = u_2 \quad \text{on } \Gamma, \\ & a_1(u_1, \mathscr{E}_1\eta) - (f, \mathscr{E}_1\eta)_{\Omega_1} = (f, \mathscr{E}_2\eta)_{\Omega_2} - a_2(u_2, \mathscr{E}_2\eta), \forall \eta \in \Lambda, \end{aligned} \tag{1.3.2}$$

where $\mathscr{E}_j$ are continuous extension operators from Λ to V_j. Equation $(1.3.2)_4$ is the correct variational discretization of $(1.3.1)_4$. Indeed, at the weak level we have to impose the continuity of the normal derivatives not strongly, but weakly in the functional sense, that is $\frac{\partial u_1}{\partial \mathbf{n}_1} = -\frac{\partial u_2}{\partial \mathbf{n}_2}$ in Λ'. Therefore we ask that

$$\left\langle \frac{\partial u_1}{\partial \mathbf{n}_1}, \eta \right\rangle = -\left\langle \frac{\partial u_2}{\partial \mathbf{n}_2}, \eta \right\rangle, \quad \forall \eta \in \Lambda.$$

Using integration by parts we obtain immediately

$$\left\langle \frac{\partial u_i}{\partial \mathbf{n}_i}, \eta \right\rangle = a_i(u_i, \mathscr{E}_i\eta) - (f, \mathscr{E}_i\eta)_{\Omega_i}, \quad i = 1,2, \tag{1.3.3}$$

and therefore we recover $(1.3.2)_4$. For a complete proof of the equivalence between (1.1.2) and (1.3.2) we refer to Lemma 1.2.1 in [139].

It is interesting that we can derive an equation involving only a variable defined on the interface Γ. To see this, we consider a finite element discretization of the weak formulation 1.1.2 which leads to the linear system

$$\begin{pmatrix} A_{II}^1 & 0 & A_{I\Gamma}^1 \\ 0 & A_{II}^2 & A_{I\Gamma}^2 \\ A_{\Gamma I}^1 & A_{\Gamma I}^2 & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_\Gamma \end{pmatrix}, \tag{1.3.4}$$

where we divided the degrees of freedom into those internal to the subdomains 'I', and those on the interface 'Γ', so that $\mathbf{u}_i$ contains the degrees of freedom inside Ω_i and $\mathbf{u}_\Gamma$

contains those on Γ. We remark that $A_{\Gamma\Gamma}$ can be split into two contributions, $A_{\Gamma\Gamma} = A^1_{\Gamma\Gamma} + A^2_{\Gamma\Gamma}$ by partitioning the contribution of the bilinear form over Ω_1 and Ω_2. Similarly, $\mathbf{f}_\Gamma = \mathbf{f}^1_\Gamma + \mathbf{f}^2_\Gamma$. Expressing $\mathbf{u}_i$ in terms of $\mathbf{u}_\Gamma$ from the first two equations, i.e. $\mathbf{u}_i = (A^i_{II})^{-1}(\mathbf{f}_i - A^i_{I\Gamma}\mathbf{u}_\Gamma)$ and substituting into the third one, we obtain the Schur complement linear system

$$\Sigma \mathbf{u}_\Gamma = \mu, \tag{1.3.5}$$

where $\Sigma := \Sigma_1 + \Sigma_2$ is called Schur complement, the local Schur complements Σ_i are defined as $\Sigma_i := A^i_{\Gamma\Gamma} - A^i_{\Gamma I}(A^i_{II})^{-1}A^i_{I\Gamma}$ and $\mu = \mu_1 + \mu_2$, with $\mu_i := \mathbf{f}^i_\Gamma - A^i_{\Gamma I}(A^i_{II})^{-1}\mathbf{f}_i$, $i = 1,2$. It is very useful to interpret the action of the local Schur complements on a given vector. Given a vector $\mathbf{u}_\Gamma$, $\mathbf{u}^0_i := -(A^i_{II})^{-1}A^i_{I\Gamma}\mathbf{u}_\Gamma$ represents the harmonic extension on Ω_i of the Dirichlet data $\mathbf{u}_\Gamma$ (see first equation in (1.3.4) assuming that $f = 0$). The term $\mathbf{u}^f_i := (A^i_{II})^{-1}\mathbf{f}_i$ corresponds instead to the solution of a homogeneous Dirichlet problem posed in Ω_i with force term equal to $\mathbf{f}_i$. We conclude that $\Sigma_i\mathbf{u}_\Gamma = A^i_{\Gamma\Gamma}\mathbf{u}_\Gamma + A^i_{\Gamma I}\mathbf{u}^0_i$ corresponds to the discretization of $a_i(u^0_{i,h}, \mathscr{E}_i\phi_\Gamma)$, where $u^0_{i,h}$ is the finite element function corresponding to the degrees of freedom stored in $\mathbf{u}^0_i$ and ϕ_Γ is a finite element function in the finite element space $\Lambda_h \subset \Lambda$. Similarly, $A^i_{\Gamma\Gamma}\mathbf{u}_\Gamma + A^i_{\Gamma I}(\mathbf{u}^0_i + \mathbf{u}^f_i)$ is exactly the discretization of $a_1(u_{1,h}, \mathscr{E}_i\phi_{i,\Gamma})$.

1.3.1 The Steklov-Poincaré operator

We derived equation (1.3.5) at the algebraic level, working exclusively with the matrices in (1.3.4). However it is possible to obtain an analogue of (1.3.5) at the continuous level. To see this, we need to introduce some operators. We define the extension operators $\mathscr{H}_i : \Lambda \to V_i$ through the relations

$$\begin{aligned} &a_i(\mathscr{H}_i\eta, v_i) = 0, \quad \forall v_i \in V^0_i, \\ &\mathscr{H}_i\eta = \eta \quad \text{on } \Gamma. \end{aligned} \tag{1.3.6}$$

In other words, $\mathscr{H}_i$ takes a function defined on the interface and it returns its harmonic extension. We remark that the operator is well defined since for any $\eta \in \Lambda$, equation (1.3.6) has a unique solution thanks to the properties of $a_i(\cdot,\cdot)$. We now set $\lambda := u_{|\Gamma}$. Due to linearity, each u_i can be written as $u_i = \mathscr{H}_i\lambda + \mathscr{G}_i f_i$, where $\mathscr{G}_i f_i$ is the solution of

$$\begin{aligned} &a_i(\mathscr{G}_i f_i, v_i) = f, \qquad \forall v_i \in H^1_0(\Omega_i), \\ &\mathscr{G}_i f_i = 0, \qquad\qquad\qquad \text{on } \partial\Omega_i. \end{aligned} \tag{1.3.7}$$

In order for u_i to be solutions of (1.3.1) we need to impose $\frac{\partial u_1}{\partial \mathbf{n}_1} = -\frac{\partial u_2}{\partial \mathbf{n}_2}$. Inserting the decomposition of u_i, we obtain

$$\mathscr{S}\lambda = \mu, \tag{1.3.8}$$

where $\mu := -\sum_{i=1}^2 \frac{\partial}{\partial \mathbf{n}_i}\mathscr{G}_i f_i$ and $\mathscr{S}\lambda := \mathscr{S}_1 + \mathscr{S}_2 = \sum_{i=1}^2 \frac{\partial}{\partial \mathbf{n}_i}\mathscr{H}_i\lambda$. The operator $\mathscr{S}$ is called Steklov-Poincaré operator. We invite the reader to compare (1.3.5) and (1.3.8). It is possible to prove that (1.3.5) is exactly the finite element approximation of (1.3.8), see [139, Chapter 2.3]. We now provide a formal definition of the Steklov-Poincaré operator.

Definition 1.3.1. Consider a domain Ω_i with Lipschitz boundary, a connected subset Γ of $\partial\Omega_i$ and a sufficiently regular Sobolev space $\Lambda(\Gamma)$ such that $\mathscr{H}_i$ is well defined. The Steklov-Poincaré operator $\mathscr{S} : \Lambda(\Gamma) \to \Lambda'(\Gamma)$ is linear and has the following variational representation

$$_{\Lambda'}\langle \mathscr{S}\eta, \mu\rangle_{\Lambda} := a_i(\mathscr{H}_i\eta, \mathscr{E}_i\mu).$$

We stress that the extension operator $\mathscr{E}_i$ can be chosen arbitrarly. The Steklov-Poincaré operator plays a key role in the convergence of nonoverlapping domain decomposition methods. We will show this in details in the next subsections and chapters. In some context, the Steklov-Poincaré operator is also called Dirichlet to Neumann operator, since it takes a function defined on the interface, it extends it harmonically inside Ω_i and then it computes the normal derivative. The operators $\mathscr{S}_i$ enjoy some nice properties derived from the bilinear form $a_i(\cdot,\cdot)$.

Lemma 1.3.2. *$\mathscr{S}_i$ is a continous, symmetric and positive definite operator.*

Proof. Since $\mathscr{E}_i$ can be chosen arbitrarly, we set $\mathscr{E}_i\mu = \mathscr{H}_i\mu$. Then we have using the symmetry of $a_i(\cdot,\cdot)$,

$$_{\Lambda'}\langle \mathscr{S}_i\eta, \mu\rangle_{\Lambda} = a_i(\mathscr{H}_i\eta, \mathscr{H}_i\mu) = a_i(\mathscr{H}_i\mu, \mathscr{H}_i\eta) =_{\Lambda'} \langle \eta, \mathscr{S}_i\mu\rangle_{\Lambda}.$$

Moreover using the continuity of $a_i(\cdot,\cdot)$ and the continuous dependence of $\mathscr{H}_i$ on the boundary data,

$$_{\Lambda'}\langle \mathscr{S}_i\eta, \mu\rangle_{\Lambda} = a_i(\mathscr{H}_i\eta, \mathscr{H}_i\mu) \le C|H_i\eta|_1|H_i\mu|_1 \le C\|\eta\|_{\Lambda}\|\mu\|_{\Lambda}.$$

Using the coercivity of $a_i(\cdot,\cdot)$ and the trace inequality between Λ and V_i^0,

$$_{\Lambda'}\langle \mathscr{S}_i\eta, \eta\rangle_{\Lambda} = a_i(\mathscr{H}_i\eta, \mathscr{H}_i\eta) > \alpha|H_i\eta|_1^2 > \alpha C\|\eta\|_{\Lambda}^2.$$

□

1.3.2 Dirichlet-Neumann method

The system formulation (1.3.1) paves the way to the definition of several domain decomposition methods. These methods are based on a relaxation of the transmission conditions $(1.3.1)_{3,4}$, generating a sequence of steps where only one of the two transmission conditions, or a combination of them is satisfied. The Dirichlet-Neumann method (DNM) is made of two sequential steps: given an interface data $\mathbf{u}_\Gamma^n$, first we solve a Dirichlet problem in Ω_1, then we compute the normal derivative along Γ and we impose it as Neumann boundary condition for the problem set in Ω_2. Finally we update with a weighted combination the new Dirichlet trace u_Γ^{n+1}. Given an initial guess u_Γ^0, this procedure corresponds

mathematically for $n \geq 0$ to

$$\begin{aligned}
-\Delta u_1^{n+1} &= f && \text{in } \Omega_1, \quad u_1^{n+1} = 0 \quad \text{on } \partial\Omega_1 \setminus \Gamma, \\
u_1^{n+1} &= u_\Gamma^n && \text{on } \Gamma, \\
-\Delta u_2^{n+1} &= f && \text{in } \Omega_2, \quad u_2^{n+1} = 0 \quad \text{on } \partial\Omega_2 \setminus \Gamma, \\
\frac{\partial u_2^{n+1}}{\partial \mathbf{n_2}} &= -\frac{\partial u_1^{n+1}}{\partial \mathbf{n_1}}, && \text{on } \Gamma, \\
u_\Gamma^{n+1} &= \theta u_2^{n+1} + (1-\theta) u_\Gamma^n && \text{on } \Gamma.
\end{aligned} \tag{1.3.9}$$

The choice of the parameter θ, $0 < \theta < 1$, influences the convergence properties of the method. We can give an algebraic formulation of (1.3.9). System (1.3.9) can be rewritten as

$$\begin{aligned}
& A_{II}^1 \mathbf{u}_1^{n+1} + A_{I\Gamma}^1 \mathbf{u}_\Gamma^n = \mathbf{f}_1, && \text{Dirichlet problem in } \Omega_1, \\
& \begin{pmatrix} A_{II}^2 & A_{I\Gamma}^2 \\ A_{\Gamma I}^2 & A_{\Gamma\Gamma}^2 \end{pmatrix} \begin{pmatrix} \mathbf{u}_2^{n+1} \\ \mathbf{u}_{2,\Gamma}^{n+1} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_2 \\ \mathbf{f}_\Gamma^2 + \mathbf{f}_\Gamma^1 - A_{I\Gamma}^1 \mathbf{u}_1^{n+1} - A_{\Gamma\Gamma}^1 \mathbf{u}_\Gamma^n \end{pmatrix}, && \text{Neumann problem in } \Omega_2, \\
& \mathbf{u}_\Gamma^{n+1} = \theta \mathbf{u}_{2,\Gamma}^{n+1} + (1-\theta) \mathbf{u}_\Gamma^n, && \text{Update step.}
\end{aligned} \tag{1.3.10}$$

The DNM described by algorithm (1.3.10) can be reformulated as a Richardson iteration to solve the Schur complement equation (1.3.5). To show this, we eliminate $\mathbf{u}_1^{n+1}$ from the right hand side of the Neumann problem,

$$\mathbf{f}_\Gamma^2 + \mathbf{f}_\Gamma^1 - A_{I\Gamma}^1 \mathbf{u}_1^{n+1} - A_{\Gamma\Gamma}^1 \mathbf{u}_\Gamma^n = \mathbf{f}_\Gamma^2 + \mathbf{f}_\Gamma^1 - (A_{\Gamma\Gamma}^1 \mathbf{u}_\Gamma^n - A_{I\Gamma}^1 (A_{\Gamma\Gamma}^1)^{-1} A_{\Gamma I}^1 \mathbf{u}_\Gamma^n) - A_{I\Gamma}^1 (A_{\Gamma\Gamma}^1)^{-1} \mathbf{f}^1 = \mathbf{f}_\Gamma^2 + \mu_1 - \Sigma_1 \mathbf{u}_\Gamma^n.$$

Eliminating now $\mathbf{u}_2^{n+1}$, we obtain

$$A_{\Gamma I}^2 \mathbf{u}_2^{n+1} + A_{\Gamma\Gamma}^2 \mathbf{u}_{2,\Gamma}^{n+1} = (A_{\Gamma\Gamma}^2 - A_{\Gamma I}^2 (A_{II}^2)^{-1} A_{I\Gamma}^2) \mathbf{u}_{2,\Gamma}^{n+1} + A_{\Gamma I}^2 (A_{II}^2)^{-1} \mathbf{f}_2 = \Sigma_2 \mathbf{u}_{2,\Gamma}^{n+1} + A_{\Gamma I}^2 (A_{II}^2)^{-1} \mathbf{f}_2.$$

Combining these two results we obtain the iteration

$$\Sigma_2 \mathbf{u}_{2,\Gamma}^{n+1} = \mu - \Sigma_1 \mathbf{u}_\Gamma^n,$$

which, plugged into the update rule, leads to

$$\Sigma_2 (\mathbf{u}^{n+1} - \mathbf{u}^n) = \theta (\mu - \Sigma \mathbf{u}^n), \tag{1.3.11}$$

which is a Richardson method to solve (1.3.5) with preconditioner Σ_2. It has been proven that for the two subdomain case, Σ_2 is an optimal preconditioner for the Schur complement system, i.e. $\kappa(\Sigma_2^{-1}\Sigma) \leq C$, where C is independent on h, see [151, Chapter 4] and [139, Chapter 2]. This implies that the condition number of the preconditioned Schur complement does not grow when the size of the finite element spaces grows. Concerning the choice of the relaxation parameter θ, for two symmetric subdomains, $\theta = 0.5$ leads to a direct method which converges in two iterations. For more than two subdomains, there are still some choices which lead to nilpotent iterations. We refer the interested reader to [32]. There are several versions of the DNM when the decomposition involves several

subdomains, since there is a certain freedom in deciding where to impose Dirichlet or Neumann boundary conditions. In Section 1.4, we analyse one of these versions for a particular geometry. We conclude this section by observing that the method we introduced is inherently a sequential method. A parallel version is easily obtained by replacing $(1.3.9)_4$ with $\frac{\partial u_2^{n+1}}{\partial \mathbf{n_2}} = -\frac{\partial u_1^n}{\partial \mathbf{n_1}}$. For two subdomains there is no advantage since the parallel version generates a sequences of approximations $\{u_i^n\}_{n\geq 1}$ $i = 1,2$, which have as subsequences the sequential approximations. For general decompositions in many subdomains, parallel and sequential versions do not preserve this relation, and indeed parallelization is needed in order to take advantage of modern high performance clusters.

1.3.3 Neumann-Neumann method

For a decomposition into two subdomains, the Neumann-Neumann method (NNM) is made of two sequential steps. Given an interface data u_Γ^n, we solve Dirichlet problems both in Ω_1 and Ω_2 imposing u_Γ^n along the interface. The two local solutions will now be continous, satisfying $(1.3.1)_3$, but their normal derivatives will not be continous in general. Thus, we compute the jump of the two normal derivatives along Γ and we solve two Neumann problems to compute local corrections. Eventually we update the new interface data u_Γ^{n+1}. In terms of differential operators, the algorithm is

$$
\begin{aligned}
-\Delta u_i^{n+1} &= f, && \text{in } \Omega_i, \quad u_i^{n+1} = 0 \quad \text{on } \partial\Omega_i \setminus \Gamma, \ i = 1,2,\\
u_i^{n+1} &= u_\Gamma^n, && \text{on } \Gamma, \ i = 1,2,\\
-\Delta \psi_i^{n+1} &= 0, && \text{in } \Omega_i, \quad \psi_i^{n+1} = 0 \quad \text{on } \partial\Omega_i \setminus \Gamma, \ i = 1,2,\\
\frac{\partial \psi_i^{n+1}}{\partial \mathbf{n_i}} &= \frac{\partial u_1^{n+1}}{\partial \mathbf{n}_1} + \frac{\partial u_2^{n+1}}{\partial \mathbf{n}_2}, && \text{on } \Gamma, \ i = 1,2,\\
u_\Gamma^{n+1} &= u_\Gamma^n - \theta(\psi_1^{n+1} + \psi_2^{n+1}), && \text{on } \Gamma.
\end{aligned}
\tag{1.3.12}
$$

The algebraic formulation of (1.3.12) corresponds to

$$
\begin{aligned}
&A_{II}^i \mathbf{u}_i^{n+1} + A_{I\Gamma}^i \mathbf{u}_\Gamma^n = \mathbf{f}_i, \ i = 1,2, \quad \text{Dirichlet problem in } \Omega_i,\\
&\begin{pmatrix} A_{II}^i & A_{I\Gamma}^i \\ A_{\Gamma I}^i & A_{\Gamma\Gamma}^i \end{pmatrix} \begin{pmatrix} \psi_i^{n+1} \\ \psi_{i,\Gamma}^{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{g}_\Gamma^n \end{pmatrix}, \quad \text{Neumann problem in } \Omega_i,\\
&\mathbf{u}_\Gamma^{n+1} = \mathbf{u}^n - \theta(\psi_{1,\Gamma}^{n+1} + \psi_{2,\Gamma}^{n+1}) \quad \text{Update step},
\end{aligned}
\tag{1.3.13}
$$

where $\mathbf{g}_\Gamma^n = \sum_{i=1}^2 A_{\Gamma I}^i \mathbf{u}_i^{n+1} + A_{\Gamma\Gamma}^i \mathbf{u}_\Gamma^n - \mathbf{f}_\Gamma^i$. Eliminating $\mathbf{u}_i^{n+1}$, $i = 1,2$ from the expression of $\mathbf{g}_\Gamma^n$, we obtain $\mathbf{g}_\Gamma^n = \Sigma \mathbf{u}_\Gamma^n - \mu$. We consider now the Neumann problems and we eliminate the interior unknowns ψ_i^{n+1} to obtain $\psi_{i,\Gamma}^{n+1} = (\Sigma_i)^{-1}\mathbf{g}_\Gamma^n = (\Sigma_i)^{-1}(\Sigma \mathbf{u}_\Gamma^n - \mu)$. Inserting these expressions into the update step, we obtain

$$
\mathbf{u}_\Gamma^{n+1} - \mathbf{u}_\Gamma^n = \theta(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu - \Sigma \mathbf{u}_\Gamma^n),
$$

which shows that the NNM is a Richardson method with preconditioner $\Sigma_1^{-1} + \Sigma_2^{-1}$. It can be shown that the NNM is an optimal preconditioner for the two subdomain case, i.e.

$\kappa((\Sigma_1^{-1}+\Sigma_2^{-1})\Sigma) \le C$ see [151, Chapter 4] and [139, Chapter 2]. For several subdomains, this method is very sensitive to the choice of the relaxation parameter θ. There are several cases in which the NNM diverges. We refer to the PhD thesis [26]. The method can be parallelized by replacing $(1.3.12)_5$ with $\frac{\partial \psi_i^{n+1}}{\partial \mathbf{n_i}} = \frac{\partial u_1^n}{\partial \mathbf{n}_1} + \frac{\partial u_2^n}{\partial \mathbf{n}_2}$. For the definition of the NN method for many subdomain decompositions, we refer to Section 1.4

1.3.4 Optimized Schwarz methods

In this paragraph we introduce optimized Schwarz methods (OSMs) which are the domain decomposition methods we will discuss the most throughout the thesis. The origins of OSMs trace back to [126] where Lions defined a first version of the algorithm and he also proposed a convergence analysis based on energy estimates, see [75, Section 5] for an historical review of other contributions. The method can be viewed as a generalization of the parallel Schwarz method in case of nonoverlapping subdomains, where instead of exchanging Dirichlet data, Robin or more general boundary conditions are imposed along the nonoverlapping interfaces. Given two initial guesses u_i^0, $i = 1,2$, the method defined in [126] computes for $n \ge 1$

$$\begin{aligned}
-\Delta u_1^{n+1} &= f && \text{in } \Omega_1, \quad u_1^{n+1} = 0 \quad \text{on } \partial\Omega_1 \setminus \Gamma,\\
\frac{\partial u_1^{n+1}}{\partial \mathbf{n}_1} + s_1 u_1^{n+1} &= \frac{\partial u_2^n}{\partial \mathbf{n}_1} + s_1 u_2^n, && \text{on } \Gamma,\\
-\Delta u_2^{n+1} &= f && \text{in } \Omega_2, \quad u_2^{n+1} = 0 \quad \text{on } \partial\Omega_2 \setminus \Gamma,\\
\frac{\partial u_2^{n+1}}{\partial \mathbf{n}_2} + s_2 u_2^{n+1} &= \frac{\partial u_1^{n+1}}{\partial \mathbf{n}_2} + s_2 u_1^{n+1}, && \text{on } \Gamma,
\end{aligned} \tag{1.3.14}$$

The variational formulation of (1.3.14) is

$$\begin{aligned}
a_1(u_1^{n+1},\mathscr{E}_1\eta) + \int_\Gamma s_1 u_1^{n+1}\eta &= \int_{\Omega_1} f\mathscr{E}_1\eta - a_2(u_2^n,\mathscr{E}_2\eta) + \int_\Gamma s_1 u_2^n\eta + \int_{\Omega_2} f\mathscr{E}_2\eta, \quad \forall \eta \in \Lambda,\\
a_2(u_2^{n+1},\mathscr{E}_2\eta) + \int_\Gamma s_2 u_2^{n+1}\eta &= \int_{\Omega_2} f\mathscr{E}_2\eta - a_1(u_1^{n+1},\mathscr{E}_1\eta) + \int_\Gamma s_2 u_1^{n+1}\eta + \int_{\Omega_1} f\mathscr{E}_1\eta, \quad \forall \eta \in \Lambda.
\end{aligned} \tag{1.3.15}$$

Already in the original paper [126], Lions stressed the importance of the acceleration parameters $s_1, s_2 \in \mathbb{R}^+$ on the rate of convergence of the algorithm. He even suggested to replace the parameters with local or nonlocal operators. These observations pushed researchers to study which are the best possible parameters/operators that lead to the fastest convergence possible. In [136], the authors showed that setting $s_1 := \mathscr{S}_1$ and $s_2 := \mathscr{S}_2$, hence using the Steklov-Poincaré operators, leads to a nilpotent method, that is the method converges in just two iterations. They actually showed that this result holds for a decomposition into many subdomains in a strip. In this case the convergence is attained in N iterations where N is the number of subdomains. A generalization of this work for arbitrary decompositions is available in [88]. However the Steklov-Poincaré operator is dense, nonlocal and expensive to compute. Given a discretization of Γ with N_s degree

of freedoms, we need to solve N_s Dirichlet problems in order to assemble the Steklov-Poincaré operator. Therefore researchers started to look for sparse and cheap approximations of these optimal operators. Some earlier work in this direction are the PhD thesis of Japhet [115] and [116]. The terminology "optimized Schwarz method" was adopted later, thanks to [74], where Gander established a solid procedure to derive optimized transmission conditions for many different problems. We provide a more comprehensive review of OSMs in Chapter 2, where we analyse this algorithm for several heterogeneous problems. In [146], it has been shown that, under some conditions, the preconditioner associated to OSMs is

$$M_{ORAS}^{-1} = \sum_{j=1}^{N} \widetilde{R}_j^\top \widetilde{A}_j^{-1} R_j,$$

where $\widetilde{A}_j$ are the finite element discretizations of the bilinear terms $a_j(u_j, v_j) + \int_\Gamma s_j u_j v_j$. A symmetrized and additive variant has been analyzed in [110].

We conclude the first part of the introduction by showing that OSMs can be seen as an alternating direction iteration (ADI) to solve the Steklov-Poincaré equation. To the best of our knownledge, this interesting point of view was first discussed in [52, Chapter 5.4]. Given the equation $\mathscr{S}\lambda = \mu$, we can decompose the operator $\mathscr{S}$ as $\mathscr{S} = S_1 + S_2$. One can choose S_1 and S_2 arbitrarily, but we will set $S_i := \mathscr{S}_i$, $i = 1,2$, equal to the local Steklov-Poincare operators. Then, given an initial guess λ^0, the ADI method with accelerating parameters s_1, s_2 to solve $\mathscr{S}\lambda = \mu$ is

$$\begin{aligned} (s_1 I + \mathscr{S}_1)\lambda^{n+\frac{1}{2}} &= (s_1 I - \mathscr{S}_2)\lambda^n + \mu, \\ (s_2 I + \mathscr{S}_2)\lambda^{n+1} &= (s_2 I - \mathscr{S}_1)\lambda^{n+\frac{1}{2}} + \mu, \end{aligned} \tag{1.3.16}$$

where I is the identity operator. This algorithm is also called Peaceman-Rachford iteration method. We refer the interested reader to [2, Chapter 7.7.3] and [153, Chapter 7]. In Section 2.5, we will discuss more about this iteration in the context of probing. In the present framework, the ADI to solve (1.3.8) is

$$\begin{aligned} \left\langle (s_1 I + \mathscr{S}_1)\lambda_1^{n+1}, \eta \right\rangle &= \left\langle \mu, \eta \right\rangle + \left\langle (s_1 I - \mathscr{S}_2)\lambda_2^n, \eta \right\rangle, \\ \left\langle (s_2 I + \mathscr{S}_2)\lambda_2^{n+1}, \eta \right\rangle &= \left\langle \mu, \eta \right\rangle + \left\langle (s_2 I - \mathscr{S}_1)\lambda_1^{n+1}, \eta \right\rangle. \end{aligned} \tag{1.3.17}$$

Interpreting λ_i as the Dirichlet traces of u_i, i.e. $\lambda_i = u_i|_\Gamma$, and assuming for simplicity that $\mu_i = 0$, $i = 1,2$, the equivalence between (1.3.16) and (1.3.15) can be understood as follows. Given an approximation λ_2^k, we compute

$$\left\langle (s_1 I - \mathscr{S}_2)\lambda_2^n, \eta \right\rangle = \int_\Gamma s_1 \lambda_2^n \eta - a_2(\mathscr{H}_2 \lambda_2^n, \mathscr{H}_2 \eta), \quad \forall \eta \in \Lambda. \tag{1.3.18}$$

We then find λ_1^{n+1} solving $(1.3.17)_1$, where substituting (1.3.18) and the definition of $\mathscr{S}_1$ leads to

$$a_1(\mathscr{H}_1 \lambda_1^{n+1}, \mathscr{H}_1 \eta) + \int_\Gamma s_1 \lambda_1^{n+1} \eta = \int_\Gamma s_1 \lambda_2^n \eta - a_2(\mathscr{H}_2 \lambda_2^n, \mathscr{H}_2 \eta), \quad \forall \eta \in \Lambda,$$

which corresponds to $(1.3.15)_1$ for $f = 0$. Thank the uniqueness of the solution of (1.3.15) we conclude that $\mathscr{H}_1\lambda_1^{n+1} = u_1^{n+1}$ and that $\lambda_1^{n+1} = u_1^{n+1}|_\Gamma$. A similar calculation concerning $(1.3.17)_2$ shows that we have $\lambda_2^{n+1} = u_2^{n+1}|_\Gamma$ and $\mathscr{H}_2\lambda_2^{n+1} = u_2^{n+1}$.

1.4 Scalability of domain decomposition methods

In the previous sections we introduced classical one-level domain decomposition methods for two subdomain decompositions. To exploit at best the large number of available cores in modern computers, decompositions into two subdomains are not attractive. Ideally, one would prefer to decompose the original domain Ω into several (hundreds, thousands, millions) of subdomains and to assign one subdomain to each core. In this way, each core would solve a small subdomain problem and hence there would be an optimal parallelization of the solution process. Thus, in view of high performance computing, an important property of a solution algorithm is the so called scalability. In the literature there are two definitions of scalability [61]. An algorithm is “strongly scalable” if the solution time of a fixed size problem is inversely proportional to the number of cores. Strong scalability is extremely hard to obtain since if one considers millions of subdomains, so that each subdomain problem becomes extremely cheap to solve, there would a bottleneck due to the communication lag to transfer data among the large number of subdomains. To mitigate the communication cost, asynchronous methods have recently gained a lot of attention [132, 27]. Thus, a more practical concept is the weak scalability. An algorithm is “weakly scalable" if the solution time is constant for a fixed ratio between the size of the problem and the number of cores. In other words, if a weakly scalable algorithm solves a problem with 100’000 unknowns in 10 seconds using 10 processors, it should be able to solve a problem with 1’000’000 unknowns in the same 10 seconds using 100 processors. In the specific context of domain decomposition methods, a domain-decomposition method is said to be weakly scalable, if its rate of convergence does not deteriorate when the number of subdomains grows (definition 1.3 in [151]). Unfortunately, one-level domain decomposition methods are in general not weakly scalable, and a coarse correction is needed to achieve scalability. To see this, we consider the model problem (1.1.1) and we divide Ω into an increasing number of subdomains using the automated partitioning tool Metis. Figure 1.2 shows an automated decomposition into 16 subdomains and the convergence of the RAS method for an increasing number of subdomains. The study of coarse corrections in two-level and multilevel paradigms will be the aim of Chapters 3 and 4.

However it has recently been observed in applications in computational chemistry that the classical one-level parallel Schwarz method is surprisingly scalable for the solution of two-dimensional chains of fixed-sized subdomains [22, 127]. Mathematically, this property has been studied rigorously for the parallel Schwarz methods in a series of papers [36, 37, 38]. In the next subsections, we study this property for other classical one-level domain decomposition methods, such as the Dirichlet-Neumann, Neumann-Neumann and optimized Schwarz methods. The content of these sections is based on [28].

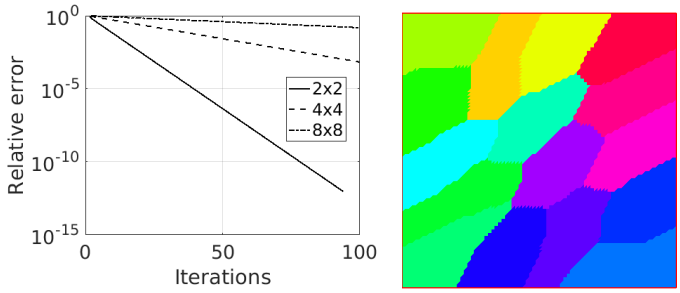


Figure 1.2: On the left, convergence of the RAS method for different decompositions with overlap equal to four times the mesh size. On the right, example of decomposition into 4x4 subdomains using Metis.

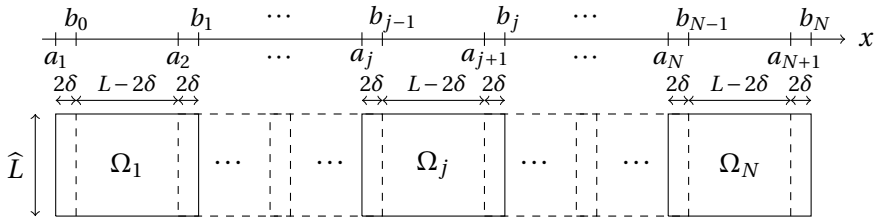


Figure 1.3: Two-dimensional chain of N rectangular fixed-sized subdomains.

1.4.1 Scalability analysis for two dimensional chains of fixed size subdomains

1.4.1.1 Scalability analysis for the optimized Schwarz method

In this subsection we study the scalability properties of OSMs for a two dimensional chain of fixed size subdomains. For the one dimensional chain analysis, we refer the interested reader to [28, Section 4.1]. Let us consider $L > 0$ and δ, $0 < \delta < \frac{L}{2}$, and define the grid points a_j for $j = 1,\ldots,N+1$ and b_j for $j = 0,\ldots,N$ as shown in Figure 1.3. The j-th subdomain of the chain is a rectangle of dimension $\Omega_j := (a_j, b_j) \times (0, \widehat{L})$, and $\Omega := \cup_{j=1}^{N} \Omega_j$. We are interested in the solution to

$$-\Delta u = f \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega. \tag{1.4.1}$$

We consider directly the error equation and we define the errors $e_j^n := u - u_j^n$, where u_j^n are the iterates of the OSM. In the error form, the overlapping OSM with Robin transmission conditions with parameter p is given by

$$\begin{aligned} -\Delta e_j^n &= 0 \text{ in } \Omega_j, \\ e_j^n(\cdot,0) = 0,\ e_j^n(\cdot,\widehat{L}) &= 0, \\ \partial_x e_j^n(a_j,\cdot) - p e_j^n(a_j,\cdot) &= \partial_x e_{j-1}^{n-1}(a_j,\cdot) - p e_{j-1}^{n-1}(a_j,\cdot), \\ \partial_x e_j^n(b_j,\cdot) + p e_j^n(b_j,\cdot) &= \partial_x e_{j+1}^{n-1}(b_j,\cdot) + p e_{j+1}^{n-1}(b_j,\cdot), \end{aligned} \tag{1.4.2}$$

for $j = 2,\ldots,N-1$, and

$$\begin{aligned}
-\Delta e_1^n &= 0 \quad \text{in } \Omega_1, & -\Delta e_N^n &= 0 \text{ in } \Omega_N,\\
e_1^n(\cdot,0) &= 0,\ e_1^n(\cdot,\widehat{L}) = 0, & e_N^n(\cdot,0) &= 0,\ e_N^n(\cdot,\widehat{L}) = 0,\\
e_1^n(a_1,\cdot) &= 0, & (\partial_x - p)e_N^n(a_N,\cdot) &= (\partial_x - p)e_{N-1}^{n-1}(a_N,\cdot),\\
(\partial_x + p)e_1^n(b_1,\cdot) &= (\partial_x + p)e_2^{n-1}(b_1,\cdot), & e_N^n(b_N,\cdot) &= 0.
\end{aligned} \tag{1.4.3}$$

To study the iteration, we use the Fourier expansion $e_j^n(x,y) = \sum_{k\in\mathcal{K}} v_j^n(x,k)\sin(ky)$ with $\mathcal{K} := \left\{\frac{\pi}{\widehat{L}}, \frac{2\pi}{\widehat{L}}, \ldots\right\}$. Inserting this expansion into (1.4.2) and (1.4.3), the Fourier coefficients v_j^n satisfy

$$\begin{aligned}
\partial_{xx} v_j^n &= k^2 v_j^n \text{ in } (a_j, b_j),\\
\partial_x v_j^n(a_j) - p v_j^n(a_j) &= \partial_x v_{j-1}^{n-1}(a_j) - p v_{j-1}^{n-1}(a_j),\\
\partial_x v_j^n(b_j) + p v_j^n(b_j) &= \partial_x v_{j+1}^{n-1}(b_j) + p v_{j+1}^{n-1}(b_j).
\end{aligned} \tag{1.4.4}$$

Defining $\mathscr{R}_-^{n-1}(a_j) := \partial_x v_{j-1}^{n-1}(a_j) - p v_{j-1}^{n-1}(a_j)$ and $\mathscr{R}_+^{n-1}(b_j) := \partial_x v_{j+1}^{n-1}(b_j) + p v_{j+1}^{n-1}(b_j)$, the solution to (1.4.4) is given by

$$\begin{aligned}
v_j^n(x,k) = \mathscr{R}_-^{n-1}(a_j)&\left[-\frac{1}{\gamma}(k-p)e^{k(x-b_j)} - \frac{1}{\gamma}(k+p)e^{k(b_j-x)}\right]\\
+ \mathscr{R}_+^{n-1}(b_j)&\left[\frac{1}{\gamma}(k+p)e^{k(x-a_j)} + \frac{1}{\gamma}(k-p)e^{k(a_j-x)}\right],
\end{aligned} \tag{1.4.5}$$

where $\gamma := (k+p)^2 e^{k(L+2\delta)} - (k-p)^2 e^{-k(L+2\delta)}$. Inserting (1.4.5) into the definitions of $\mathscr{R}_-^n(a_j)$ and $\mathscr{R}_+^n(b_j)$ we obtain

$$\begin{bmatrix}\mathscr{R}_-^n(a_j)\\ \mathscr{R}_+^n(b_j)\end{bmatrix} = T_1\begin{bmatrix}\mathscr{R}_-^{n-1}(a_{j-1})\\ \mathscr{R}_+^{n-1}(b_{j-1})\end{bmatrix} + T_2\begin{bmatrix}\mathscr{R}_-^{n-1}(a_{j+1})\\ \mathscr{R}_+^{n-1}(b_{j+1})\end{bmatrix}, \tag{1.4.6}$$

where

$$T_1 := \begin{bmatrix} g_3 - pg_1 & g_4 - pg_2\\ 0 & 0\end{bmatrix}, \qquad T_2 := \begin{bmatrix} 0 & 0\\ g_4 - pg_2 & g_3 - pg_1\end{bmatrix},$$

with

$$\begin{aligned}
g_1 &:= -\frac{1}{\gamma}(k-p)e^{-2\delta k} - \frac{1}{\gamma}(k+p)e^{2\delta k}, & g_2 &:= \frac{1}{\gamma}(k+p)e^{kL} + \frac{1}{\gamma}(k-p)e^{-kL},\\
g_3 &:= \frac{-k}{\gamma}(k-p)e^{-2\delta k} + \frac{k}{\gamma}(k+p)e^{2\delta k}, & g_4 &:= \frac{k}{\gamma}(k+p)e^{kL} - \frac{k}{\gamma}(k-p)e^{-kL}.
\end{aligned} \tag{1.4.7}$$

Similar arguments allow us to obtain for the subdomains Ω_1, Ω_2, Ω_{N-1}, and Ω_N the relations

$$\begin{aligned}
\begin{bmatrix}0\\ \mathscr{R}_+^n(b_1)\end{bmatrix} &= T_2\begin{bmatrix}\mathscr{R}_-^{n-1}(a_2)\\ \mathscr{R}_+^{n-1}(b_2)\end{bmatrix}, & \begin{bmatrix}\mathscr{R}_-^n(a_2)\\ \mathscr{R}_+^n(b_2)\end{bmatrix} &= \widetilde{T}_1\begin{bmatrix}0\\ \mathscr{R}_+^{n-1}(b_1)\end{bmatrix} + T_2\begin{bmatrix}\mathscr{R}_-^{n-1}(a_3)\\ \mathscr{R}_+^{n-1}(b_3)\end{bmatrix},\\
\begin{bmatrix}\mathscr{R}_-^n(a_N)\\ 0\end{bmatrix} &= T_1\begin{bmatrix}\mathscr{R}_-^{n-1}(a_{N-1})\\ \mathscr{R}_+^{n-1}(b_{N-1})\end{bmatrix}, & \begin{bmatrix}\mathscr{R}_-^n(a_{N-1})\\ \mathscr{R}_+^n(b_{N-1})\end{bmatrix} &= T_1\begin{bmatrix}\mathscr{R}_-^{n-1}(a_{N-2})\\ \mathscr{R}_+^{n-1}(b_{N-2})\end{bmatrix} + \widetilde{T}_2\begin{bmatrix}\mathscr{R}_-^{n-1}(a_n)\\ 0\end{bmatrix},
\end{aligned} \tag{1.4.8}$$

where

$$\widetilde{T}_1 := \begin{bmatrix} 0 & \frac{(k+p)e^{-kL}+(k-p)e^{kL}}{(k+p)e^{k(L+2\delta)}+(k-p)e^{-k(L+2\delta)}} \\ 0 & 0 \end{bmatrix}, \quad \widetilde{T}_2 := \begin{bmatrix} 0 & 0 \\ \frac{(k+p)e^{-kL}+(k-p)e^{kL}}{(k+p)e^{k(L+2\delta)}+(k-p)e^{-k(L+2\delta)}} & 0 \end{bmatrix}.$$

Now we define $\mathbf{r}^n := [0, \mathscr{R}_+^n(b_1), \mathscr{R}_-^n(a_2), \mathscr{R}_+^n(b_2), \ldots, \mathscr{R}_-^n(a_j), \mathscr{R}_+^n(b_j), \ldots, \mathscr{R}_+^n(b_{N-1}), \mathscr{R}_-^n(a_N), 0]^\top$, and using (1.4.6)-(1.4.8), we obtain the iteration relation $\mathbf{r}^n = T_{2D}^O \mathbf{r}^{n-1}$, where T_{2D}^O is the block matrix

$$T_{2D}^O := \begin{bmatrix} & T_2 & & & & & & \\ \widetilde{T}_1 & & T_2 & & & & & \\ & T_1 & & T_2 & & & & \\ & & T_1 & & T_2 & & & \\ & & & \ddots & & \ddots & & \\ & & & & T_1 & & T_2 & \\ & & & & & T_1 & & \widetilde{T}_2 \\ & & & & & & T_1 & \end{bmatrix}. \tag{1.4.9}$$

We are ready to prove the scalability of the OSM in the overlapping case.

Theorem 1.4.1. *Recall* (1.4.7) *and define* $\varphi(k,\delta,p) := |g_3 - pg_1| + |g_4 - pg_2|$. *The overlapping OSM* (1.4.2) *is scalable, in the sense that*

$$\rho(T_{2D}^O(k,\delta,p)) \le \|T_{2D}^O(k,\delta,p)\|_\infty \le \max_k \max\{\varphi(k,\delta,p), \|\widetilde{T}_1(k,\delta,p)\|_\infty\} < 1,$$

independently of N for every $p \ge 0$.

Proof. Because of the structure of T_{2D}^O, the norm $\|T_{2D}^O\|_\infty$ is given by

$$\|T_{2D}^O\|_\infty = \max_i \sum_j |(T_{2D}^O)_{i,j}| = \max\{\varphi(k,\delta,p), \|\widetilde{T}_1\|_\infty, \|\widetilde{T}_2\|_\infty\}.$$

We start showing that $\varphi(k,\delta,p) < 1$ for any $k, p \in [0,\infty)$ and $\delta > 0$. To do so, we notice that

$$\begin{aligned} |g_3 - pg_1| &= |\frac{1}{\gamma}(-k(k-p)e^{-2\delta k} + k(k+p)e^{2\delta k} + p(k-p)e^{-2\delta k} + p(k+p)e^{2\delta k})| \\ &= \frac{1}{\gamma}|(k+p)^2 e^{2\delta k} - (k-p)^2 e^{-2\delta k})| = \frac{1}{\gamma}((k+p)^2 e^{2\delta k} - (k-p)^2 e^{-2\delta k}) \end{aligned}$$

and

$$\begin{aligned} |g_4 - pg_2| &= \frac{1}{\gamma}|(k(k-p)e^{kL} - k(k-p)e^{-kL} - p(k+p)e^{kL} - p(k-p)e^{-kL})| \\ &= \frac{1}{\gamma}(k+p)|k-p|(e^{kL} - e^{-kL}), \end{aligned}$$

which implies that

$$\varphi(k,\delta,p) = \frac{(k+p)^2 e^{2\delta k} - (k-p)^2 e^{-2\delta k} + (k+p)|k-p|(e^{kL} - e^{-kL})}{(k+p)^2 e^{kL+2k\delta} - (k-p)^2 e^{-kL-2k\delta}}.$$

By computing the derivative of φ with respect to p we find

$$\frac{\partial\varphi}{\partial p} = -\frac{2ke^{2\delta k+2kL} - 2ke^{2\delta k}}{k^2\, e^{4\delta k} + 2ke^{4\delta k} + p^2\, e^{4\delta k}\, e^{2kL} + 2k^2\, e^{2\delta k} - 2e^{2\delta k}\, p^2\, e^{kL} + (p-k)^2} \quad \text{for } p<k,$$

$$\frac{\partial\varphi}{\partial p} = \frac{2ke^{2\delta k+2kL} - 2ke^{2\delta k}}{k^2\, e^{4\delta k} + 2ke^{4\delta k} + p^2\, e^{4\delta k}\, e^{2kL} + 2p^2\, e^{2\delta k} - 2e^{2\delta k}\, k^2\, e^{kL} + (k-p)^2} \quad \text{for } p>k.$$

Analyzing the signs of these derivatives, we see that $\varphi(k,\delta,p)$ is strictly decreasing for $p<k$ and it is strictly increasing for $p>k$, thus it reaches a minimum for $p=k$. Therefore the maximum of $\varphi(k,\delta,p)$ with respect to the variable p is obtained for $p=0$ and for $p\to+\infty$:

$$\varphi(k,\delta,p) \le \max\big\{\varphi(k,\delta,0)\,,\ \lim_{p\to\infty}\varphi(k,\delta,p)\big\}.$$

For $p=0$, $\delta>0$ and $L>0$ we have

$$\begin{aligned}\varphi(k,\delta,p=0) &= \frac{e^{2\delta k} - e^{-2\delta k} + e^{kL} - e^{-kL}}{e^{kL+2\delta k} - e^{-kL-2\delta k}} \\ &= \frac{\sinh(2\delta k) + \sinh(kL)}{\sinh(kL)\cosh(2\delta k) + \sinh(2\delta k)\cosh(kL)} < 1,\end{aligned}$$

and, under the same conditions,

$$\lim_{p\to\infty}\varphi(k,\delta,p) = \frac{\sinh(2\delta k) + \sinh(kL)}{\sinh(kL)\cosh(2\delta k) + \sinh(2\delta k)\cosh(kL)} = \varphi(k,\delta,0) < 1.$$

Hence, it holds that $\varphi(k,\delta,p) \le \varphi(k,\delta,0) < 1$. We now focus on $\|\widetilde{T}_1\|_\infty$ and $\|\widetilde{T}_2\|_\infty$. Notice that $\|\widetilde{T}_1\|_\infty = \|\widetilde{T}_2\|_\infty$ and

$$\begin{aligned}\|\widetilde{T}_1\|_\infty &= \left|\frac{(k+p)e^{-kL} + (k-p)e^{kL}}{(k+p)e^{k(L+2\delta)} + (k-p)e^{-k(L+2\delta)}}\right| \\ &= \left|\frac{k\cosh(kL) - p\sinh(kL)}{k\cosh(k(L+2\delta)) + p\sinh(k(L+2\delta))}\right| < 1.\end{aligned}$$

In order to get a bound independently of k, we observe that $\lim_{k\to\infty}\varphi(k,\delta,p) = \lim_{k\to\infty}\|\widetilde{T}_1\|_\infty = 0$ if $\delta>0$. Therefore defining $\bar\rho(\delta) := \max_k \max\{\varphi(k,\delta,p), \|\widetilde{T}_1(k,\delta,p)\|_\infty\}$, we see that $\|T_{2D}^O\|_\infty = \max\{\varphi, \|\widetilde{T}_1\|, \|\widetilde{T}_2\|\} < \bar\rho(\delta) < 1$, for every $\delta, p > 0$. □

For the case without overlap, we need a further argument because for $\delta=0$ both $\rho(T_{2D}^O)$ and $\|T_{2D}^O\|_\infty$ are less than one for any finite frequency k, but tend to one as $k\to\infty$. One can therefore construct a situation where the method would not be scalable as follows: suppose we have N subdomains, and on the j-th subdomain we choose as initial guess e_j^0 the j-th frequency $e_j^0 = \hat e_j^0 \sin(j\frac{\pi}{\widetilde{L}}y)$. Then the convergence of the method is determined by the frequency which maximizes $\rho(T_{2D}^O(k))$. When the number of subdomains N becomes large, this maximum is attained for the largest frequency $k_N = N\frac{\pi}{\widetilde{L}_y}$ since

$\rho(T_{2D}^O(k)) \to 1$ as $k \to \infty$. Thus, every time we add a subdomain to the chain with a new initial condition on the interface $N+1$ according to our rule, the convergence rate of the method deteriorates from $\rho(T_{2D}^O(N\frac{\pi}{\tilde{L}}))$ to $\rho(T_{2D}^O((N+1)\frac{\pi}{\tilde{L}}))$ and the scalability property is lost. Theorem 1.4.2 gives however a sufficient condition such that the OSM is weakly scalable also without overlap, and to see this we introduce the vector $\mathbf{e}^n$ with $e_k^n = \|\mathbf{r}^n(k)\|_\infty$ where $\mathbf{r}^n(k)$ contains the Robin traces at the interfaces of the k-th Fourier mode.

Theorem 1.4.2. *Given a tolerance* `Tol`*, and supposing there exists a* $\tilde{k}$ *that does not depend on N such that* $e_k^0 <$ `Tol` *for every* $k > \tilde{k}$*, then the OSM without overlap,* $\delta = 0$*, and* $p > 0$ *is weakly scalable.*

Proof. Suppose that the initial guess satisfies $\|\mathbf{e}^0\|_\infty >$ `Tol`, since otherwise there is nothing to prove. Then, due to the hypothesis, we have that $\max_{\frac{\pi}{L} \le k \le \tilde{k}} e_k^0 >$ `Tol`. We now show that the method contracts with a ρ independent of the number of subdomains up to the tolerance `Tol`, and therefore we have scalability. Indeed, for every k such that $\frac{\pi}{L} \le k \le \tilde{k}$

$$e_k^n = \|\mathbf{r}^n(k)\|_\infty \le \|T_{2D}^O(k)\|_\infty \|\mathbf{r}^{n-1}(k)\|_\infty \le \|T_{2D}^O(\bar{k})\|_\infty \|\mathbf{r}^{n-1}(k)\|_\infty = \|T_{2D}^O(\bar{k})\|_\infty e_k^{n-1},$$

where $\|T_{2D}^O(\bar{k})\|_\infty = \max_{\frac{\pi}{L} \le k \le \tilde{k}} \|T_{2D}^O(k)\|_\infty < 1$ because $\|T_{2D}^O(k)\|_\infty$ is strictly less than 1 for every finite k. Now for $k > \tilde{k}$,

$$e_k^n = \|\mathbf{r}^n(k)\|_\infty \le \|T_{2D}^O(k)\|_\infty \|\mathbf{r}^{n-1}(k)\|_\infty \le \|\mathbf{r}^{n-1}(k)\|_\infty = e_k^{n-1},$$

since $\|T_{2D}^O(k)\|_\infty \le 1$. Therefore we observe that the method does not increase the error for the frequencies $k > \tilde{k}$ while it contracts for the other frequencies with a contraction factor of at least $\bar{\rho} = \|T_{2D}^O(\bar{k})\|_\infty < 1$. Hence, as long as $\|\mathbf{e}^n\|_\infty >$ `Tol`, we have $\|\mathbf{e}^n\|_\infty \le \bar{\rho}^n \|\mathbf{e}^0\|_\infty$ with $\bar{\rho}$ independent of N. □

The technical assumption in Theorem 1.4.2 on the frequency content of the initial error is not restrictive, since in a numerical implementation we have a maximum frequency $k_{\max}$ which can be represented by the grid. Choosing $\tilde{k} = k_{\max}$, the hypothesis of Theorem 1.4.2 is verified. Note also that without overlap, $\delta = 0$, we have that $\|T_{2D}^O\|_\infty = 1$ for $p = 0$ or $p \to \infty$. Therefore we can not conclude that the method is scalable in these two cases. For $p = 0$, the OSM exchanges only partial derivatives information on the interface. For $p \to \infty$, we obtain the classical Schwarz algorithm and it is well known that without overlap ($\delta = 0$), the method does not converge. We finally show the behaviour of $p \mapsto \|T_{2D}^O(k, \delta, p)\|_\infty$ for a fixed pair (δ, k) in Figure 1.4. According to the proof of Theorem 1.4.1, the minimum of the function $p \mapsto \varphi(k, \delta, p)$ is located at $p = k$. Even though it is a minimum for $\varphi(k, \delta, p)$ and not necessarily for $\|T_{2D}^O(k, \delta, p)\|_\infty$ or $\rho(T_{2D}^O)$, we might deduce from Figure 1.4 that in order to eliminate the k-th frequency, a good choice would be to set $p := k$ in the OSM. For the Laplace equation, it has been shown for two subdomains that setting $p := k$ leads to a vanishing convergence factor $\rho(k)$ for the frequency k [74]. In the case of many subdomains, a similar result has not been proved yet, but Figure 1.4 indicates that it might hold as well.

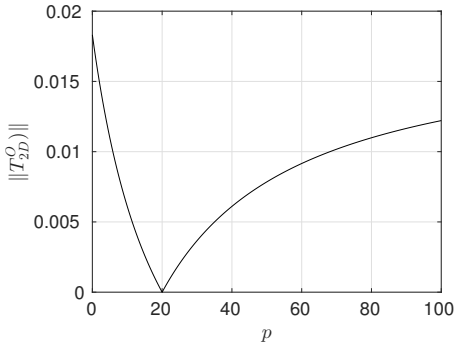


Figure 1.4: Infinity norm of the iteration matrix T_{2D}^O as a function of p for $L = 1, \widehat{L} = 1, \delta = 0.1, k = 20, N = 50$.

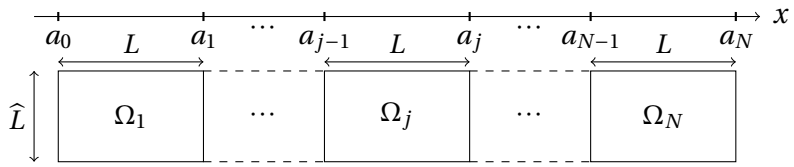


Figure 1.5: Nonoverlapping domain decomposition in two dimensions. Notice that $a_j = jL$.

1.4.1.2 Scalability analysis for the Dirichlet-Neumann method

We now consider a two dimensional problem decomposed into nonoverlapping subdomains as shown in Figure 1.5. The error form of the parallel Dirichlet Neumann method (PDNM) is given by

$$\begin{aligned}
-\Delta e_j^n &= 0 \quad \text{in } \Omega_j, \\
e_j^n(\cdot,0) = 0,\ & e_j^n(\cdot,\widehat{L}) = 0, \\
e_j^n(a_j,\cdot) &= (1-\theta)e_j^{n-1}(a_j,\cdot) + \theta e_{j+1}^{n-1}(a_j,\cdot), \\
\partial_x e_j^n(a_{j-1},\cdot) &= (1-\mu)\partial_x e_j^{n-1}(a_{j-1},\cdot) + \mu\partial_x e_{j-1}^{n-1}(a_{j-1},\cdot),
\end{aligned}$$

for $j = 2,\ldots,N-1$, and

$$\begin{aligned}
-\Delta e_1^n &= 0 \quad \text{in } \Omega_1, \\
e_1^n(\cdot,0) = 0,\ & e_1^n(\cdot,\widehat{L}) = 0, \\
e_1^n(a_0,\cdot) &= 0, \\
e_1^n(a_1,\cdot) &= (1-\theta)e_1^{n-1}(a_1,\cdot) + \theta e_2^{n-1}(a_1,\cdot),
\end{aligned}$$

and

$$\begin{aligned}
-\Delta e_N^n &= 0 \quad \text{in } \Omega_N,\\
e_N^n(\cdot,0) = 0,\ & e_N^n(\cdot,\widehat{L}) = 0,\\
\partial_x e_N^n(a_{N-1},\cdot) &= (1-\mu)\partial_x e_N^{n-1}(a_{N-1},\cdot) + \mu\partial_x e_{N-1}^{n-1}(a_{N-1},\cdot),\\
e_N^n(a_N,\cdot) &= 0,
\end{aligned}$$

where $\theta,\mu \in (0,1)$. We consider again the Fourier expansion of e_j^n, where the Fourier coefficients $v_j^n(x,k)$ solve

$$\begin{aligned}
\partial_{xx} v_j^n &= k^2 v_j^n \text{ in } (a_j,b_j),\\
v_j^n(a_j) &= (1-\theta)v_j^{n-1}(a_j) + \theta v_{j+1}^{n-1}(a_j),\\
\partial_x v_j^n(a_{j-1}) &= (1-\mu)\partial_x v_j^{n-1}(a_{j-1}) + \mu\partial_x v_{j-1}^{n-1}(a_{j-1}).
\end{aligned}$$

Defining

$$\begin{aligned}
\mathscr{D}_j^n &:= (1-\theta)v_j^{n-1}(a_j) + \theta v_{j+1}^{n-1}(a_j),\\
\mathscr{N}_j^n &:= (1-\mu)\partial_x v_j^{n-1}(a_{j-1}) + \mu\partial_x v_{j-1}^{n-1}(a_{j-1}),
\end{aligned}$$

we get

$$v_j^n(x,k) = \frac{1}{k\gamma_2}\left[ke^{-k[(j-1)L-x]}\mathscr{D}_j^n + e^{-k(jL-x)}\mathscr{N}_j^n + ke^{k[(j-1)L-x]}\mathscr{D}_j^n - e^{k(jL-x)}\mathscr{N}_j^n\right],$$

for $j = 2,\dots,N-1$, and

$$v_1^n(x,k) = \frac{D_1^n}{\gamma_1}[e^{kx} - e^{-kx}], \quad v_N^n(x,k) = \frac{1}{k\gamma_2}\left[e^{-k(NL-x)}N_N^n - e^{k(NL-x)}N_N^n\right],$$

where $\gamma_1 := e^{-kL} - e^{kL}$ and $\gamma_2 := e^{kL} + e^{-kL}$. Using the expressions of $\mathscr{N}_j^n$ and $\mathscr{D}_j^n$, we get

$$\begin{aligned}
\begin{bmatrix}0\\ \mathscr{D}_1^n\end{bmatrix} &= \widehat{T}_1\begin{bmatrix}\mathscr{N}_1^{n-1}\\ \mathscr{D}_1^{n-1}\end{bmatrix} + T_2\begin{bmatrix}\mathscr{N}_2^{n-1}\\ \mathscr{D}_2^{n-1}\end{bmatrix},\\
\begin{bmatrix}\mathscr{N}_2^n\\ \mathscr{D}_2^n\end{bmatrix} &= \widetilde{T}_0\begin{bmatrix}0\\ \mathscr{D}_1^{n-1}\end{bmatrix} + T_1\begin{bmatrix}\mathscr{N}_2^{n-1}\\ \mathscr{D}_2^{n-1}\end{bmatrix} + T_2\begin{bmatrix}\mathscr{N}_3^{n-1}\\ \mathscr{D}_3^{n-1}\end{bmatrix},\\
\begin{bmatrix}\mathscr{N}_j^n\\ \mathscr{D}_j^n\end{bmatrix} &= T_0\begin{bmatrix}\mathscr{N}_{j-1}^{n-1}\\ \mathscr{D}_{j-1}^{n-1}\end{bmatrix} + T_1\begin{bmatrix}\mathscr{N}_j^{n-1}\\ \mathscr{D}_j^{n-1}\end{bmatrix} + T_2\begin{bmatrix}\mathscr{N}_{j+1}^{n-1}\\ \mathscr{D}_{j+1}^{n-1}\end{bmatrix}, \text{ for } j = 3,\dots,N-2,\\
\begin{bmatrix}\mathscr{N}_{N-1}^n\\ \mathscr{D}_{N-1}^n\end{bmatrix} &= T_0\begin{bmatrix}\mathscr{N}_{N-2}^{n-1}\\ \mathscr{D}_{N-2}^{n-1}\end{bmatrix} + T_1\begin{bmatrix}\mathscr{N}_{N-1}^{n-1}\\ \mathscr{D}_{N-1}^{n-1}\end{bmatrix} + \widetilde{T}_2\begin{bmatrix}\mathscr{N}_N^{n-1}\\ 0\end{bmatrix},\\
\begin{bmatrix}\mathscr{N}_N^n\\ 0\end{bmatrix} &= \widetilde{T}_1\begin{bmatrix}\mathscr{N}_N^{n-1}\\ \mathscr{D}_N^{n-1}\end{bmatrix} + \widetilde{T}_2\begin{bmatrix}\mathscr{N}_{N-1}^{n-1}\\ \mathscr{D}_{N-1}^{n-1}\end{bmatrix},
\end{aligned} \tag{1.4.10}$$

where

$$T_0 := \begin{bmatrix} \frac{2}{\gamma_2} & \frac{k\gamma_1}{\gamma_2} \\ 0 & 0 \end{bmatrix},\ T_1 := \begin{bmatrix} 1-\mu & 0 \\ 0 & 1-\theta \end{bmatrix},\ T_2 := \begin{bmatrix} 0 & 0 \\ \frac{-\theta\gamma_1}{k\gamma_2} & \frac{2\theta}{\gamma_2} \end{bmatrix},$$
$$\widetilde{T}_0 := \begin{bmatrix} 0 & \frac{\mu k\gamma_2}{\gamma_1} \\ 0 & 0 \end{bmatrix},\ \widehat{T}_1 := \begin{bmatrix} 0 & 0 \\ 0 & 1-\theta \end{bmatrix},\ \widetilde{T}_1 := \begin{bmatrix} 1-\mu & 0 \\ 0 & 0 \end{bmatrix},\ \widetilde{T}_2 := \begin{bmatrix} 0 & 0 \\ -\frac{\theta\gamma_1}{k\gamma_2} & 0 \end{bmatrix}.$$

Defining $\mathbf{e}^n := \left[0, \mathscr{D}_1^n, \mathscr{N}_2^n, \mathscr{D}_2^n, \ldots, \mathscr{N}_j^n, \mathscr{D}_j^n, \ldots, \mathscr{N}_{N-1}^n, \mathscr{D}_{N-1}^n, \mathscr{N}_N^n, 0\right]^\top$, the iteration relations (1.4.10) may be rewritten as

$$\mathbf{e}^n = T_{2D}^{DN} \mathbf{e}^{n-1},$$

where

$$T_{2D}^{DN} := \begin{bmatrix} \widehat{T}_1 & T_2 & & & & & & \\ \widetilde{T}_0 & T_1 & T_2 & & & & & \\ & T_0 & T_1 & T_2 & & & & \\ & & T_0 & T_1 & T_2 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & T_0 & T_1 & T_2 & \\ & & & & & T_0 & T_1 & \widetilde{T}_2 \\ & & & & & & T_0 & \widetilde{T}_1 \end{bmatrix}. \tag{1.4.11}$$

Numerically we observe that $\rho(T_{2D}^{DN}) < 1$, but in general $\|T_{2D}^{DN}\|_\infty > 1$. Hence, the infinity-norm is not suitable to bound the spectral radius and conclude convergence and scalability. Nevertheless in Theorem 1.4.3, we prove scalability of the PDNM under certain assumptions on the parameters μ, θ and using similarity arguments..

Theorem 1.4.3. *Denote by $k_{\min}$ the minimum frequency and define $\alpha(x) := 1/\cosh(x)$. If $\theta = \mu$, then*

$$\rho(T_{2D}^{DN}) \leq \bar{\rho}(\mu) := \sqrt{1-\mu+\mu^2} + \mu\alpha(k_{\min}L),$$

where $\bar{\rho}(\mu)$ is independent of N. Furthermore, if $\cosh(k_{\min}L) > 2$, then $\bar{\rho}(\mu) < 1$ for any positive μ such that $\mu < \frac{1-2\alpha(k_{\min}L)}{1-\alpha(k_{\min}L)^2}$, which implies that the PDNM is convergent and scalable.

We show in Figure 1.6 the function $\bar{\rho}(\mu)$ for the case $\widehat{L} = 1$, that is $k_{\min} = \pi$. The proof of Theorem 1.4.3 relies on the following lemma.

Lemma 1.4.4. *Let $\alpha(x) := 1/\cosh(x)$. Then for any $x \in (0,\infty)$ such that $\cosh(x) > 2$ it holds that $\frac{1-2\alpha(x)}{1-\alpha(x)^2} \in (0,1)$. Moreover, for any $x \in (0,\infty)$ and $\mu \in (0,1)$ such that $\cosh(x) > 2$ and $\mu < \frac{1-2\alpha(x)}{1-\alpha(x)^2}$, it holds that $\sqrt{1-\mu+\mu^2} + \alpha(x)\mu < 1$.*

Proof. Let $x \in (0,\infty)$, then $\alpha(x) = 1/\cosh(x) < 1$. Hence we have $0 < 1-\alpha(x)^2 < 1$. Now, take any $x \in (0,\infty)$ such that $\cosh(x) > 2$. First, we have $\frac{1}{2} > \frac{1}{\cosh(x)}$, which implies that

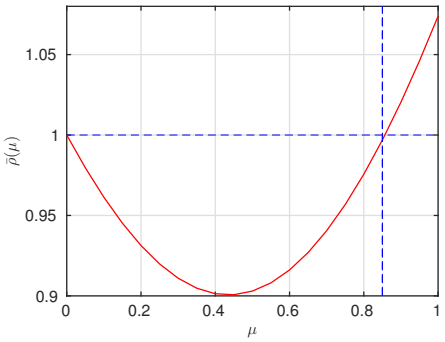


Figure 1.6: Function $\mu \mapsto \bar{\rho}(\mu)$ for $L = 1$ and $\widehat{L} = 1$. Notice that for $\mu < 0.831$ (vertical dashed line) it holds that $\bar{\rho}(\mu) < 1$.

$1 - 2\alpha(x) > 0$. Second, we note that $1 - 2\alpha(x) < 1 - \alpha(x) < 1 - \alpha(x)^2$. Therefore, we obtain that $\frac{1-2\alpha(x)}{1-\alpha(x)^2} \in (0,1)$. Now take any $\mu \in (0,1)$ such that $\mu < \frac{1-2\alpha(x)}{1-\alpha(x)^2}$. This implies that $1 - 2\alpha(x) + (\alpha(x)^2 - 1)\mu > 0$. Multiplying this by μ, with a direct calculation, we get $-\mu + \mu^2 < -2\alpha(x)\mu + \alpha(x)^2\mu^2$. Adding 1 to both sides and taking the square root then leads to the claim. □

We are now ready to prove Theorem 1.4.3.

Proof. If $\mu = \theta$, the matrix T_{2D}^{DN} has the structure

$$T_{2D}^{DN} := \begin{bmatrix} 0 & & & & & & & & & \\ & \widetilde{B}_{1,1} & \widetilde{B}_{1,2} & \frac{2\mu}{\gamma_2} & & & & & & \\ & \widetilde{B}_{2,1} & \widetilde{B}_{2,2} & & & & & & & \\ & & & \widehat{B}_{1,1} & \widehat{B}_{1,2} & \frac{2\mu}{\gamma_2} & & & & \\ & & \frac{2\mu}{\gamma_2} & \widehat{B}_{2,1} & \widehat{B}_{2,2} & & & & & \\ & & & \ddots & \ddots & \ddots & \ddots & & & \\ & & & & & \widehat{B}_{1,1} & \widehat{B}_{1,2} & \frac{2\mu}{\gamma_2} & & \\ & & & & \frac{2\mu}{\gamma_2} & \widehat{B}_{2,1} & \widehat{B}_{2,2} & & & \\ & & & & & & \overline{B}_{1,1} & \overline{B}_{1,2} & \frac{2\mu}{\gamma_2} & \\ & & & & & \frac{2\mu}{\gamma_2} & \overline{B}_{2,1} & \overline{B}_{2,2} & & \\ & & & & & & & & & 0 \end{bmatrix},$$

where $\widetilde{B}, \widehat{B}, \overline{B} \in \mathbb{R}^{2\times 2}$. We introduce an invertible block diagonal matrix

$$G := \begin{bmatrix} g & 0 & & & & & \\ 0 & \widetilde{G} & 0 & & & & \\ & 0 & \widehat{G} & 0 & & & \\ & & 0 & \widehat{G} & 0 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & 0 & \widehat{G} & 0 \\ & & & & & 0 & g \end{bmatrix}, \text{ with } \widehat{G} := \begin{bmatrix} \widehat{d}_1 & 0 \\ 0 & \widehat{d}_2 \end{bmatrix} \text{ and } \widetilde{G} := \begin{bmatrix} \widetilde{d}_1 & 0 \\ 0 & \widetilde{d}_2 \end{bmatrix},$$

where the elements $g, \widehat{d}_1, \widehat{d}_2, \widetilde{d}_1, \widetilde{d}_2 \in \mathbb{R} \setminus \{0\}$ will be chosen in such a way that the matrix $G^{-1} T_{2D}^{DN} G$ can be bounded in some suitable norm. We have

$$G^{-1} T_{2D}^{DN} G := \begin{bmatrix} 0 & & & & & & & & & \\ & \widetilde{C}_{1,1} & \widetilde{C}_{1,2} & \frac{2\mu}{\gamma_2} & & & & & & \\ & \widetilde{C}_{2,1} & \widetilde{C}_{2,2} & & & & & & & \\ & & & \widehat{C}_{1,1} & \widehat{C}_{1,2} & \frac{2\mu}{\gamma_2} & & & & \\ & & \frac{2\mu}{\gamma_2} & \widehat{C}_{2,1} & \widehat{C}_{2,2} & & & & & \\ & & & \ddots & \ddots & \ddots & \ddots & & & \\ & & & & & \widehat{C}_{1,1} & \widehat{C}_{1,2} & \frac{2\mu}{\gamma_2} & & \\ & & & & \frac{2\mu}{\gamma_2} & \widehat{C}_{2,1} & \widehat{C}_{2,2} & & & \\ & & & & & & & \overline{C}_{1,1} & \overline{C}_{1,2} & \frac{2\mu}{\gamma_2} \\ & & & & & & \frac{2\mu}{\gamma_2} & \overline{C}_{2,1} & \overline{C}_{2,2} & \\ & & & & & & & & & 0 \end{bmatrix},$$

where

$$\widetilde{C} = \widetilde{G}^{-1} \widetilde{B} \widetilde{G}, \quad \widehat{C} = \widehat{G}^{-1} \widehat{B} \widehat{G}, \quad \overline{C} = \widehat{G}^{-1} \overline{B} \widehat{G}.$$

Now, we split $G^{-1} T_{2D}^{DN} G$ into a sum, i.e. $G^{-1} T_{2D}^{DN} G = T_{\text{diag}} + T_{\text{off}}$, where T_{diag} contains the diagonal blocks, that is

$$T_{\text{diag}} = \begin{bmatrix} 0 & & & & & & \\ & \widetilde{C} & & & & & \\ & & \widehat{C} & & & & \\ & & & \ddots & & & \\ & & & & \widehat{C} & & \\ & & & & & \overline{C} & \\ & & & & & & 0 \end{bmatrix},$$

and $T_{\text{off}} = G^{-1} T_{2D}^{DN} G - T_{\text{diag}}$ contains the remaining off-diagonal elements $\frac{2\mu}{\gamma_2}$. Then we have

$$\begin{aligned} \rho(T_{2D}^{DN}) = \rho(G^{-1} T_{2D}^{DN} G) &\le \| G^{-1} T_{2D}^{DN} G \|_2 = \| T_{\text{diag}} + T_{\text{off}} \|_2 \\ &\le \| T_{\text{diag}} \|_2 + \| T_{\text{off}} \|_2 \le \sqrt{\rho(T_{\text{diag}}^\top T_{\text{diag}})} + \sqrt{\rho(T_{\text{off}}^\top T_{\text{off}})}. \end{aligned} \tag{1.4.12}$$

Notice that

$$T_{\text{off}}^{\top} T_{\text{off}} = \text{diag}\left(0, 0, \frac{4\mu^2}{\gamma_2^2}, \ldots, \frac{4\mu^2}{\gamma_2^2}, 0, 0\right),$$

and hence $\sqrt{\rho(T_{\text{off}}^{\top} T_{\text{off}})} = \frac{2\mu}{\gamma_2}$. Now, we focus on the term $\rho(T_{\text{diag}}^{\top} T_{\text{diag}})$. The block diagonal structure of $T_{\text{diag}}^{\top} T_{\text{diag}}$ allows us to write

$$\rho(T_{\text{diag}}^{\top} T_{\text{diag}}) = \sqrt{\max\{\rho(\widetilde{C}^{\top}\widetilde{C}), \rho(\widehat{C}^{\top}\widehat{C}), \rho(\overline{C}^{\top}\overline{C})\}}. \tag{1.4.13}$$

The evaluation of the spectral radii $\rho(\widetilde{C}^{\top}\widetilde{C})$, $\rho(\widehat{C}^{\top}\widehat{C})$, and $\rho(\overline{C}^{\top}\overline{C})$ leads to the analysis of cumbersome formulas, and we thus bound instead the spectral radii by the corresponding infinity-norms. To do so, setting $\widetilde{d}_1 := \gamma_1$ and $\widetilde{d}_2 := k\gamma_2$, we obtain

$$\rho(\widetilde{C}^{\top}\widetilde{C}) = \rho(\widetilde{G}\widetilde{B}^{\top}\widetilde{G}^{-1}\widetilde{G}^{-1}\widetilde{B}\widetilde{G}) \leq \|\widetilde{G}\widetilde{B}^{\top}\widetilde{G}^{-1}\widetilde{G}^{-1}\widetilde{B}\widetilde{G}\|_{\infty} = 2\mu^2 - 2\mu + 1.$$

Next, we set $\widehat{d}_1 := \gamma_2$ and $\widehat{d}_2 := k\gamma_1$ and get

$$\begin{aligned}\rho(\overline{C}^{\top}\overline{C}) &= \rho(\widehat{G}\overline{B}^{\top}\widehat{G}^{-1}\widehat{G}^{-1}\overline{B}\widehat{G}) \leq \|\widehat{G}\overline{B}^{\top}\widehat{G}^{-1}\widehat{G}^{-1}\overline{B}\widehat{G}\|_{\infty} \\ &= \max\left\{1-\mu, 1-\mu+\frac{\mu^2(e^{-kL}-e^{kL})^4}{(e^{-kL}+e^{kL})^4}\right\} \leq 1-\mu+\mu^2,\end{aligned}$$

where the fact that $\frac{(e^{-kL}-e^{kL})^4}{(e^{-kL}+e^{kL})^4} \leq 1$ for any k is used. Now, a direct calculation shows that

$$2\mu^2 - 2\mu + 1 \leq 2\mu^2 - 2\mu + 1 + \frac{4\mu(1-\mu)}{(e^{k_{\min}L}+e^{-k_{\min}L})^2} \leq 1 - \mu + \mu^2,$$

for any $\mu \in (0,1)$. Therefore, we obtain

$$\|T_{\text{diag}}\|_2 = \rho(T_{\text{diag}}^{\top} T_{\text{diag}}) \leq \sqrt{1-\mu+\mu^2}.$$

Recalling (1.4.12) and (1.4.13), we conclude that

$$\begin{aligned}\rho(T_{2D}^{DN}) &\leq \|T_{\text{diag}}\|_2 + \|T_{\text{off}}\|_2 \leq \sqrt{1-\mu+\mu^2} + \frac{2\mu}{\gamma_2} \\ &\leq \sqrt{1-\mu+\mu^2} + \frac{2\mu}{(e^{k_{\min}L}+e^{-k_{\min}L})} =: \bar{\rho}(\mu),\end{aligned}$$

which is the first statement of the theorem. The second part follows now from Lemma 1.4.4 by observing that if $\bar{\rho}(\mu) < 1$, then $\rho(T_{2D}^{DN}) \leq \bar{\rho}(\mu) < 1$ where $\bar{\rho}(\mu)$ is independent of N. □

1.4.1.3 Scalability analysis for the Neumann-Neumann method

Finally we study the convergence of the Neumann-Neumann method (NNM). For our model problem, the error form for the NNM is the following: first solve

$$\begin{aligned} -\Delta e_j^n &= 0 \text{ in } \Omega_j, \\ e_j^n(\cdot,0) = 0,\ & e_j^n(\cdot,L) = 0, \\ e_j^n(a_{j-1},\cdot) = \mathscr{D}_{j-1}^n,\ & e_j^n(a_j,\cdot) = \mathscr{D}_j^n, \end{aligned}$$

for $j = 2,\ldots,N-1$ and

$$\begin{aligned} -\Delta e_1^n &= 0 \text{ in } \Omega_1, & -\Delta e_N^n &= 0 \text{ in } \Omega_N, \\ e_1^n(\cdot,0) = 0,\ & e_1^n(\cdot,L) = 0, & e_N^n(\cdot,0) = 0,\ & e_N^n(\cdot,L) = 0, \\ e_1^n(a_0,\cdot) = 0,\ & e_1^n(a_1,\cdot) = \mathscr{D}_1^n, & e_N^n(a_{N-1},\cdot) = \mathscr{D}_{N-1}^n,\ & e_N^n(a_N,\cdot) = 0, \end{aligned}$$

then solve

$$\begin{aligned} -\Delta \psi_j^n &= 0 \quad \text{in } \Omega_j, \\ \partial_x \psi_j^n(\cdot,0) = 0,\ & \psi_j^n(\cdot,L) = 0, \\ \partial_x \psi_j^n(a_{j-1},\cdot) &= \partial_x e_j^n(a_{j-1},\cdot) - \partial_x e_{j-1}^n(a_{j-1},\cdot), \\ \partial_x \psi_j^n(a_j,\cdot) &= \partial_x e_j^n(a_j,\cdot) - \partial_x e_{j+1}^n(a_j,\cdot), \end{aligned}$$

for $j = 2,\ldots,N-1$ and

$$\begin{aligned} -\Delta \psi_1^n &= 0 \quad \text{in } \Omega_1, \\ \psi_1^n(\cdot,0) = 0,\ & \psi_1^n(\cdot,L) = 0,\ \psi_1^n(a_0,\cdot) = 0, \\ \partial_x \psi_1^n(a_1,\cdot) &= \partial_x e_1^n(a_1,\cdot) - \partial_x e_2^n(a_1,\cdot), \end{aligned}$$

and

$$\begin{aligned} -\Delta \psi_N^n &= 0 \quad \text{in } \Omega_N, \\ \psi_N^n(\cdot,0) = 0,\ & \psi_N^n(\cdot,L) = 0,\ \psi_N^n(a_N,\cdot) = 0, \\ \partial_x \psi_N^n(a_{N-1},\cdot) &= \partial_x e_N^n(a_{N-1},\cdot) - \partial_x e_{N-1}^n(a_{N-1},\cdot), \end{aligned}$$

and finally set

$$\mathscr{D}_j^{n+1} := \mathscr{D}_j^n - \vartheta(\psi_{j+1}^n(a_j,\cdot) + \psi_j^n(a_j,\cdot)), \tag{1.4.14}$$

for $j = 1,\cdots,N-1$, where $\vartheta > 0$. We expand both e_j^n and ψ_j^n in Fourier series ,

$$e_j^n(x,y) = \sum_{m=1}^{\infty} v_j^n(x,k)\sin(ky), \quad \psi_j^n(x,y) = \sum_{m=1}^{\infty} w_j^n(x,k)\sin(ky),$$

where $k \in \mathscr{K}$. The Fourier coefficients $v_j^n(x,k)$ and $w_j^n(x,k)$ solve the problems

$$\begin{aligned} k^2 v_j^n - \partial_{xx} v_j^n &= 0 \quad \text{in } (a_{j-1},a_j), & k^2 w_j^n - \partial_{xx} w_j^n &= 0 \quad \text{in } (a_{j-1},a_j), \\ v_j^n(a_{j-1},k) &= \mathscr{D}_{j-1}^n, & \partial_x w_j^n(a_{j-1},k) &= \partial_x v_j^n(a_{j-1},k) - \partial_x v_{j-1}^n(a_{j-1},k), \\ v_j^n(a_j,k) &= \mathscr{D}_j^n, & \partial_x w_j^n(a_j,k) &= \partial_x v_j^n(a_j,k) - \partial_x v_{j+1}^n(a_j,k), \end{aligned}$$

for $j = 2, \ldots, N-1$, and

$$\begin{aligned} &k^2 v_1^n - \partial_{xx} v_j^n = 0 \quad \text{in } (a_0, a_1), \qquad && k^2 w_1^n - \partial_{xx} w_1^n = 0 \quad \text{in } (a_0, a_1), \\ &\quad v_1^n(a_0, k) = 0, && \quad \widetilde{w}_1^n(a_0, k) = 0, \\ &\quad v_1^n(a_1, k) = \mathscr{D}_1^n, && \partial_x w_1^n(a_1, k) = \partial_x v_1^n(a_1, k) - \partial_x v_2^n(a_1, k), \end{aligned}$$

and

$$\begin{aligned} &k^2 v_N^n - \partial_{xx} v_N^n = 0 \quad \text{in } (a_{N-1}, a_N), \qquad && k^2 w_N^n - \partial_{xx} w_N^n = 0 \quad \text{in } (a_{N-1}, a_N), \\ &\quad v_N^n(a_{N-1}, k) = \mathscr{D}_{N-1}^n, && \partial_x w_N^n(a_{N-1}, k) = \partial_x v_N^n(a_{N-1}, k) - \partial_x v_{N-1}^n(a_{N-1}, k), \\ &\quad v_N^n(a_N, k) = 0, && \quad w_N^n(a_N, k) = 0, \end{aligned}$$

for the first and last subdomains. For the sake of notation, we set $\mathscr{D}_0^n = \mathscr{D}_N^n = 0$ and defining $\gamma_1 := e^{kL} - e^{-kL}$, the solution v_j^n can be written as

$$v_j^n(x,k) = \frac{1}{\gamma_1}\left[\mathscr{D}_j^n\left(e^{k(x-(j-1)L)} - e^{k((j-1)L-x)}\right) + \mathscr{D}_{j-1}^n\left(e^{k(jL-x)} - e^{k(x-jL)}\right)\right],$$

which is used to solve the problems in w_j^n, and we obtain

$$\begin{aligned} w_j^n(x,k) = &\frac{1}{\gamma_1^2}\left(2\mathscr{D}_{j-1}^n(e^{kL} + e^{-kL}) - 2\mathscr{D}_j^n - 2\mathscr{D}_{j-2}^n\right)\left(e^{k(x-jL)} + e^{k(jL-x)}\right) \\ &+ \frac{1}{\gamma_1^2}\left(2\mathscr{D}_j^n(e^{kL} + e^{-kL}) - 2\mathscr{D}_{j-1}^n - 2\mathscr{D}_{j+1}^n\right)\left(e^{k(x-(j-1)L)} + e^{k((j-1)L-x)}\right), \end{aligned}$$

for $j = 2, \ldots, N-1$, and

$$\begin{aligned} w_1^n(x,k) &= \frac{1}{\gamma_1\gamma_2}\left(2\mathscr{D}_1^n(e^{kL} + e^{-kL}) - 2\mathscr{D}_2^n\right)\left(e^{kx} - e^{-kx}\right), \\ w_N^n(x,k) &= \frac{1}{\gamma_1\gamma_2}\left(-2\mathscr{D}_{N-1}^n(e^{kL} + e^{-kL}) + 2\mathscr{D}_{N-2}^n\right)\left(e^{k(x-NL)} - e^{k(NL-x)}\right), \end{aligned}$$

where $\gamma_2 := e^{kL} + e^{-kL}$. Using equation (1.4.14) we get

$$\mathscr{D}_j^{n+1} = \mathscr{D}_j^n - \frac{\vartheta}{\gamma_1^2}\left[4\mathscr{D}_j^n\left(\left(e^{kL} + e^{-kL}\right)^2 - 2\right) - 4\mathscr{D}_{j-2}^n - 4D_{j+2}^n\right], \tag{1.4.15}$$

for $j = 2, \ldots, N-2$, and

$$\begin{aligned} \mathscr{D}_1^{n+1} = &\mathscr{D}_1^n - \frac{\vartheta}{\gamma_2}\left(2(e^{kL} + e^{-kL})\mathscr{D}_1^n - 2\mathscr{D}_2^n\right) \\ &- \frac{\vartheta}{\gamma_1^2}\left(2((e^{kL} + e^{-kL})^2 - 2)\mathscr{D}_1^n + 2(e^{kL} + e^{-kL})\mathscr{D}_2^n - 4\mathscr{D}_3^n\right), \\ \mathscr{D}_{N-1}^{n+1} = &\mathscr{D}_{N-1}^n - \frac{\vartheta}{\gamma_2}\left(2(e^{kL} + e^{-kL})\mathscr{D}_{N-2}^n - 2\mathscr{D}_{N-2}^n\right) \\ &- \frac{\vartheta}{\gamma_1^2}\left(2((e^{kL} + e^{-kL})^2 - 2)\mathscr{D}_{N-1}^n + 2(e^{kL} + e^{-kL})\mathscr{D}_{N-2}^n - 4\mathscr{D}_{N-3}^n\right). \end{aligned} \tag{1.4.16}$$

We define $\mathbf{e}^n = \left[\mathscr{D}_1^n, \mathscr{D}_2^n, \cdots, \mathscr{D}_{N-1}^n\right]^\top$, and write equations (1.4.15)-(1.4.16) as $\mathbf{e}^{n+1} = T_{2D}^{NN}\mathbf{e}^n$, where the iteration matrix T_{2D}^{NN} is given by

$$T_{2D}^{NN} = \begin{bmatrix} \widetilde{\alpha} & \widetilde{\gamma} & \widetilde{\beta} & & & & \\ 0 & \alpha & 0 & \beta & & & \\ \beta & 0 & \alpha & 0 & \beta & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \alpha & 0 & \beta \\ & & & \beta & 0 & \alpha & 0 \\ & & & & \widetilde{\beta} & \widetilde{\gamma} & \widetilde{\alpha} \end{bmatrix},$$

with $\alpha := 1 - \frac{4\vartheta}{\gamma_1^2}\left(\left(e^{kL} + e^{-kL}\right)^2 - 2\right)$, $\beta := \frac{4\vartheta}{\gamma_1^2}$, $\widetilde{\alpha} := 1 - \frac{2\vartheta}{\gamma_2}(e^{kL} + e^{-kL}) - \frac{2\vartheta}{\gamma_1^2}((e^{kL} + e^{-kL})^2 - 2)$, $\widetilde{\gamma} := \frac{2\vartheta}{\gamma_2} - \frac{2\vartheta}{\gamma_1^2}(e^{kL} + e^{-kL})$, $\widetilde{\beta} := \frac{4\vartheta}{\gamma_1^2}$.

Theorem 1.4.5. *If $\frac{L}{\widehat{L}} > \frac{\ln(1+\sqrt{2})}{\pi}$, then the NNM with $\vartheta = \frac{1}{4}$ is scalable, in the sense that $\rho(T_{2D}^{NN}) \leq \|T_{2D}^{NN}\|_\infty = \frac{4}{\gamma_1^2} < 1$.*

Proof. The infinity-norm of T_{2D}^{NN} is given by

$$\|T_{2D}^{NN}\|_\infty = \max\left\{|\widetilde{\alpha}| + |\widetilde{\gamma}| + |\widetilde{\beta}|, |\alpha| + 2|\beta|\right\}.$$

Using $\vartheta = \frac{1}{4}$ and exploiting the definition of γ_1, the coefficient α in T_{2D}^{NN} becomes

$$\alpha = 1 - \frac{1}{\gamma_1^2}\left(\left(e^{kL} + e^{-kL}\right)^2 - 2\right) = 1 - \frac{\cosh(kL)^2}{\sinh(kL)^2} + \frac{1}{2}\frac{1}{\sinh(kL)^2} = -\frac{2}{\gamma_1^2}.$$

Similarly, one obtains that $\widetilde{\alpha} = -\frac{1}{\gamma_1^2}$. Moreover, we have $\beta = \widetilde{\beta} = \frac{1}{\gamma_1^2}$. Therefore, we get

$$\|T_{2D}^{NN}\|_\infty = \max\left\{\left(2 + \frac{2}{\gamma_2}\right)\frac{1}{\gamma_1^2}, \frac{4}{\gamma_1^2}\right\} = \frac{4}{\gamma_1^2},$$

since $\gamma_2 > 1$. This shows that $\|T_{2D}^{NN}\|_\infty$ is strictly smaller than one if the condition $\frac{4}{\gamma_1^2} < 1$ holds, meaning that $\gamma_1 > 2$, and since the map $k \mapsto \gamma_1 = 2\sinh(kL)$ is strictly increasing in k, it suffices that $\gamma_1 > 2$ is satisfied for just $k = \frac{\pi}{\widehat{L}}$. Hence the condition becomes $\sinh(kL) > 1$ or equivalently $kL > \operatorname{arcsinh}(1) = \ln(1 + \sqrt{2})$, which concludes the proof. □

1.4.1.4 Numerical results

We close this subsection with a numerical experiment. We start with a random initial guess and we apply the different methods to solve (1.4.1) with $f = g = 0$. We set the geometric parameters equal to $\widehat{L} = L = 1$ and we discretize each subdomain square with $N_h = 100$ interior unknowns. For the overlapping OSM we choose $\delta = 10h$, where $h = \frac{1}{N_h+1}$ and we set $p = \pi$. For the PDNM, we set $\theta = \lambda = \frac{1}{2}$ while for the PNNM $\theta = \frac{1}{4}$. In Table 1.1,

N	10	20	30	40	50
OSM	13	13	13	13	13
PDNM	70	70	70	70	70
PNNM	6	6	6	6	6

Table 1.1: Number of iterations to reach convergence as the number of subdomains N increases.

we report the number of iterations to reach convergence with a tolerance of Tol := 10^{-12} for the different methods as the number N of the subdomains increases. We can observe that every method requires a constant number of iterations to reach convergence. Therefore, these numerical experiments are in agreement with the theoretical results presented in this subsection. According to Table 1.1, it seems that the PNNM is the fastest method. However, we remark that each iteration of the PNNM requires two subdomain solves, so its cost is comparable with the OSM. Moreover, the PNNM is extremely sensitive on the choice of θ, see [26].

1.4.2 Scalability analysis for Discrete Fracture Network

Discrete fracture networks (DFNs) are advanced mathematical models to study flows in fractured media. A DFN usually consists of complex three-dimensional structures characterized by the intersections of planar polygonal fractures generated stochastically. Thus, fractures are represented by two dimensional planes which intersect randomly in the three dimensional space, giving rise to highly complex structures, that are coupled through the conservation of physical quantities of interest. The generation of accurate meshes is not trivial for DFNs and in the last decade, there has been a great interest in developing robust discretization techniques for DNFs, ranging from the virtual element method [7] to XFEM [8] and optimization approaches [9]. Much less attention has been devoted to the development of fast and robust ad-hoc iterative solvers. In this section, we limit our study to the scalability properties of optimized Schwarz methods applied to DFNs. Part of this work has been carried out during a research visit at the GEOSCORE group led by Prof. Berrone at Politecnico di Torino.

We consider a simplified DFN composed by the union of one-dimensional fractures F_i, $i = 1,...,N$ depicted in Fig 1.7. Denoting the DFN by Ω, we have $\Omega = \cup_{i=1}^{N} F_i$. The boundary of the fractures is denoted with ∂F_i and it holds that $\partial\Omega = \cup_{i=1}^{N} \partial F_i$. Furthermore, the boundary $\partial\Omega$ can be decomposed into a Dirichlet boundary Γ_D and a Neumann boundary Γ_N, so that $\partial\Omega = \Gamma_D \cup \Gamma_N$. The intersections between fractures are called traces and are denoted by S_m, $m = 1,...,N-1 =: M$. We suppose that the vertical fractures have two traces located at $y = \gamma_1$ and $y = \gamma_2$, while the horizontal ones are intersected at $x = \gamma_1$ and

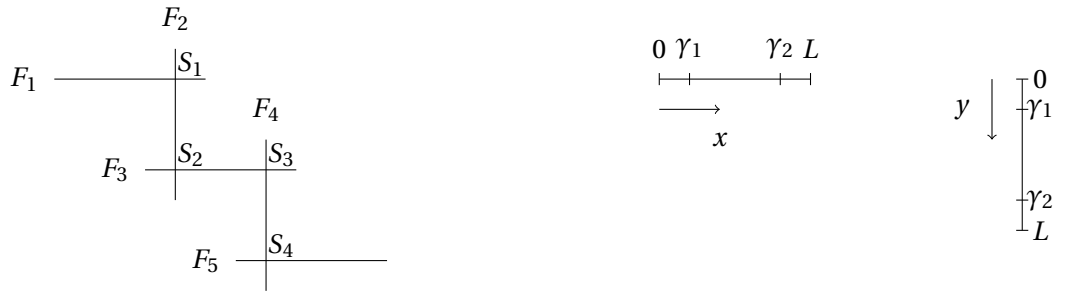


Figure 1.7: On the left, we show the geometry of a discrete fracture network with five fractures intersecting in four traces. On the right we specify the geometry of the fractures.

$x = \gamma_2$, except the first and last fracture which have only one trace. Our goal is to solve

$$\begin{aligned} -\nu(\tau)\partial_{\tau\tau} u = f \quad \text{in } \Omega, \quad \mathscr{B}(u) = 0 \quad \text{on } \partial\Omega, \\ u_{|F_i} = u_{|F_{i+1}} \quad \text{on } S_i,\ i = 1,...,M, \\ \left[\left[\frac{\partial u_i}{\partial \tau_i}\right]\right] + \left[\left[\frac{\partial u_{i+1}}{\partial \tau_{i+1}}\right]\right] = 0 \quad \text{on } S_i,\ i = 1,...,M, \end{aligned} \tag{1.4.17}$$

where τ_i is the local variable on the fracture F_i, the operator $\mathscr{B}$ represents the boundary conditions, $\nu(\tau)$ is the diffusion coefficient which might change between fractures and $[[v]]$ is the jump of v across the intersection of fractures.

We now define an OSM for (1.4.17). Given a DFN Ω composed of N fractures, we decompose Ω into a set of N_{sub} nonoverlapping subdomains Ω_i, $\Omega = \cup_{i=1}^{N_{\text{sub}}} \Omega_i$. Each subdomain can correspond to a single fracture, or to a set of fractures. In the following we suppose that $\Omega_i = F_i$ and thus, $N_{\text{sub}} = N$. To solve equation (1.4.17), starting with an initial guess u_j^0, $j = 1,...,N$, the OSM computes at each iteration $n = 1,2,...$ until convergence

$$\begin{aligned} -\nu_j \partial_{\tau_j\tau_j} u_j^n = f_j \quad \text{in } F_j, \quad \mathscr{B}_j(u_j^n) = 0 \quad \text{on } \partial F_i, \\ \left[\left[\frac{\partial u_j^n}{\partial \tau_j}\right]\right] + s_{j-1}^+ u_j^n = -\left[\left[\frac{\partial u_{j-1}^{n-1}}{\partial \tau_{j-1}}\right]\right] + s_{j-1}^+ u_{j-1}^{n-1} \text{ on } S_{j-1}, \\ \left[\left[\frac{\partial u_j^n}{\partial \tau_j}\right]\right] + s_j^- u_j^n = -\left[\left[\frac{\partial u_{j+1}^{n-1}}{\partial \tau_{j+1}}\right]\right] + s_j^- u_{j+1}^{n-1} \text{ on } S_j. \end{aligned} \tag{1.4.18}$$

for $j = 2,...,N-1$, while for $j = 1, N$,

$$\begin{aligned} -\nu_1 \partial_{\tau_1\tau_1} u_1^n = f_1 \text{ in } F_1, \quad \mathscr{B}_1(u_1^n) = 0, \quad -\nu_N \partial_{\tau_N\tau_N} u_N^n = f_N \text{ in } F_N, \quad \mathscr{B}_N(u_N^n) = 0, \\ \left[\left[\frac{\partial u_1^n}{\partial \tau_1}\right]\right] + s_1^- u_1^n = -\left[\left[\frac{\partial u_2^{n-1}}{\partial \tau_2}\right]\right] + s_1^- u_2^{n-1} \text{ on } S_1, \\ \left[\left[\frac{\partial u_N^n}{\partial \tau_N}\right]\right] + s_{N-1}^+ u_N^n = -\left[\left[\frac{\partial u_{N-1}^{n-1}}{\partial \tau_{N-1}}\right]\right] + s_{N-1}^+ u_{N-1}^{n-1} \text{ on } S_{N-1}. \end{aligned} \tag{1.4.19}$$

The functions f_j are the restriction of the force term on the fracture F_j and $s_j^{+,-}$, $j = 1,...,M$ are positive parameters. We remark that equations $(1.4.18)_{2,3}$ are obtained taking a linear combination with parameter $s_j^{+,-}$ of the coupling conditions expressed in $(1.4.17)_{2,3}$. We obtain what are usually called transmission conditions $(1.4.18)_{2,3}$ in the domain decomposition literature to distinguish them from the physical coupling conditions $(1.4.17)_{2,3}$. The two formulations are equivalent in the sense that, at convergence, $u_j = u|_{F_j}$.

1.4.2.1 Convergence and scalability analysis

In this subsection we perform a convergence and scalability analysis of OSMs applied to DFNs. For the sake of simplicity we suppose that $s_j^{+,-} = p \in \mathbb{R}^+$ for $j = 1,...,N-1$ and $\nu_j = 1$, $j = 1,...,N$. We first discuss the case in which every $\mathscr{B}_j$ represents a Dirichlet boundary condition. Second, we treat the more realistic case in which we have Neumann boundary conditions everywhere, expect at the left boundary of F_1 and at the right boundary of F_N. Intermediate cases can be derived straightforwardly from our analysis.

Equations (1.4.18) and (1.4.19) define a sequence $\left\{u_j^n\right\}_n$. To study the convergence of the proposed method we use the linearity of the problem defining $e_j^n := u - u_j^n$, and we consider the error equations

$$\begin{aligned}
&-\partial_{\tau_j\tau_j} e_j^n = 0, \quad \text{in } F_j, \quad \mathscr{B}_j(e_j^n) = 0 \quad \text{on } \partial F_i,\\
&\left[\left[\frac{\partial e_j^n}{\partial \tau_j}\right]\right] + p e_j^n = -\left[\left[\frac{\partial e_{j-1}^{n-1}}{\partial \tau_{j-1}}\right]\right] + p e_{j-1}^{n-1} \text{ on } S_{j-1},\\
&\left[\left[\frac{\partial e_j^n}{\partial \tau_j}\right]\right] + p e_j^n = -\left[\left[\frac{\partial e_{j+1}^{n-1}}{\partial \tau_{j+1}}\right]\right] + p e_{j+1}^{n-1} \text{ on } S_j.
\end{aligned} \tag{1.4.20}$$

for $j = 2,...,N-1$, while for $j = 1, N$,

$$-\partial_{\tau_1\tau_1} e_1^n = 0, \quad \text{in } F_1, \quad \mathscr{B}_1(e_1^n) = 0, \qquad \left[\left[\frac{\partial e_1^n}{\partial \tau_1}\right]\right] + p e_1^n = -\left[\left[\frac{\partial e_2^{n-1}}{\partial \tau_2}\right]\right] + p e_2^{n-1} \text{ on } S_1,$$

$$-\partial_{\tau_N\tau_N} e_N^n = 0, \quad \text{in } F_N, \quad \mathscr{B}_N(e_N^n) = 0, \qquad \left[\left[\frac{\partial e_N^n}{\partial \tau_N}\right]\right] + p e_N^n = -\left[\left[\frac{\partial e_{N-1}^{n-1}}{\partial \tau_{N-1}}\right]\right] + p e_{N-1}^{n-1} \text{ on } S_{N-1}.$$

Inside each fracture the analytical solutions are

$$e_1^n = \frac{\hat{e}_1^n \tau_1}{\gamma_2}\chi([0,\gamma_2]) + \frac{\hat{e}_1^n (L-\tau_1)}{L-\gamma_2}\chi([\gamma_2, L]), \tag{1.4.21}$$

$$e_j^n = \frac{\hat{e}_j^{1,n}\tau_j}{\gamma_1}\chi([0,\gamma_1]) + \left(\frac{\hat{e}_j^{1,n}(\gamma_2-\tau_j)}{\gamma_2-\gamma_1} + \frac{\hat{e}_j^{2,n}(\tau_j-\gamma_1)}{\gamma_2-\gamma_1}\right)\chi([\gamma_1,\gamma_2]) + \frac{\hat{e}_j^{2,n}(L-\tau_j)}{L-\gamma_2}\chi([\gamma_2, L]), \quad j = 2,...,N-1,$$

$$e_N^n = \frac{\hat{e}_N^n \tau_N}{\gamma_1}\chi([0,\gamma_1]) + \frac{\hat{e}_N^n (L-\tau_N)}{L-\gamma_1}\chi([\gamma_1, L]). \tag{1.4.22}$$

The functions $\chi([a,b])$ are characteristic functions which satisfy $\chi(\tau) = 1$ if $\tau \in [a,b]$ and zero otherwise. We remark that the unknown coefficients $\hat{e}_j^{i,n}$, $i = 1,2$, $j = 1,...,N$ represent the value of the error functions at the traces $S_{j+(i-3)+1}$ on the fracture j. We now

insert these expressions into the transmission conditions, and we aim to express the coefficients of the error in fracture j at iteration n in terms of the coefficients of the errors in fractures $j-1$ and $j+1$ at iteration $n-1$. For instance, for F_1,F_2 and F_3 we find directly

$$\begin{aligned}
&\hat{e}_1^n\left(\frac{L}{\gamma_2(L-\gamma_2)}+p\right)=\hat{e}_2^{1,n-1}\left(p-\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}\right)+\frac{\hat{e}_1^{2,n-1}}{\gamma_2-\gamma_1} \qquad (1.4.23)\\
&\hat{e}_2^{1,n}\left(p+\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}\right)-\frac{\hat{e}_1^{2,n}}{\gamma_2-\gamma_1}=\hat{e}_1^{n-1}\left(p-\frac{L}{\gamma_2(L-\gamma_2)}\right),\\
&-\frac{\hat{e}_1^{1,n}}{\gamma_2-\gamma_1}+\hat{e}_2^{2,n}\left(p+\frac{L-\gamma_1}{(L-\gamma_2)(\gamma_2-\gamma_1)}\right)=\hat{e}_3^{1,n-1}\left(p-\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}\right)+\hat{e}_3^{2,n-1}\frac{1}{\gamma_2-\gamma_1},\\
&\hat{e}_3^{1,n}\left(p+\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}\right)-\frac{\hat{e}_3^{2,n}}{\gamma_2-\gamma_1}=\frac{\hat{e}_2^{n-1,1}}{\gamma_2-\gamma_1}+\hat{e}_2^{n-1,1}\left(p-\frac{L-\gamma_1}{(\gamma_2-\gamma_1)(L-\gamma_2)}\right),\\
&-\frac{\hat{e}_3^{1,n}}{\gamma_2-\gamma_1}+\left(\frac{L-\gamma_1}{(L-\gamma_2)(\gamma_2-\gamma_1)}+p\right)\hat{e}_3^{2,n}=\hat{e}_4^{1,n-1}\left(p-\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}\right)+\frac{1}{\gamma_2-\gamma_1}\hat{e}_4^{2,n-1}.
\end{aligned}$$

Carrying out the calculations for N fractures and defining the vector $\mathbf{v}^n=(\hat{e}_1^n,\hat{e}_2^{1,n},\hat{e}_2^{2,n},...,\hat{e}_N^{1,n})^\top$, $\mathbf{v}^n\in\mathbb{R}^{\widetilde{N}}$, $\widetilde{N}:=2(N-2)+2$, which contains all the values at the traces at iteration n, we find the recurrence relation

$$\mathbf{v}^n=T_N^D\mathbf{v}^{n-1}=M_N^{-1}N_N\mathbf{v}^{n-1}. \qquad (1.4.24)$$

The matrix $M_N\in\mathbb{R}^{\widetilde{N},\widetilde{N}}$ has the following block structure

$$M_N:=\begin{pmatrix}F_1&&&&&\\&F_2&&&&\\&&F_2&&&\\&&&\ddots&&\\&&&&F_2&\\&&&&&F_4\end{pmatrix},\quad\text{with blocks } F_2:=\begin{pmatrix}p+\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}&-\frac{1}{\gamma_2-\gamma_1}\\-\frac{1}{\gamma_2-\gamma_1}&p+\frac{L-\gamma_1}{(L-\gamma_2)(\gamma_2-\gamma_1)}\end{pmatrix}, \qquad (1.4.25)$$

$F_1:=p+\frac{L}{\gamma_2(L-\gamma_2)}$ and $F_4:=p+\frac{L}{\gamma_1(L-\gamma_1)}$. The block F_2 appears $N-2$ times on the diagonal. The matrix N_N has the following structure

$$N_N:=\begin{pmatrix}&a&b&&&&&\\d&&&&&&&\\&&&a&b&&&\\&b&c&&&&&\\&&&&&\ddots&\ddots&\\&&&b&c&&&\\&&&&\ddots&\ddots&a&b\\&&&&&&&d\\&&&&&&b&c&\end{pmatrix} \qquad (1.4.26)$$

where $a:=p-\frac{\gamma_2}{\gamma_1(\gamma_2-\gamma_1)}$, $b:=\frac{1}{\gamma_2-\gamma_1}$, $c:=p-\frac{L-\gamma_1}{(L-\gamma_2)(\gamma_2-\gamma_1)}$, $d:=p-\frac{L}{\gamma_1(L-\gamma_1)}$.

Theorem 1.4.6. *Suppose that $\gamma_1+\gamma_2 = L$ and $s_j^{+,-} = p$, $\forall\, j$. Then, the optimized Schwarz method is scalable for the solution of problem* (1.7) *with Dirichlet boundary conditions on each F_i, in the sense that $\rho(T_N^D) \le C < 1$, independently of N for every $p > 0$.*

Proof. We want to show that $\rho(T_N^D) \le C < 1$ for every $p > 0$. To do so, we observe that $\rho(T_N^D) = \rho(M_N^{-1}N_N) = \rho(N_N M_N^{-1}) \le \|N_N M_N^{-1}\|_\infty$. Direct calculations show that

$$\|N_N M_N^{-1}\|_\infty = \max\left\{\left|\frac{p\gamma_2(L-\gamma_2)-L}{p\gamma_2(L-\gamma_2)+L}\right|, \frac{2p(L-\gamma_2)^2+\left|L+(L-2\gamma_2)(L-\gamma_2)^2p^2\right|}{(p(L-\gamma_2)+1)(p(L-\gamma_2)(2\gamma_2-L)+L)}\right\}.$$

The first term is clearly less than 1 for every $p > 0$. Considering the second term, we distinguish two cases: if $L+(L-2\gamma_2)(L-\gamma_2)^2p^2 < 0$, then $p > \frac{\sqrt{L}}{\sqrt{2\gamma_2-L}(L-\gamma_2)}$ and the second term simplifies to $\frac{-1+(L-\delta 2)p}{1+(L-\delta 2)p}$ which is positive in the admissible range of p and strictly less than 1. Similarly, if $L+(L-2\gamma_2)(L-\gamma_2)^2p^2 \ge 0$, then we have $\frac{2p(L-\gamma_2)^2+\left|L+(L-2\gamma_2)(L-\gamma_2)^2p^2\right|}{(p(L-\gamma_2)+1)(p(L-\gamma_2)(2\gamma_2-L)+L)} = \frac{p(L-\delta 2)(2\gamma_2-L)-L}{p(L-\delta 2)(-2\delta 2+L)-L} < 1$. Thus we conclude that $\|N_N M_N^{-1}\|_\infty < C$, with $C < 1$ for every $p > 0$ and independent of N. □

We have shown that for a simplified DFN, the OSM is scalable for the solution of problem (1.7) with Dirichlet boundary conditions on the boundary of the fractures. We remind that for one-dimensional chains of fixed size-subdomains, the OSM does not scale [28]. What happens if we consider Neumann boundary conditions? Unfortunately, as one could guess from the discussion in Section 3.1 and 8 of [28], the method does not scale and thus the spectral radius of the iteration matrix tends to one as the number of fractures increases. Imposing Neumann boundary conditions except on the left of the first fracture and on the right of the last one, the general solution becomes

$$e_1^n = \frac{\hat{e}_1^n\tau_1}{\gamma_2}\chi([0,\gamma_2]) + \hat{e}_1^n\chi([\gamma_2,L]), \tag{1.4.27}$$

$$e_j^n = \hat{e}_j^{1,n}\chi([0,\gamma_1]) + \left(\frac{\hat{e}_j^{1,n}(\gamma_2-\tau_j)}{\gamma_2-\gamma_1} + \frac{\hat{e}_j^{2,n}(\tau_j-\gamma_1)}{\gamma_2-\gamma_1}\right)\chi([\gamma_1,\gamma_2]) + \hat{e}_j^{2,n}\chi([\gamma_2,L]), \quad j = 2,\dots,N-1,$$

$$e_N^n = \hat{e}_N^n\chi([0,\gamma_1]) + \frac{\hat{e}_N^n(L-\tau_N)}{L-\gamma_1}\chi([\gamma_1,L]). \tag{1.4.28}$$

We can again obtain a recurrence relation $\mathbf{v}^n = T_N^N\mathbf{v}^{n-1} = \widetilde{M}_N^{-1}\widetilde{N}_N\mathbf{v}^{n-1}$, where the matrices $\widetilde{M}_N$ $\widetilde{N}_N$ have the same structure of (1.4.25) and (1.4.26), but with blocks $\widetilde{F}_1 := p+\frac{1}{\gamma_2}$, $\widetilde{F}_4 := p+\frac{1}{L-\gamma_1}$, $\widetilde{F}_2 := \begin{pmatrix} p+\frac{1}{\gamma_2-\gamma_1} & -\frac{1}{\gamma_2-\gamma_1} \\ -\frac{1}{\gamma_2-\gamma_1} & p+\frac{1}{(\gamma_2-\gamma_1)}\end{pmatrix}$, $\widetilde{a} := p-\frac{1}{\gamma_2-\gamma_1}$, $\widetilde{b} := \frac{1}{\gamma_2-\gamma_1}$, $\widetilde{c} := \widetilde{a}$, $\widetilde{d} := p-\frac{1}{(L-\gamma_1)}$.

In Fig 1.8, we show how the spectral radii of T_N^D and T_N^N vary as the number of fractures increases. We clearly observe that while the spectral radius of T_N^D remains bounded below

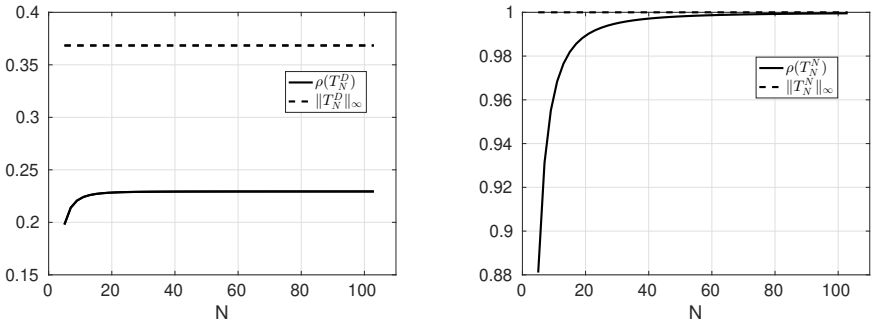


Figure 1.8: Behaviour of the spectral radii of T_N^D and T_N^N when increasing the number of fractures. Parameters: $L = 1$,$\gamma_1 = 0.2$ and $\gamma_2 = 0.6$.

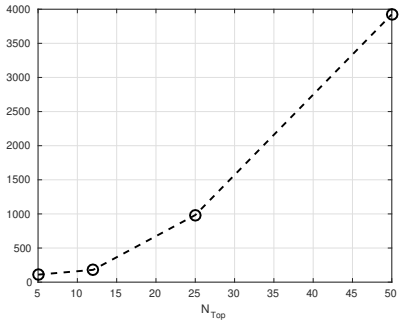


Figure 1.9: Number of iterations to reach a tolerance of tol = 10^{-10} as N_{Top} increases. The DFN is made by 2003 fractures.

one, the spectral radius of T_N^N tends rapidly to one, indicating that we need a larger number of iterations to reach a fixed tolerance as N increases. This phenomenon has been already explained in [28] and it is due to the fact that the constants require about $N/2$ iterations to start contracting in the fractures in the middle of the network, see Figure 3 in [28]. Fortunately, the situation is generally more favourable in realistic simulations. First of all, every fracture which intersects the boundary of the cube/square, where the fractures are generated, will have Dirichlet boundary conditions. Second, the number of subdomains is generally less than N, since each subdomain will be composed by several fractures. Therefore the convergence of the method will depend on the maximum over the subdomains of the minimum distance between each subdomain and the Dirichlet boundary. To imitate this behaviour, we simulate a one dimensional DFN which has Neumann boundary conditions everywhere except on the first and last fracture. Then we modify the DFN imposing Dirichlet boundary conditions every N_{Top} fractures. In Fig. 1.9, we show how the convergence of the method depends on N_{Top}. A subdomain Ω_j such that $\partial\Omega_j \cap \Gamma_D = \emptyset$ is called "floating subdomain". We conclude that the decomposition of the DFN into subdomains should try to minimize the number of floating subdomains. The use of a coarse level to obtain scalability even with Neumann boundary conditions is

currently under study.

1.4.2.2 Optimization of the transmission conditions

The rate of convergence of OSMs depends strongly on the transmission conditions. Therefore it is important to have good estimates of the parameters $s_j^{+,-}$. In the rest of the subsection, we derive optimized transmission conditions for two fractures and we show that these optimized transmission conditions are very efficient in the many fracture case as well.

We consider two fractures F_1 and F_2, and we suppose that on the two fractures we may have discontinuous diffusion coefficients ν_1 and ν_2. The OSM for the error equation is

$$-\nu_1\partial_{\tau_1\tau_1}e_1^n=0,\quad \text{in}\quad F_1,\quad e_1^n(0)=0 \qquad -\nu_2\partial_{\tau_2\tau_2}e_2^n=0,\quad \text{in}\quad F_2,\quad e_2^n(L)=0,$$
$$\left[\left[\frac{\partial e_1^n}{\partial\tau_1}\right]\right]+s_1^-e_1^n=-\left[\left[\frac{\partial e_2^{n-1}}{\partial\tau_2}\right]\right]+s_1^-e_2^{n-1},\quad \text{on } S_1, \qquad \left[\left[\frac{\partial e_2^n}{\partial\tau_2}\right]\right]+s_1^+e_2^n=-\left[\left[\frac{\partial e_1^{n-1}}{\partial\tau_1}\right]\right]+s_1^+e_1^{n-1},\quad \text{on } S_1.$$

We now want to find the best parameters s_1^-, s_1^+, to have the fastest convergence possible. The general solutions are given by

$$e_1^n=\frac{\hat{e}_1^n\tau_1}{\gamma_2}\chi([0,\gamma_2])+\frac{\hat{e}_1^n(L-\tau_1)}{L-\gamma_2}\chi([\gamma_2,L]), \tag{1.4.29}$$
$$e_2^n=\frac{\hat{e}_2^n\tau_1}{\gamma_1}\chi([0,\gamma_1])+\frac{\hat{e}_2^n(L-\tau_1)}{L-\gamma_1}\chi([\gamma_1,L]), \tag{1.4.30}$$

where the unknowns are the two coefficients $\hat{e}_1^n$ and $\hat{e}_2^n$. Inserting these solutions in the transmission conditions we obtain

$$\hat{e}_1^n\left(\frac{\nu_1L}{\gamma_2(L-\delta 2)}+s_1^-\right)=\hat{e}_2^{n-1}\left(-\frac{\nu_2L}{\gamma_1(L-\delta 1)}+s_1^-\right),$$
$$\hat{e}_2^n\left(+\frac{\nu_2L}{\gamma_1(L-\delta 1)}+s_1^+\right)=\hat{e}_1^{n-1}\left(\frac{\nu_1L}{\gamma_2(L-\delta 2)}+s_1^+\right).$$

Rescaling the index, one obtains $\hat{e}_1^n=\rho(s_1^-,s_1^+,\nu_1,\nu_2)\hat{e}_1^{n-2}$ and $\hat{e}_2^n=\rho(s_1^-,s_1^+,\nu_1,\nu_2)\hat{e}_2^{n-2}$, where

$$\rho(s_1^-,s_1^+,\nu_1,\nu_2)=\frac{\left(\frac{\nu_2L}{\gamma_1(L-\delta 1)}-s_1^-\right)\left(\frac{\nu_1L}{\gamma_2(L-\delta 2)}-s_1^+\right)}{\left(\frac{\nu_1L}{\gamma_2(L-\delta 2)}+s_1^-\right)\left(\frac{\nu_2L}{\gamma_1(L-\delta 1)}+s_1^+\right)}.$$

It follows from the structure of $\rho(s_1^-,s_1^+,\nu_1,\nu_2)$ that if we chose $s_1^-=s_1^{-,\text{opt}}:=\frac{\nu_2L}{\gamma_1(L-\delta 1)}$ and $s_1^+=s_1^{+,\text{opt}}:=\frac{\nu_1L}{\gamma_2(L-\delta 2)}$, we would have $\rho(s_1^-,s_1^+,\nu_1,\nu_2)=0$. In other words we would have a nilpotent method, i.e. a method which converges to the exact solution in a finite number of steps, in this case two steps. We refer to [32] for a discussion about the nilpotent property of some domain decomposition methods. We conclude this subsection showing that the two fracture analysis can provide good estimates for the Robin parameters also in the multifracture case. In Figure 1.10, we plot the spectral radius of T_N^D for 2,5,103 and 203 fractures with Dirichlet boundary conditions.

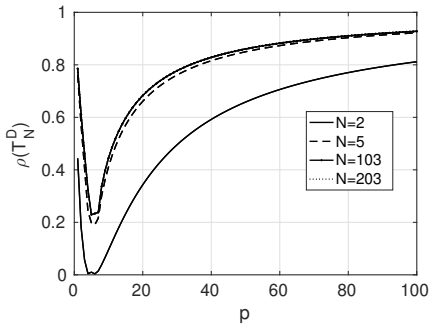


Figure 1.10: Behaviour of the spectral radii of T_N^D when varying the Robin parameter p. Parameters: $L = 1, \gamma_1 = 0.2$, $\gamma_2 = 0.6$ and $\nu_1 = \nu_2 = 1$.

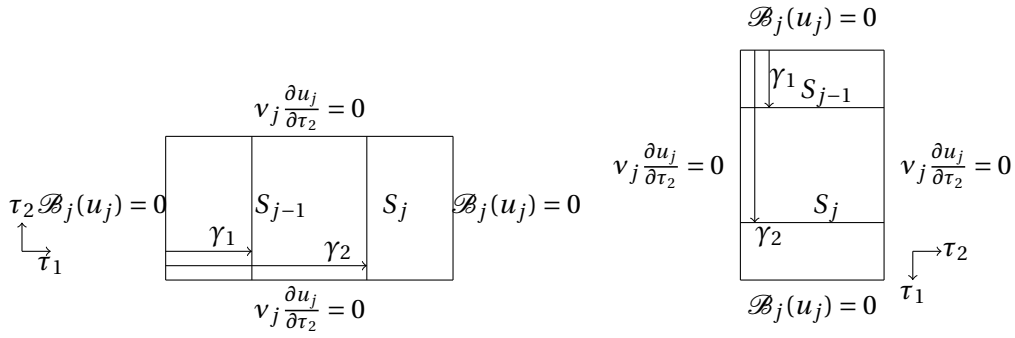


Figure 1.11: Geometry of a two dimensional fracture.

1.4.2.3 A simplified 2-D model

In this section we consider the two dimensional extension of the 1D model. We suppose that each fracture is a two dimensional plane which can be rotated and translated such that it corresponds to Fig 1.11. The intersections between planes are called traces, denoted by S_j, and are straight segments crossing the whole fracture. On each fracture we consider a local reference system of coordinates $\{\tau_1, \tau_2\}$. The coordinate τ_1 is perpendicular to the traces while τ_2 is parallel. We consider the case in which all $\mathscr{B}_j$ represent Dirichlet boundary conditions.

1.4.2.4 Convergence and scalability analysis

We work directly on the error equation. Under the geometry hypothesis, the error can be expanded in Fourier series in each fracture, i.e. $e_j = \sum_{k=0}^{\infty} \tilde{e}_j(\tau_1, k) \cos(\frac{k\pi}{L}\tau_2)$. The Fourier coefficients $\tilde{e}_j(\tau_1, k)$ are obtained imposing the boundary conditions and the transmis-

sion conditions. The general solutions can be written for $k>0$

$$\tilde{e}_1^n(\tau_1,k)=\hat{e}_1^n(k)\frac{\sinh(\frac{k\pi}{L}\tau_1)}{\sinh(\frac{k\pi}{L}\gamma_2)}\chi([0,\gamma_2])+\hat{e}_1^n(k)\frac{\sinh(\frac{k\pi}{L}(L-\tau_1))}{\sinh(\frac{k\pi}{L}(L-\gamma_2))}\chi([\gamma_2,L]), \tag{1.4.31}$$

$$\begin{aligned}\tilde{e}_j(\tau_1,k)=&\hat{e}_j^{1,n}(k)\frac{\sinh(\frac{k\pi}{L}\tau_1)}{\sinh(\frac{k\pi}{L}\gamma_1)}\chi([0,\gamma_1])+\hat{e}_j^{2,n}(k)\frac{\sinh(\frac{k\pi}{L})(L-\tau_1)}{\sinh(\frac{k\pi}{L}(L-\gamma_2))}\chi([\gamma_2,L]) \\ &+\left(\hat{e}_j^{1,n}(k)\frac{\sinh(\frac{k\pi}{L}(\gamma_2-\tau_1))}{\sinh(\frac{k\pi}{L}(\gamma_2-\gamma_1))}+\hat{e}_2^{2,n}(k)\frac{\sinh(\frac{k\pi}{L}(\tau_1-\gamma_1))}{\sinh(\frac{k\pi}{L}(\gamma_2-\gamma_1))}\right)\chi([\gamma_1,\gamma_2]),\quad j=2,...,N-1,\end{aligned} \tag{1.4.32}$$

$$\tilde{e}_N^n(\tau_1,k)=\hat{e}_N^n(k)\frac{\sinh(\frac{k\pi}{L}\tau_1)}{\sinh(\frac{k\pi}{L}\gamma_1)}\chi([0,\gamma_1])+\hat{e}_N^n(k)\frac{\sinh(\frac{k\pi}{L}(L-\tau_1))}{\sinh(\frac{k\pi}{L}(L-\gamma_1))}\chi([\gamma_1,L]), \tag{1.4.33}$$

while for $k=0$,

$$\tilde{e}_1^n(\tau_1,0)=\frac{\hat{e}_1^n(0)\tau_1}{\gamma_2}\chi([0,\gamma_2])+\frac{\hat{e}_1^n(0)(L-\tau_1)}{L-\gamma_2}\chi([\gamma_2,L]), \tag{1.4.34}$$

$$\begin{aligned}\tilde{e}_j^n(\tau_1,0)=&\frac{\hat{e}_j^{1,n}(0)\tau_j}{\gamma_1}\chi([0,\gamma_1])+\left(\frac{\hat{e}_j^{1,n}(0)(\gamma_2-\tau_j)}{\gamma_2-\gamma_1}+\frac{\hat{e}_j^{2,n}(0)(\tau_j-\gamma_1)}{\gamma_2-\gamma_1}\right)\chi([\gamma_1,\gamma_2]) \\ &+\frac{\hat{e}_j^{2,n}(0)(L-\tau_j)}{L-\gamma_2}\chi([\gamma_2,L]),\quad j=2,...,N-1,\end{aligned} \tag{1.4.35}$$

$$\tilde{e}_N^n(\tau_1,0)=\frac{\hat{e}_N^n(0)\tau_N}{\gamma_1}\chi([0,\gamma_1])+\frac{\hat{e}_N^n(0)(L-\tau_N)}{L-\gamma_1}\chi([\gamma_1,L]). \tag{1.4.36}$$

We remark that the unknowns $\hat{e}_j^{i,n}(k)$ are the values attained by the k-th mode of the Fourier expansions at each trace. The Fourier expansion is generally truncated at $k_{\max}=N_h\approx\frac{1}{h}$, where N_h is the number of discretization points on the traces. Indeed, the numerical grid is only capable of representing a certain number of frequencies which depends on the mesh size. Similarly to the 1D case, one can obtain recurrence relations which link the Fourier coefficients of one fracture at iteration n as functions of the Fourier coefficients of the neighbouring fractures at iteration $n-1$. In particular for $k=0$, $\mathbf{v}_0^n:=\left(\hat{e}_1^n(0),\hat{e}_2^{1,n}(0),\hat{e}_2^{2,n}(0),\ldots,\hat{e}_N^n(0)\right)$ satisfies $\mathbf{v}_0^n=T_N^D\mathbf{v}_0^{n-1}$, where T_N^D is the matrix of the 1D system with Dirichlet boundary conditions. For $k>0$, we obtain instead $\mathbf{v}_k^n=T_N^{2D}(k)\mathbf{v}_k^{n-1}$, where $T_N^{2D}=M_{2D}^{-1}N_{2D}$ has the same block structure of the 1D case but with blocks defined as

$$F_2:=\begin{pmatrix}p+\coth(\frac{k\pi}{L}\gamma_1)+\coth(\frac{k\pi}{L}(\gamma_2-\gamma_1)) & -\frac{1}{\coth(\frac{k\pi}{L}(\gamma_2-\gamma_1)}\\ -\frac{1}{\sinh(\frac{k\pi}{L}(\gamma_2-\gamma_1)} & p+\coth(\frac{k\pi}{L}(L-\gamma_2))+\coth(\frac{k\pi}{L}(\gamma_2-\gamma_1))\end{pmatrix},$$

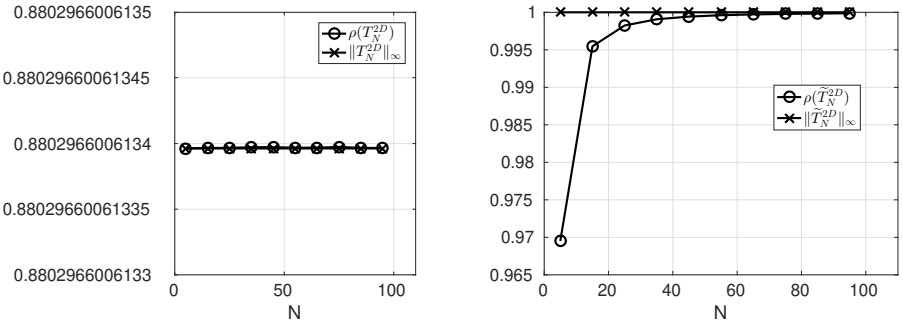


Figure 1.12: Behaviour of the spectral radii of the iteration matrix for OSMs applied to the 2D DFN. On the left with Dirichlet boundary conditions on the vertical edges and on the right with Neumann boundary conditions everywhere. Parameters: $L = 1, \gamma_1 = 0.4, \gamma_2 = 0.8$ and $p = 20$.

$F_1 := p + \coth(\frac{k\pi}{L}\gamma_2) + \coth(\frac{k\pi}{L}(L-\gamma_2))$ and $F_4 := p + \coth(\frac{k\pi}{L}(L-\gamma_1)) + \coth(\frac{k\pi}{L}(\gamma_1))$. On the other hand the coefficients of N_{2D} are

$$a := p - \coth\left(\frac{k\pi}{L}\gamma_1\right) - \coth\left(\frac{k\pi}{L}(\gamma_2-\gamma_1)\right), \quad b := \frac{1}{\sinh(\frac{k\pi}{L}(\gamma_2-\gamma_1)}, \tag{1.4.37}$$

$$c := p - \coth\left(\frac{k\pi}{L}(L-\gamma_2)\right) - \coth\left(\frac{k\pi}{L}(\gamma_2-\gamma_1)\right), \quad d := p - \coth\left(\frac{k\pi}{L}(L-\gamma_2)\right) - \coth\left(\frac{k\pi}{L}(\gamma_2-\gamma_1)\right). \tag{1.4.38}$$

Fig 1.12 shows numerically that the OSM is scalable also for a 2D DFN with Dirichlet boundary conditions. Observing that the frequency $k = 0$ behaves according to the 1D analysis, one can easily guess that the OSM with Neumann boundary conditions on each fracture except the first and last fracture does not scale. Repeating the calculations one finds an iteration matrix $\widetilde{T}_N^{2D}$ and Fig 1.12 confirms this intuition.

1.4.2.5 Optimization of the transmission conditions

We now analyse in more details the convergence behaviour of OSMs for two 2D fractures, focusing on establishing optimized transmission conditions. We consider the problem

$$-\nu_1\Delta e_1^n = 0, \quad \text{in} \quad F_1, \quad e_1^n = 0 \text{ on } \partial F_1 \qquad -\nu_2\Delta e_2^n = 0, \quad \text{in} \quad F_2, \quad e_2^n(L) = 0 \text{ on } \partial F_1,$$

$$\left[\left[\frac{\partial e_1^n}{\partial \tau_1}\right]\right] + s_1^- e_1^n = -\left[\left[\frac{\partial e_2^{n-1}}{\partial \tau_1}\right]\right] + s_1^- e_2^{n-1}, \quad \text{on } S_1, \quad \left[\left[\frac{\partial e_2^n}{\partial \tau_1}\right]\right] + s_1^+ e_2^n = -\left[\left[\frac{\partial e_1^{n-1}}{\partial \tau_1}\right]\right] + s_1^+ e_1^{n-1}, \quad \text{on } S_1.$$

On the one hand, inserting the Fourier expansion into the transmission conditions and defining

$$f_1(k) := \frac{\nu_2 k\pi}{L}\left(\coth\left(\frac{k\pi}{L}\gamma_2\right) + \coth\left(\frac{k\pi}{L}(L-\gamma_2)\right)\right), \quad f_2(k) := \frac{\nu_1 k\pi}{L}\left(\coth\left(\frac{k\pi}{L}\gamma_1\right) + \coth\left(\frac{k\pi}{L}(L-\gamma_1)\right)\right),$$

we obtain for $k>0$,

$$\hat{e}_1^n(k) f_2(k) + s_1^- = -\hat{e}_2^{n-1}(k) f_1(k) + s_1^-,$$
$$\hat{e}_2^n(k) f_1(k) + s_1^+ = -\hat{e}_2^{n-1}(k) f_2(k) + s_1^+.$$

Rescaling the index we get $\hat{e}_j^n(k) = \rho(k, s_1^-, s_1^+)\hat{e}_j^{n-2}(k)$, for $k>0, j=1,2$, where $\rho(k, s_1^-, s_1^+) := \frac{f_1(k)-s_1^-}{f_2(k)+s_1^-}\frac{f_2(k)-s_1^+}{f_1(k)+s_1^+}$. On the other hand, concerning the constant mode $k=0$, we recover the 1D result, $\hat{e}_1^n(0) = \rho_{1D}(s_1^-, s_1^+)\hat{e}_1^{n-2}(0)$ where

$$\rho_{1D}(s_1^-, s_1^+) = \frac{\left(\frac{\nu_2 L}{\gamma_1(L-\delta 1)} - s_1^-\right)\left(\frac{\nu_1 L}{\gamma_2(L-\delta 2)} - s_1^+\right)}{\left(\frac{\nu_1 L}{\gamma_2(L-\delta 2)} + s_1^-\right)\left(\frac{\nu_2 L}{\gamma_1(L-\delta 1)} + s_1^+\right)}.$$

To derive optimized transmission conditions, we have to solve the min-max problem

$$\min_{s_1^-, s_1^+ \in \mathbb{R}^+} \max\left\{\rho_{1D}(s_1^-, s_1^+,), \max_{k\in[1,k_{\max}]} \rho(k, s_1^-, s_1^+)\right\}. \tag{1.4.39}$$

To solve (1.4.39) we use some classical tools from the theory of OSMs. Guided by [95], we set $s_1^- = f_1(p)$ and $s_1^+ = f_2(p)$, for some $p \in \mathbb{R}^+$. This ansatz allows us to rescale the Robin parameters of the two fractures according to the physical properties of the problem such as the diffusion coefficients ν_1 and ν_2. We will see in Chapter 2 that rescaling the parameters s_1^- and s_1^+ permits to take advantage of heterogeneity, i.e. the OSM becomes faster the more different ν_1 and ν_2 are. Moreover we remark that f_j, $j=1,2$ being positive functions, we satisfy the constraint $s_1^-, s_1^+ \in \mathbb{R}^+$. Hence we have simplified problem (1.4.39) to

$$\min_{p\in\mathbb{R}^+} \max\left\{\rho_{1D}(p), \max_{k\in[1,k_{\max}]} \rho(k,p)\right\}. \tag{1.4.40}$$

We now observe that $\rho(k,p)$ is not defined at $k=0$, since the hyperbolic cotangent has a singularity. However we observe that $\lim_{k\to 0}\rho(k,p) = \frac{\left(\frac{\nu_2 L}{\gamma_1(L-\delta 1)} - f_1(p)\right)\left(\frac{\nu_1 L}{\gamma_2(L-\delta 2)} - f_2(p)\right)}{\left(\frac{\nu_1 L}{\gamma_2(L-\delta 2)} + f_1(p)\right)\left(\frac{\nu_2 L}{\gamma_1(L-\delta 1)} + f_2(p)\right)} = \rho_{1D}(p)$. Thus we introduce the function $\widetilde{\rho}(k,p) = \rho(k,p)$ for $k>0$ and $\widetilde{\rho}(0,p) = \rho_{1D}(p)$. Using this result we can further simply the min-max problem to

$$\min_{p\in\mathbb{R}^+} \max_{k\in[0,k_{\max}]} \widetilde{\rho}(k,p). \tag{1.4.41}$$

and we proof the following result.

Theorem 1.4.7. *The solution of the min-max problem* (1.4.41) *is given by the unique* p^* *which satisfies* $\widetilde{\rho}(0,p) = \widetilde{\rho}(k_{\max}, p)$.

Proof. We first observe that both f_1 and f_2 are positive and increasing functions for $k>0$, while $\widetilde{\rho}(k,p) = 0$ only if $k=p$. We compute the derivative with respect to p and we obtain

$$\frac{\partial\widetilde{\rho}(k,p)}{\partial p} = \frac{(f_1(k)+f_2(k))(f_2(p)-f_2(k))(f_1(k)+f_2(p))f_1'(p) + f_2'(p)(f_2(k)+f_1(p))(f_1(p)-f_1(k))}{(f_2(k)+f_1(p))^2(f_1(k)+f_2(p))^2},$$

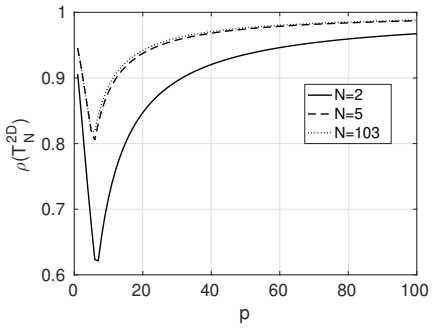


Figure 1.13: Behaviour of the spectral radii of the T_N^D when varying the Robin parameter p. Parameters: $L = 1$,$\gamma_1 = 0.2$, $\gamma_2 = 0.6$ and $\nu_1 = \nu_2 = 1$

where $f_j'(p)$ stands for the derivative of f_j with respect to p. We set $h(k,p) := \frac{\partial \widetilde{\rho}(k,p)}{\partial p}$. We remark that if $p > k_{\max}$ then, $\frac{\partial \widetilde{\rho}(k,p)}{\partial p} > 0$ for all $k \in [0, k_{\max}]$, and we are not at the optimum. Thus we conclude that $p \in (0, k_{\max})$. For symmetry we have $\frac{\partial \widetilde{\rho}(k,p)}{\partial k} = h(p,k)$, and we conclude that the function is decreasing until $k = p$ and then it is increasing until $k = k_{\max}$, hence the function has two local maxima located in $k = 0$ and $k = k_{\max}$. Therefore, $\max_{k\in[0,k_{\max}]} |\widetilde{\rho}(k,p)| = \max\{\widetilde{\rho}(0,p), \widetilde{\rho}(k_{\max},p)\}$. Now since $\frac{\partial \widetilde{\rho}(0,p)}{\partial p} > 0$ and $\frac{\partial \widetilde{\rho}(k_{\max},p)}{\partial p} < 0$ $\forall\, p \in (0, k_{\max}]$, by continuity the optimal p^* satisfies $\widetilde{\rho}(0,p^*) = \widetilde{\rho}(k_{\max},p^*)$. The uniqueness of p^* follows from the strict sign of $\frac{\partial \widetilde{\rho}(k,p)}{\partial p}$ for $k = 0, k_{\max}$. □

In Figure 1.13 we plot the spectral radius of the iteration matrix for 2,5 and 103 fractures as function of p. We remark that, as in the one dimensional case, the analysis for only two fractures provides good estimates for the optimal parameters in the many fracture case.

CHAPTER 2

Heterogeneous optimized Schwarz methods for Second Order PDEs

"We begin with a few remarks concerning the choice of parameters (γ_{ij}): first of all, it is possible to replace, say in the case $m = 2$, $\lambda = \lambda_{12}$ and $\mu = \lambda_{21}$, by two arbitrary constants,(...), or even by two proportional functions,(...), or even by local or nonlocal operators(...)"

— Lions, On the Schwarz alternating method III: a variant for nonoverlapping subdomains, *1990*

Due to their property of convergence in the absence of overlap, optimized Schwarz methods (OSMs) are the natural domain decomposition framework for heterogeneous problems, where the spatial decomposition is provided by the multi-physics of the phenomena. Their origins lie in the pioneering paper [126], in which Lions proposed a convergent nonoverlapping algorithm using Robin transmission conditions. In a small paragraph of [126], reported in the epigraph of this Chapter, Lions discusses briefly generalizations of this method, using more general transmission conditions on the interfaces. It is really incredible how seminal these few lines are and how much research derived from them! Nowadays, there are standard procedures to obtain so called optimized transmission conditions. The problem of interest is posed in a simplified setting where one can use the Fourier transform [74], for unbounded domains, or Fourier series expansion or more generally separation of variables [100, 95], for bounded domains, to transform the PDE into a set of ODEs parametrized by the frequencies k. Then, solving the ODEs and using the transmission conditions, one can get a recursive relation for the Fourier coefficients and obtain a closed formula for the convergence factor which contains some free parameters to optimize. The literature regarding OSMs for homogeneous problems is well developed. Optimized transmission conditions have been obtained for many problems such as Helmholtz equations [91, 81], Maxwell equations [58, 121, 137], advection diffusion problems [65, 95], Navier Stokes equations [12], shallow water equations [133]

and Euler equations [62]. In all these works, homogeneous problems are analyzed, in the sense that a unique physics is considered in the whole domain, and therefore the coupling on the interfaces regards equations of the same nature. First attempts to generalize this situation have been carried out in [131], [78], where Laplace equations with different diffusion coefficients were considered, and in [60], which was devoted to Maxwell equations with discontinuous coefficients. Let us remark that at least two possible interpretations of heterogeneous domain decomposition methods exist. The first one concerns problems where the same physical phenomenon is taking place in the whole domain, but it can be convenient to use a cheaper approximation in some parts of the domain in order to save computational resources. This might be the case in the presence of boundary layers, or for example in CFD simulations where a potential flow is used far away from the zone of interest while the Navier-Stokes equations are fully solved near, for instance, an aircraft. In this situation, good transmission conditions can be obtained through a factorization approach, see [82] for further details. The second interpretation assumes that two different physical phenomena are present in the domain and they interact through an interface. In this case some physical coupling conditions must be satisfied along the common interface, such as the continuity of the function and its normal derivative for second order PDEs, or the continuity of normal stresses for fluid-structure problems. An examples in this direction can be found in [102] where a partial optimization procedure was carried out for a fluid-structure problem. For this kind of heterogeneous problems, a domain decomposition approach can be extremely useful since it allows to reuse specific solvers designed for the different physical phenomena present in the domain. For instance, one can use a finite volume solver where a strong advection is present while using a multigrid solver where diffusion dominates or an ad-hoc linear elasticity solver combined with a CFD code for the Navier-Stokes equations. In this perspective, OSMs lead to a significantly better convergence of the coupling routine with respect to other domain decomposition algorithms (e.g. Dirichlet-Neumann, Neumann-Neumann) since they take into account the physical properties in their transmission conditions. We refer the interested reader to [122, 123] for the application of OSMs for the coupling of atmospheric and oceanic computational simulation models. Optimized transmission conditions have also been applied in time dependent PDEs [158] and in electrical circuits [85, 87, 86, 119].

This chapter is based on [97] and [96]. In these works we studied heterogeneous problems which arise from the coupling of second order PDEs. In [97], we focused on elliptic PDEs and we proved theoretical results and asymptotic formulae both for single and double sided optimizations. From our analysis, it follows that OSMs do not suffer from the heterogeneity, it is the opposite, they are faster the stronger the heterogeneity is. It is even possible to have h independent convergence choosing two independent Robin parameters. This property was proved for a Laplace equation with discontinous coefficients, but only conjectured for more general couplings in [78]. We then focus on the coupling between the Helmholtz equation and the Laplace equation [96]. In this case, the well-possessedness of the problem is not obvious and thus we investigate it in detail. Then we derive optimized transmission conditions. The last part of this Chapter aims to answer

the question: what can we do if the theoretical analysis is not applicable to my case of interest? Am I doomed to use other domain decomposition methods? No, actually you are not. In fact, we show that using a technique called probing it is possible to obtain numerically optimized transmission conditions for very complicated and general problems in a inexpensive way.

2.1 Reaction Diffusion-Diffusion coupling

Let us consider two domains $\Omega_1 := (-\infty, 0) \times (0, L)$ and $\Omega_2 := (0, +\infty) \times (0, L)$ and the interface $\Gamma := \{0\} \times (0, L)$. In this Section we study a reaction-diffusion equation with discontinuous coefficients along the interface Γ,

$$(\eta^2(x) - \nu(x)\Delta)u = f \quad \text{in} \quad \Omega, \tag{2.1.1}$$

where $\Omega := \Omega_1 \cup \Omega_2$, $\eta^2(x) = \eta^2 \geq 0$ in Ω_1 and $\eta(x) = 0$ in Ω_2, while $\nu(x) = \nu_1$ in Ω_1 and $\nu(x) = \nu_2$ in Ω_2, with $\nu_1, \nu_2 \in \mathbb{R}^+$. Equation (2.1.1) is closed by homogeneous Dirichlet boundary conditions on the horizontal edges and assuming $\lim_{x\to\pm\infty} u = 0$. The OSM for this problem is

$$\begin{aligned}(\eta^2 - \nu_1\Delta)u_1^n = f \text{ in } \quad \Omega_1,\ (\nu_1\partial_x + S_1)(u_1^n)(0,\cdot) = (\nu_2\partial_x + S_1)(u_2^{n-1})(0,\cdot),\\ -\nu_2\Delta u_2^n = f \text{ in } \quad \Omega_2,\ (\nu_2\partial_x - S_2)(u_2^n)(0,\cdot) = (\nu_1\partial_x - S_2)(u_1^{n-1})(0,\cdot),\end{aligned}$$

where S_j, $j = 1, 2$ are linear operators along the interface Γ in the y direction. The goal is to find which operators guarantee the best performance in terms of convergence speed. We consider the error equation whose unknowns are $e_i^n := u_{|\Omega_i} - u_i^n$, $i = 1, 2$, and we expand the solutions in the Fourier basis in the y direction, $e_i^n = \sum_{k\in\mathcal{V}} \hat{e}_i^n(x,k)\sin(ky)$, $i = 1, 2$ with $\mathcal{V} := \{\frac{\pi}{L}, \frac{2\pi}{L}, \dots\}$. Moreover we suppose that the operator S_j are diagonalizable, with eigenvectors $\psi_k(y) := \sin(ky)$, such that $S_j\psi_k = \sigma_j(k)\psi_k$, where $\sigma_j(k)$ are the eigenvalues of S_j. Under these assumptions, we find that the coefficients $\hat{e}_i^n$ satisfy,

$$\begin{array}{rcll}(\eta^2 - \nu_1\partial_{xx} + \nu_1 k^2)(\hat{e}_1^n) &=& 0, & k \in \mathcal{V},\ x < 0,\\ (\nu_1\partial_x + \sigma_1(k))(\hat{e}_1^n)(0,k) &=& (\nu_2\partial_x + \sigma_1(k))(\hat{e}_2^{n-1})(0,k), & k \in \mathcal{V},\\ (-\nu_2\partial_{xx} + \nu_2 k^2)(\hat{e}_2^n) &=& 0, & k \in \mathcal{V},\ x > 0,\\ (\nu_2\partial_x - \sigma_2(k))(\hat{e}_2^n)(0,k) &=& (\nu_1\partial_x - \sigma_2(k))(\hat{e}_1^{n-1})(0,k), & k \in \mathcal{V}.\end{array} \tag{2.1.2}$$

Solving the two differential equations parametrized by k in (2.1.2), imposing that the solutions remain bounded for $x \to \pm\infty$ and defining $\lambda(k) := \sqrt{k^2 + \tilde{\eta}^2}$ and $\gamma(k) := k$, we obtain

$$\begin{array}{rcll}\hat{e}_1^n &=& \hat{e}_1^n(0,k)e^{\sqrt{k^2+\tilde{\eta}^2}x} = \hat{e}_1^n(0,k)e^{\lambda(k)x} & \text{in} \quad \Omega_1,\\ \hat{e}_2^n &=& \hat{e}_2^n(0,k)e^{-kx} = \hat{e}_2^n(0,k)e^{-\gamma(k)x} & \text{in} \quad \Omega_2,\end{array} \tag{2.1.3}$$

where $\tilde{\eta}^2 = \frac{\eta^2}{\nu_1}$. The transmission conditions in (2.1.2) allow us to express the Fourier coefficient at iteration n of the solution in one subdomain as function of the coefficient of the solution in the other subdomain at the previous iteration $n-1$, namely

$$\hat{e}_1^n(0,k) = \frac{-\nu_2\gamma(k) + \sigma_1(k)}{\nu_1\lambda(k) + \sigma_1(k)}\hat{e}_2^{n-1}(0,k), \tag{2.1.4}$$

and

$$\hat{e}_2^n(0,k) = \frac{\nu_1\lambda(k) - \sigma_2(k)}{-\nu_2\gamma(k) - \sigma_2(k)}\hat{e}_1^{n-1}(0,k). \tag{2.1.5}$$

Combining (2.1.4) and (2.1.5) we get

$$\hat{e}_1^n(0,k) = \frac{-\nu_2\gamma(k) + \sigma_1(k)}{\nu_1\lambda(k) + \sigma_1(k)} \cdot \frac{\nu_1\lambda(k) - \sigma_2(k)}{-\nu_2\gamma(k) - \sigma_2(k)}\hat{e}_1^{n-2}(0,k).$$

By induction we deduce

$$\hat{e}_1^{2n}(0,k) = \rho^n\hat{e}_1^0(0,k), \quad \hat{e}_2^{2n}(0,k) = \rho^n\hat{e}_2^0(0,k),$$

where the convergence factor ρ is defined by

$$\rho := \rho(k,\sigma_1,\sigma_2) = \frac{-\nu_2\gamma(k) + \sigma_1(k)}{\nu_1\lambda(k) + \sigma_1(k)} \cdot \frac{\nu_1\lambda(k) - \sigma_2(k)}{-\nu_2\gamma(k) - \sigma_2(k)}.$$

Expressing the dependence on the Fourier frequency k we get

$$\rho(k,\sigma_1,\sigma_2) = \frac{-\nu_2 k + \sigma_1(k)}{\nu_1\sqrt{k^2+\tilde{\eta}^2} + \sigma_1(k)} \cdot \frac{\nu_1\sqrt{k^2+\tilde{\eta}^2} - \sigma_2(k)}{-\nu_2 k - \sigma_2(k)}. \tag{2.1.6}$$

A closer inspection of (2.1.6) leads us to conclude that if we chose the operators S_j such that their eigenvalues are

$$\sigma_1^{\text{opt}}(k) := \nu_2 k \quad \text{and} \quad \sigma_2^{\text{opt}}(k) := \nu_1\sqrt{k^2+\tilde{\eta}^2}, \tag{2.1.7}$$

then we would have $\rho \equiv 0$. In this case the algorithm would converge in just two iterations. This option, even tough it is optimal, leads to non local operators S_j^{opt}, which correspond to the Schur complements [136], and they are expensive from the computational point of view. Indeed, the operator associated to the eigenvalues $\sigma_1^{\text{opt}}(k) := \nu_2 k$ corresponds to the square root of the Laplacian on the interface Γ, i.e. $S_1^{\text{opt}} = \nu_2(-\Delta_\Gamma)^{\frac{1}{2}}$ which is a fractional and non local operator. The non-local property of S_1^{opt} can also be understood considering a discretization of the straight interface Γ and the discrete counterpart of S_1^{opt}, i.e. $S_{1h}^{\text{opt}} := \nu_2(-\Delta_{y,h})^{\frac{1}{2}}$ where $-\Delta_{y,h} = \text{Diag}(-1,2,-1)$ is the classical 1-D Laplacian. A direct implementation shows that the matrix S_{1h}^{opt} is dense. Even though the use of S_{1h}^{opt} would destroy the sparsity of the subdomain matrices, theoretically it could still be used as a transmission condition and the method would then converge in two iterations. However, the major drawback is that in general we do not know the operator S_j^{opt} and therefore we would have to assemble numerically the Schur complements. This is an operation which requires the knowledge of the inverse of the subdomain operators and therefore it is computationally expensive.

We thus look for classes of convenient transmission conditions which are amenable to easy implementation, and then to find which transmission conditions among a specific class lead to the best convergence factor. We consider here zeroth order approximations

of the optimal operators in (2.1.7) which correspond to classical Robin conditions on the interface. In order to get the best transmission conditions in terms of convergence speed, we have to minimize the maximum of the convergence factor over all the frequencies k. Defining $\mathscr{D}_1, \mathscr{D}_2$ as the classes of transmission conditions, we are looking for a couple $(\sigma_1^*, \sigma_2^*) \in \mathscr{D} := \mathscr{D}_1 \times \mathscr{D}_2$ such that

$$(\sigma_1^*, \sigma_2^*) = \underset{(\sigma_1,\sigma_2)\in\mathscr{D}}{\arg\min} \Big(\max_{k_{\min}\le k\le k_{\max}} |\rho(k,\sigma_1,\sigma_2)| \Big). \tag{2.1.8}$$

The lower and upper bounds $k_{\min}$, $k_{\max}$ depend on the problem under study: $k_{\min}$ is given by the Fourier expansion and here it is equal to $k_{\min} = \frac{\pi}{L}$. The presence of $k_{\min}$ in (2.1.8), is the "memory" that our problem has of the boundness of the domain. We refer the interested reader to [76, 101, 100, 95] for more details on the influence of the domain on OSMs. The upper bound $k_{\max}$ is instead the maximum frequency that can be resolved by the grid and it is typically estimated as $k_{\max} = \frac{\pi}{h}$ where h is a measure of the grid spacing.

2.1.1 Zeroth order single sided optimized transmission conditions

Let p be a free parameter, we define

$$\sigma_1(k) = \nu_2 p, \quad \sigma_2(k) = \nu_1 \sqrt{\tilde{\eta}^2 + p^2}. \tag{2.1.9}$$

We have made this choice because the optimal operators in (2.1.7) are clearly rescaled according to the diffusion constants of the two subdomains and thus we imitate this behaviour. Furthermore we introduce the parameter $\tilde{\eta}^2$ in the definition of $\sigma_2(k)$ in order to make the problem amenable to analytical treatment. With this choice, we have $\sigma_j(k) = \sigma_j^{opt}(k)$ for $k = p$; in other words, for the frequency $k = p$, the transmission conditions lead to an exact solver which converges in two iterations. The idea of introducing free parameters such that the eigenvalues $\sigma_j(k)$ are identical to the optimal ones for a certain frequency is essential, because as we will see in the following, it allows us to solve the min-max problems which, for a generic choice of σ_j, are extremely hard to solve.

Inserting (2.1.9) into (2.1.6), the min-max problem (2.1.8) becomes

$$\min_{p\in\mathbb{R}} \max_{k_{\min}\le k\le k_{\max}} \left| \frac{k-p}{k+\lambda\sqrt{p^2+\tilde{\eta}^2}} \cdot \frac{\sqrt{k^2+\tilde{\eta}^2} - \sqrt{p^2+\tilde{\eta}^2}}{\sqrt{k^2+\tilde{\eta}^2}+\frac{p}{\lambda}} \right|, \tag{2.1.10}$$

where $\lambda = \frac{\nu_1}{\nu_2}$. We define $\rho(k,p) := \frac{k-p}{k+\lambda\sqrt{p^2+\tilde{\eta}^2}} \cdot \frac{\sqrt{k^2+\tilde{\eta}^2}-\sqrt{p^2+\tilde{\eta}^2}}{\sqrt{k^2+\tilde{\eta}^2}+\frac{p}{\lambda}}$. We are now solving the min-max problem (2.1.10). The main steps are the following:

- Restricting the range in which we are searching for p.
- Identifying the candidates for the maxima in the variable k.

- Studying how the maxima behave when varying the parameter p.

Lemma 2.1.1 (Restriction for the interval of p)**.** *If p^* is a solution to problem* (2.1.10) *then p^* belongs to the interval* $[k_{\min}, k_{\max}]$.

Proof. First we note that $|\rho(k,p)| < |\rho(k,-p)|$ for every $p \geq 0$. Therefore we can assume $p^* \in \mathbb{R}^+$. Moreover the function is always positive and equal to zero only for $k = p$. Thus we can neglect the absolute value. Direct calculations show that $\frac{\partial \rho(k,p)}{\partial p} = h(k,p)$ where

$$h(k,p) := \frac{(p-k)\lambda p(\sqrt{k^2+\tilde{\eta}^2}\lambda + k)}{(k+\lambda\sqrt{p^2+\tilde{\eta}^2})^2(\sqrt{k^2+\tilde{\eta}^2}\lambda + p)\sqrt{p^2+\tilde{\eta}^2}} + \frac{(\sqrt{p^2+\tilde{\eta}^2} - \sqrt{k^2+\tilde{\eta}^2})\lambda(\sqrt{k^2+\tilde{\eta}^2}\lambda + k)}{(k+\lambda\sqrt{p^2+\tilde{\eta}^2})(\sqrt{k^2+\tilde{\eta}^2}\lambda + p)^2}. \quad (2.1.11)$$

We observe that if $p^* < k_{\min}$ then $\frac{\partial \rho}{\partial p}(k,p^*) < 0$ for all $k \in [k_{\min}, k_{\max}]$, hence we are for sure not at the optimum since increasing p^* would decrease the convergence factor for all the frequencies $k \in [k_{\min}, k_{\max}]$.
On the other hand if $p^* > k_{\max}$ then we have $\frac{\partial \rho}{\partial p}(k,p^*) > 0 \ \forall k \in [k_{\min}, k_{\max}]$. Hence we cannot be at the optimum either since decreasing p^* would decrease $\rho(k,p) \ \forall k \in [k_{\min}, k_{\max}]$. Thus we can conclude that if p^* is a solution of (2.1.10), then p^* lies in the interval $[k_{\min}, k_{\max}]$. □

Now we focus on the search of the maxima of $\rho(p,k)$ with respect to k knowing that $p \in [k_{\min}, k_{\max}]$.

Lemma 2.1.2 (Local maxima in k)**.** *For any fixed value of $p \in [k_{\min}, k_{\max}]$, the function $k \to \rho(k,p)$ assumes its maximum either at $k = k_{\min}$ or at $k = k_{\max}$.*

Proof. We consider the derivative of $\rho(k,p)$ with respect to k and we recall that $\rho(k,p)$ is always positive so we may neglect the absolute value. Direct calculations show that $\frac{\partial \rho}{\partial k} = h(p,k)$. Thus considering (2.1.11) we have that letting $p \in (k_{\min}, k_{\max})$, $\frac{\partial \rho}{\partial k} < 0, \forall k < p$, and $\frac{\partial \rho}{\partial k} > 0, \forall k > p$. Therefore the maximum is attained on the boundary, either at $k = k_{\min}$ or $k = k_{\max}$.
On the other hand, if $p = k_{\min}$, $\rho(k, k_{\min})$ has a zero in $k = k_{\min}$. For all the other values of k in the interval $[k_{\min}, k_{\max}]$, the function is strictly increasing and therefore the maximum is attained at $k = k_{\max}$. The case $p = k_{\max}$ is identical and hence the result follows. □

We now have all the ingredients to solve the min-max problem (2.1.10).

Theorem 2.1.3. *The unique optimized Robin parameter p^* solving the min-max problem* (2.1.10) *is given by the unique root of the non linear equation*

$$|\rho(k_{\min}, p^*)| = |\rho(k_{\max}, p^*)|. \quad (2.1.12)$$

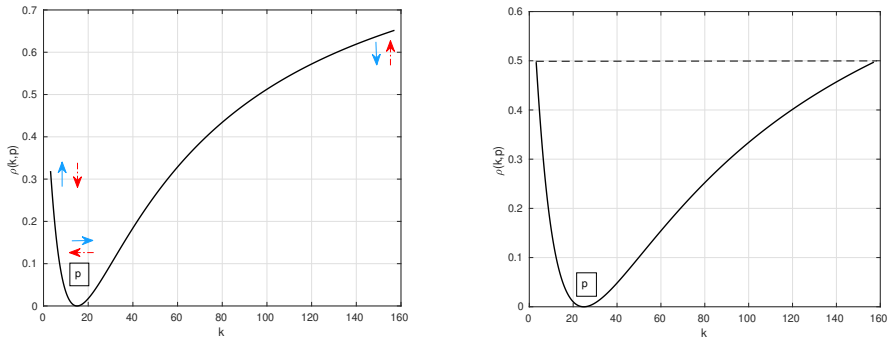


Figure 2.1: Illustration of the equioscillation property described in Theorem 2.1.3.

Proof. From the previous lemmas, we know that we can rewrite problem (2.1.10) as

$$\min_{p\in[k_{\min},k_{\max}]} \max\left\{\rho(k_{\min},p),\rho(k_{\max},p)\right\},$$

i.e. the maximum is either attained at $k=k_{\min}$ or $k=k_{\max}$. We now show that the optimal p^* satisfies a classical equioscillation property [150], see Fig 2.1 for a graphical representation. We first note that $\rho(k_{\min},p)=0$ for $p=k_{\min}$, and $\frac{\partial\rho(k_{\min},p)}{\partial p}>0, \forall\, p\in(k_{\min},k_{\max}]$. Therefore increasing p, $\rho(k_{\min},p)$ strictly increases until it reaches its maximum value for $p=k_{\max}$. On the other hand, we have that $\rho(k_{\max},k_{\min})$ is strictly greater than zero, and while p increases from $k_{\min}$ to $k_{\max}$, $\rho(k_{\max},p)$ decreases, because $\frac{\partial\rho(k_{\max},p)}{\partial p}<0, \forall\, p\in[k_{\min},k_{\max})$. Furthermore we have that $\rho(k_{\max},k_{\max})=0$.
Hence, thanks to the strict monotonicity of both $\rho(k_{\min},p)$ and $\rho(k_{\max},p)$, there exists by continuity a unique value p^* such that $\rho(k_{\min},p^*)=\rho(k_{\max},p^*)$. This value is clearly the optimum, because perturbing p^* would increase the value of ρ at one of the two extrema and therefore the maximum of ρ over all k. □

Even though a closed form solution of (2.1.12) is not known, it is interesting to study asymptotically how the algorithm performs. Therefore we keep ν_1, ν_2 and $\tilde{\eta}^2$ fixed, and $k_{\max}=\frac{\pi}{h}$ while letting $h\to 0$. We introduce the notation $f(h)\sim g(h)$ as $h\to 0$ if and only if $\lim_{h\to 0}\frac{f(h)}{g(h)}=1$.

Theorem 2.1.4. *Let $D:=\sqrt{k_{\min}^2+\tilde{\eta}^2}$. Then if $\nu_1,\nu_2,\tilde{\eta}^2$ are kept fixed, $k_{\max}=\frac{\pi}{h}$ and h is small enough, then the optimized Robin parameter p^* is given by*

$$p^*\sim C\cdot h^{-\frac{1}{2}},\quad C:=\sqrt{\frac{(\lambda D+k_{\min})\pi}{(\lambda+1)}}. \tag{2.1.13}$$

Furthermore the asymptotic convergence factor of the heterogeneous OSM is

$$\max_{k_{\min}\le k\le\pi/h}|\rho(k,p^*)|\sim 1-h^{\frac{1}{2}}\left[\frac{\lambda D}{C}+\frac{D}{C}+\frac{k_{\min}}{\lambda C}+\frac{k_{\min}}{C}\right]. \tag{2.1.14}$$

Proof. We make the ansatz $p = C \cdot h^{-\alpha}$ in the equation (2.1.12). Expanding for small h, we get that

$$|\rho(k_{\min}, p)| \sim 1 - h^{\alpha}\left[\frac{\lambda D}{C} + \frac{D}{C} + \frac{k_{\min}}{\lambda C} + \frac{k_{\min}}{C}\right].$$

On the other hand,

$$|\rho(k_{\max}, p)| \sim 1 - h^{1-\alpha}\left[\frac{\lambda C}{\pi} + \frac{2C}{\pi} + \frac{C}{\lambda\pi}\right].$$

Comparing the first two terms we get the result. □

Remark 2.1.5. Note that if we set $\tilde{\eta}^2 = 0$, then we recover the results for the coupling of two Laplace equations with different diffusion constants, see [78]. In that case,

$$\rho \sim 1 - h^{\frac{1}{2}}\sqrt{\frac{k_{\min}}{\pi}}\left[\frac{(\lambda+1)^2}{\lambda}\right], \qquad p^* = \sqrt{k_{\min}\pi}h^{-\frac{1}{2}}.$$

Moreover we have that the convergence factor (2.1.14) satisfies for $\lambda = \frac{\nu_1}{\nu_2} \to \infty$, $|\rho| \sim 1 - h^{\frac{1}{2}}\lambda\sqrt{\frac{D}{\pi}}$ and for $\lambda \to 0$, $|\rho| \sim 1 - h^{\frac{1}{2}}\frac{1}{\lambda}\sqrt{\frac{k_{\min}}{\pi}}$. On the other hand as $\tilde{\eta} \to \infty$ we have $|\rho| \sim 1 - h^{\frac{1}{2}}\sqrt{\tilde{\eta}}\frac{(\lambda+1)^{\frac{3}{2}}}{\sqrt{\lambda\pi}}$. It follows that for all strong heterogeneity limits, the constant in front of the asymptotic term $h^{\frac{1}{2}}$ becomes larger, therefore the deterioration is slower and the method is more efficient.

2.1.2 Zeroth order two sided optimized transmission conditions

Let us consider now the more general case for Robin transmission conditions, with two free parameters p and q such that the operators S_j have eigenvalues

$$\sigma_1(k) = \nu_2 p, \quad \sigma_2(k) = \nu_1\sqrt{q^2 + \tilde{\eta}^2}.$$

We remark that $\sigma_1(k)$ is exact for the frequency $k = p$ while $\sigma_2(k)$ is exact for frequency $k = q$. Therefore from (2.1.6) we deduce the method converges in two iterations for two frequencies. Letting again $\lambda = \frac{\nu_1}{\nu_2}$, we get

$$\min_{p,q}\max_{k_{\min}\le k\le k_{\max}}|\rho(k,p,q)| = \min_{p,q}\max_{k_{\min}\le k\le k_{\max}}\left|\frac{(k-p)(\sqrt{k^2+\tilde{\eta}^2} - \sqrt{q^2+\tilde{\eta}^2})}{(k+\lambda\sqrt{q^2+\tilde{\eta}^2})(\sqrt{k^2+\tilde{\eta}^2}+\frac{p}{\lambda})}\right|. \quad (2.1.15)$$

Following the same philosophy of the previous section, we start restricting the range in which we need to search for the parameters p and q. Then we focus on the maxima with respect to k and finally we analyse how these maxima behave with respect to p and q.

Lemma 2.1.6 (Restriction for the interval of p, q)**.** *If the couple (p^*, q^*) is a solution to the min-max problem* (2.1.15)*, then we have that both p^* and q^* belong to the interval $[k_{\min}, k_{\max}]$.*

Proof. For $p > 0$, we observe that $|\rho(k,p,q|) < |\rho(k,-p,q)|$ and q is always squared so we can restrict both parameters to be positive without loss of generality. Next we consider the partial derivatives of $|\rho|$ with respect to p and q:

$$\operatorname{sign}\left(\frac{\partial|\rho|}{\partial p}\right) = -\operatorname{sign}(k-p), \quad \operatorname{sign}\left(\frac{\partial|\rho|}{\partial q}\right) = -\operatorname{sign}(k-q). \tag{2.1.16}$$

Repeating the same argument of Lemma 2.1.1, we conclude that we are not at the optimum unless both p and q belong to $[k_{\min}, k_{\max}]$. □

Next we analyse the behaviour of $|\rho(k,p,q)|$ with respect to the variable k.

Lemma 2.1.7 (Local maxima in k)**.** *For $p, q \in [k_{\min}, k_{\max}]$,*

$$\max_{k_{\min}\le k\le k_{\max}} |\rho(k,p,q)| = \max\{|\rho(k_{\min},p,q)|, |\rho(\widetilde{k},p,q)|, |\rho(k_{\max},p,q)|\},$$

where $\widetilde{k}$ is an interior maximum between $[\min(p,q), \max(p,q)]$.

Proof. We first observe that $|\rho(k,p,q)|$ has two zeros, one at $k = p$ and the other at $k = q$. Next we consider the derivative of $\rho(k,p,q)$ with respect to k, and assuming that $p \neq q$[1] we get,

$$\begin{aligned}\frac{\partial\rho(k,p,q)}{\partial k} = &\frac{(\sqrt{k^2+\tilde\eta^2} - \sqrt{q^2+\tilde\eta^2})(\sqrt{k^2+\tilde\eta^2})(\sqrt{k^2+\tilde\eta^2}+\frac{p}{\lambda})(\lambda\sqrt{q^2+\tilde\eta^2}+p)}{D(k,p)} + \\ &+ \frac{(k-p)(k+\lambda\sqrt{q^2+\tilde\eta^2})k(\frac{p}{\lambda}+\sqrt{q^2+\tilde\eta^2})}{D(k,p)}.\end{aligned} \tag{2.1.17}$$

The denominator $D(k,p)$ is always positive. Now we consider the two cases in which $k < \min(p,q)$ and $k > \max(p,q)$: in both we have that $\rho(k,p,q) > 0$, and analyzing equation (2.1.17) we conclude that for $k < \min(p,q)$, $\frac{\partial\rho(k,p)}{\partial k} < 0$ and for $k > \max(p,q)$, $\frac{\partial\rho}{\partial k} > 0$. Hence by continuity of $\partial_k\rho(k,p)$, there exits at least one $\widetilde{k}$, which is a local minimum of $\rho(k,p)$ and a local maximum for $|\rho(k,p)|$ see Fig. 2.2, such that $\partial_k\rho = 0$, and all of them lie in the interval $[\min(p,q), \max(p,q)]$ for p and q fixed. Now we prove that the interior maximum is unique. Indeed the interior maxima for $|\rho(k,p,q)|$ are given by the roots of the equation $\partial_k\rho(k,p) = 0$ which corresponds to

$$\frac{\sqrt{q^2+\tilde\eta^2} - \sqrt{\tilde\eta^2+k^2}}{k+\lambda\sqrt{\tilde\eta^2+q^2}} = \frac{\left(k-p\right)k}{\left(\lambda\sqrt{k^2+\tilde\eta^2}+p\right)\sqrt{k^2+\tilde\eta^2}}. \tag{2.1.18}$$

First we suppose that $p < k < q$. Then we have that the left hand side of (2.1.18) is positive in $k = p$, it is strictly decreasing in k, and it reaches zero at $k = q$. The right hand side

[1] If $p = q$ we are considering the optimization problem discussed in the previous subsection.

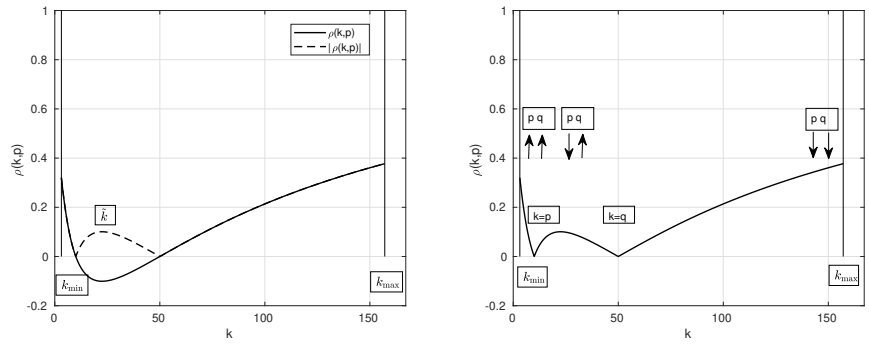


Figure 2.2: The left panel shows an example of the convergence factor with its three local maxima localized at $k = k_{\min}$, $k = k_{\max}$ and $k = \tilde{k}$. On the right we summarizes how these local maxima behave as function of p and q.

of (2.1.18) instead starts from zero and it is strictly increasing. We conclude that there is a unique point $\widetilde{k}$ such that the two sides are equal and hence a unique interior maximum $\hat{k}$ for $|\rho(k,p,q)|$. If instead $q < k < p$, changing the sign of (2.1.18) and dividing by $k/\sqrt{k^2+\tilde{\eta}^2}$, the right hand side is strictly decreasing while the left hand side, computing the derivative, is strictly increasing and hence the same conclusion holds.
We may conclude that the function assumes its maximum either at the interior point $\widetilde{k}$, or at the boundaries of the interval, i.e. $k_{\min}$, $k_{\max}$. □

In the next lemma we prove that the end points $k_{\min}$ and $k_{\max}$ satisfy an equioscillation property as in the previous case of a single parameter p.

Lemma 2.1.8 (Equioscillation at the end points)**.** *The optimized convergence factor* $|\rho(k,p,q)|$ *must satisfy equioscillation at the endpoints, i.e.*

$$|\rho(k_{\min},p^*,q^*)| = |\rho(k_{\max},p^*,q^*)|.$$

Proof. We study how $|\rho(k_{\min},p,q)|$, $|\rho(\widetilde{k},p,q)|$ and $|\rho(k_{\max},p,q)|$ behave as p,q vary and we show that if we do not have equioscillation at the boundary points, we can always improve the convergence factor until equioscillation is reached. Taking into account (2.1.16) we have for every $p,q \in [k_{\min},k_{\max}]$

$$\frac{\partial|\rho(k_{\min},p,q)|}{\partial p} > 0, \quad \frac{\partial|\rho(k_{\min},p,q)|}{\partial q} > 0,$$

$$\frac{\partial|\rho(k_{\max},p,q)|}{\partial p} < 0, \quad \frac{\partial|\rho(k_{\max},p,q)|}{\partial q} < 0.$$

In other words, increasing independently p,q increases $|\rho(k_{\min},p,q)|$ and decreases $|\rho(k_{\max},p,q)|$. We now compute the total derivative of $|\rho(\widetilde{k},p,q)|$ with respect to p and q, which since

we have $\partial_k|\rho(\widetilde{k},p,q)|=0$, corresponds to the partial derivative with respect to the two arguments. One then finds that the sign of $\frac{\partial|\rho(\widetilde{k},p,q)|}{\partial p}$ and $\frac{\partial|\rho(\widetilde{k},p,q)|}{\partial q}$ depends on the position of $\tilde{k}$ with respect to p and q. Indeed it holds

$$\operatorname{sign}\left(\frac{\partial|\rho(\widetilde{k},p,q)|}{\partial p}\right)=\operatorname{sign}(p-\widetilde{k}),\quad \operatorname{sign}\left(\frac{\partial|\rho(\widetilde{k},p,q)|}{\partial q}\right)=\operatorname{sign}(q-\widetilde{k}).$$

The right panel of Fig. 2.2 summarizes the dependence of the local maxima with respect to p and q. Let us suppose that $p<q$, q fixed, and $|\rho(k_{\min},p,q)|<|\rho(k_{\max},p,q)|$. The other cases are treated similarly. We do not make any assumptions on the value of $|\rho(\widetilde{k},p,q)|$. Now if we increase p we decrease $\max\{|\rho(k_{\min},p,q)|,|\rho(\widetilde{k},p,q)|,|\rho(k_{\max},p,q)|\}$ as long as $|\rho(k_{\min},p,q)|\leq|\rho(k_{\max},p,q)|$ and $p\leq q$. If $|\rho(k_{\min},p,q)|=|\rho(k_{\max},p,q)|$ for a certain $p<q$, then we obtain the desired result since we have improved uniformly the convergence factor. Suppose instead that when $p=q$, and therefore $|\rho(\widetilde{k},p,q)|=0$, we still have $|\rho(k_{\min},p,q)|<|\rho(k_{\max},p,q)|$. Thus the convergence factor is equal to $|\rho(k_{\max},p,q)|$. We now set up a process which improves $\max_{[k_{\min},k_{\max}]}|\rho(k,p,q)|$ until we get equioscillation at the boundary points. As long as $|\rho(k_{\min},p,q)|<|\rho(k_{\max},p,q)|$, we increase $p>q$ until $|\rho(\widetilde{k},p,q)|\leq|\rho(k_{\max},p,q)|$. When we reach $|\rho(\widetilde{k},p,q)|=|\rho(k_{\max},p,q)|$, we then increase q until $q=p$. If while increasing q we still have $|\rho(k_{\min},p,q)|<|\rho(k_{\max},p,q)|$, then we repeat the process. Continuing this process, we must reach equioscillation at some point by continuity since when p approaches $k_{\max}$, we must have $|\rho(k_{\min},k_{\max},q)|>|\rho(k_{\max},k_{\max},q)|=0$. At the same time we improved surely the convergence factor since, in spite of the initial value of $|\rho(\widetilde{k},p,q)|$, we have that $\max_{[k_{\min},k_{\max}]}|\rho(k,p,q)|\leq|\rho(k_{\max},p,q)|$ which is decreasing along the process. □

We now have enough tools and insights to prove the main results of this subsection:

Theorem 2.1.9. *There are two pairs of parameters (p_1^*,q_1^*) and (p_2^*,q_2^*) such that we obtain equioscillation between all the three local maxima,*

$$|\rho(k_{\min},p_j^*,q_j^*)|=|\rho(k_{\max},p_j^*,q_j^*)|=|\rho(\hat{k},p_j^*,q_j^*)|\quad j=1,2. \tag{2.1.19}$$

The optimal pair of parameters is the one which realizes the

$$\min_{(p_j^*,q_j^*),j=1,2}|\rho(k_{\min},p_j^*,q_j^*)|. \tag{2.1.20}$$

Proof. Let us define $F_1(p,q):=\rho(k_{\min},p,q)$ and $F_2(p,q):=\rho(k_{\max},p,q)$. Due to Lemma 2.1.8, we know that there exist values (p,q) such that $F:=|F_1(p,q)|-|F_2(p,q)|=0$. We can thus express one parameter, for example q, as a function of the other one, i.e. $q=q(p)$. Although the expression is too complicated to be used for analytical computations, we are able to infer about the structure of $q(p)$. First of all we can state that $q(p=k_{\min})=k_{\max}$ since $|F_1(k_{\min},q(k_{\min}))|=0$ implies that $|F_2(k_{\min},q(k_{\min})|=0$ but then the only choice

possible is $q(k_{\min}) = k_{\max}$. Similarly we have $q(k_{\max}) = k_{\min}$. We next use implicit differentiation to infer about the behaviour of q with respect to p.
Following classical arguments we have that, since $F(p, q(p)) = 0$,

$$0 = \frac{dF(p,q(p))}{dp} = \frac{dF_1(p,q(p)) - dF_2(p,q(p))}{dp} = \frac{\partial F_1 - \partial F_2}{\partial p} + \frac{\partial F_1 - \partial F_2}{\partial q} q'(p),$$

and therefore

$$q'(p) = \frac{\frac{\partial F_2}{\partial p} - \frac{\partial F_1}{\partial p}}{\frac{\partial F_1}{\partial q} - \frac{\partial F_2}{\partial q}}. \tag{2.1.21}$$

Analyzing the sign of each term, we conclude that $q'(p) < 0 \quad \forall\, p \in (k_{\min}, k_{\max})$. Therefore we state that $q(p)$ is a strictly decreasing function which starts from $q(p = k_{\min}) = k_{\max}$ and reaches its minimum at $q(k_{\max}) = k_{\min}$.
Now we have only one free parameter p, since q is constrained to vary such that the equioscillation between the end points is achieved, thus we look for values of p such that we obtain equioscillation between $k_{\min}$ and the interior maximum $\widetilde{k}$.
Let us first study how $\widetilde{F}(p, q) := \rho(\widetilde{k}, p, q(p))$ behaves while p varies. As long as $p \le \widetilde{k} \le q(p)$, we have

$$\operatorname{sign}\left(\frac{\partial |\widetilde{F}(p,q(p))|}{\partial p}\right) = \operatorname{sign}(\sqrt{q(p)^2 + \tilde{\eta}^2} - \sqrt{\widetilde{k}^2 + \tilde{\eta}^2}) \cdot \operatorname{sign}(\widetilde{F}(p,q(p)) < 0,$$
$$\operatorname{sign}\left(\frac{\partial |\widetilde{F}(p,q(p))|}{\partial q}\right) = \operatorname{sign}(p - \widetilde{k}) \cdot \operatorname{sign}(\widetilde{F}(p,q(p)) > 0.$$

Then, keeping in mind the $q'(p) < 0$, $\widetilde{F}(p, q(p))$ is strictly decreasing for all the values of p such that $p < \widetilde{k} < q(p)$,

$$\frac{d|\widetilde{F}(p,q(p))|}{dp} = \frac{\partial |\widetilde{F}(p,q(p))|}{\partial p} + \frac{\partial |\widetilde{F}(p,q(p))|}{\partial q} \cdot q'(p) < 0.$$

Similarly it is straightforward to verify that for $q(p) < \widetilde{k} < p$

$$\frac{d|\widetilde{F}(p,q(p))|}{dp} = \frac{\partial |\widetilde{F}(p,q(p))|}{\partial p} + \frac{\partial |\widetilde{F}(p,q(p))|}{\partial q} \cdot q'(p) > 0.$$

Moreover we have that for $p = \widetilde{k} = q(p)$, $|\widetilde{F}(p, q(p))| = 0$ and $\frac{d|\widetilde{F}(p,q(p))|}{dp} = 0$.
Focusing next on $|F_1(p, q(p))|$ we can state that, neglecting the $\operatorname{sign}(F_1(p, q(p)))$, because it is always positive or zero, the derivatives at the left and right boundary extrema are equal to

$$\frac{d|F_1(k_{\min}, k_{\max})|}{dp} = \frac{\partial |F_1(k_{\min}, k_{\max})|}{\partial p} + \frac{\partial |F_1(k_{\min}, k_{\max})|}{\partial q} q'(p) = \frac{\partial |F_1(k_{\min}, k_{\max})|}{\partial p} > 0,$$

and

$$\frac{d|F_1(k_{\max}, k_{\min})|}{dp} = \frac{\partial |F_1(k_{\max}, k_{\min})|}{\partial p} + \frac{\partial |F_1(k_{\max}, k_{\min})|}{\partial q} q'(p) = \frac{\partial |F_1(k_{\max}, k_{\min})|}{\partial p} < 0.$$

So for values of p in a right neighbourhood of $p = k_{\min}$, $|F_1(p, q(p))|$ increases, while for values of p in a left neighbourhood of $p = k_{\max}$, $|F_1(p, q(p))|$ decreases. Using the monotonicity of $|F(\widetilde{k}, p, q(p))|$ and the fact that when $\widetilde{k} = p = q(p)$, $|F(\widetilde{k}, p, q(p))| = 0$, while $|F(k_{\min}, p, q(p))| > 0$, we conclude that there exists at least one pair (p, q) such that $|F(k_{\min}, p, q(p))| = |F(\widetilde{k}, p, q(p))|$.
We still have to prove that actually there exist only two couples (p_j, q_j) such that equioscillation is achieved. Indeed, if we imagine that $|F_1(p, q(p))|$ had a certain behaviour, for example it oscillates, then we might have more than two pairs. Nevertheless we show that $|F_1(p, q(p))|$ has a unique local maximum for $p \in [k_{\min}, k_{\max}]$ so that only two equioscillations are allowed among all the three local maxima: one while $|\widetilde{F}(p, q(p))|$ decreases, the other one for increasing $|\widetilde{F}(p, q(p))|$.
To do so, we consider $\frac{d|F_1(p,q(p))|}{dp}$ again and substitute (2.1.21),

$$\frac{d|F_1(p, q(p))|}{dp} = \frac{\frac{\partial F_1}{\partial q} \cdot \frac{\partial F_2}{\partial p} - \frac{\partial F_2}{\partial q} \cdot \frac{\partial F_1}{\partial p}}{\frac{\partial F_1}{\partial q} \cdot \frac{\partial F_2}{\partial q}}.$$

The zeros of the derivative are given by the non linear equation

$$(p - k_{\min})(\sqrt{k_{\max}^2 + \tilde{\eta}^2} - \sqrt{k_{\min}^2 + \tilde{\eta}^2}) \frac{\sqrt{k_{\min}^2 + \tilde{\eta}^2} + \frac{p}{\lambda}}{\sqrt{k_{\max}^2 + \tilde{\eta}^2} + \frac{p}{\lambda}} =$$
$$(k_{\max} - p)(\sqrt{q^2 + \tilde{\eta}^2} - \sqrt{k_{\min}^2 + \tilde{\eta}^2}) \frac{k_{\min} + \lambda\sqrt{q^2 + \tilde{\eta}^2}}{k_{\max} + \lambda\sqrt{q^2 + \tilde{\eta}^2}}.$$

It is sufficient to observe that the left hand side starts from 0 and it is strictly increasing in p, while the right hand side starts from a positive value, it decreases with p and it reaches 0 for $p = k_{\max}$. So the equation admits only one solution and therefore the local maximum with respect to p of $|F_1(p, q(p))|$ is unique. The solution to the min-max problem (2.1.15) is the pair of parameters (p^*, q^*) which allows equioscillation among the three local maxima and realizes (2.1.20). Every other pair of parameter would led to the increase of at least one of the local maxima and therefore of the maximum of $|\rho|$ over k. □

In [78], the authors proved a similar result for the Laplace equation with discontinuous coefficients without the presence of the further optimality condition (2.1.20). Their result was based on the possibility to restrict the interval of interest for the parameters to $p < q$ or $q < p$ according to the value of λ. In the present case this is not possible because of the presence of $\tilde{\eta}^2$ which breaks the symmetry of the convergence factor. Therefore we cannot discard a priori one of the two possible equioscillations and the further condition (2.1.20) must be added. Nevertheless in the asymptotic regime for $h \to 0$ and $k_{\max} \to \infty$, the next result allows us to clearly choose the optimal pair as a function of λ, recovering the property of the results for the simplified situation treated in [78].

Theorem 2.1.10. *Let $D := \sqrt{k_{\min}^2 + \tilde{\eta}^2}$. Then if the physical parameters $\tilde{\eta}^2, \nu_1, \nu_2$ are fixed, $k_{\max} = \frac{\pi}{h}$ and h goes to zero, the optimized two-sided Robin parameters are for $\lambda \geq 1$,*

$$\begin{aligned} p_1^* &\sim \tfrac{\lambda(k_{\min}+D)}{\lambda-1} - \tfrac{2\sqrt{2}(1+\lambda)(\lambda D+k_{\min})\lambda^2\sqrt{\pi(k_{\min}+D)}}{\pi\lambda(\lambda-1)^3} h^{\frac{1}{2}}, \\ q_1^* &\sim \tfrac{\pi(\lambda-1)}{2\lambda} h^{-1} + \tfrac{\sqrt{2}(1+\lambda)^2\sqrt{\pi(k_{\min}+D)}}{2\lambda(\lambda-1)} h^{-\frac{1}{2}}, \\ \max_{k_{\min}\leq k\leq \pi/h} &|\rho(k, p_1^*, q_1^*)| \sim \tfrac{1}{\lambda} - \tfrac{2\sqrt{2}(1+\lambda)\sqrt{(k_{\min}+D)}}{\sqrt{\pi}\lambda(\lambda-1)} h^{\frac{1}{2}}, \end{aligned} \tag{2.1.22}$$

and for $\lambda < 1$ we have

$$\begin{aligned} p_2^* &\sim \tfrac{1}{2}\pi(1-\lambda)h^{-1} + \tfrac{\sqrt{2}(1+\lambda)^2\sqrt{\pi(D+k_{\min})}}{2(1-\lambda)} h^{-\frac{1}{2}}, \\ q_2^* &\sim \sqrt{\left(\tfrac{D+k_{\min}}{1-\lambda}\right)^2 - \tilde{\eta}^2} - \tfrac{2\sqrt{2}(D+k_{\min})^2(\lambda+1)(\lambda D+k_{\min})}{(\lambda-1)^4\sqrt{\pi(D+k_{\min})}\sqrt{\frac{D+k_{\min}}{1-\lambda}-\tilde{\eta}^2}} h^{\frac{1}{2}}, \\ \max_{k_{\min}\leq k\leq \pi/h} &|\rho(k, p_2^*, q_2^*)| \sim \lambda - \tfrac{2\sqrt{2}\lambda(1+\lambda)\sqrt{(k_{\min}+D)}}{\sqrt{\pi}(1-\lambda)} h^{\frac{1}{2}}. \end{aligned} \tag{2.1.23}$$

Proof. Guided by numerical experiments, for $\lambda \geq 1$ we make the ansatz $p \sim C_p + Ah^{\frac{1}{2}}$, $q \sim Qh^{-1} + Bh^{-\frac{1}{2}}$, and $\hat{k} = C_k h^{-\frac{1}{2}}$. First of all considering the equation $\partial_k \rho(\widetilde{k}, p, q) = 0$, we find setting to zero the first non zero term $C_k = \sqrt{C_p \cdot Q}$. Inserting this into (2.1.19) and comparing the two leading terms, we get the result. Similarly for $\lambda < 1$, we make the ansatz $p \sim C_p h^{-1} + Ah^{-\frac{1}{2}}$, $q \sim Q + Bh^{\frac{1}{2}}$ and $\hat{k} = C_k h^{-\frac{1}{2}}$ and we get $C_k = \sqrt{C_p\sqrt{Q^2+\tilde{\eta}^2}}$. Substituting and matching the leading order terms we obtain the result. □

If we set $\tilde{\eta}^2 = 0$, then $D = k_{\min}$ and we recover the results of [78]. Note that in contrast to the one sided case, the convergence factor does not deteriorate to 1 as $h \to 0$, but it is bounded either by $\frac{1}{\lambda}$ if $\lambda \geq 1$ or by λ if $\lambda < 1$, so we obtain a non-overlapping OSM that converges independently of the mesh size h. We emphasize that the heterogeneity makes the method faster instead of presenting a difficulty. A heuristic explanation is that the heterogeneity tends to decouple the problems, making them less dependent one from the other. In contrast with other domain decomposition methods, OSMs can be tuned according to the physics and therefore they can benefit from this decoupling.

2.2 Advection Reaction Diffusion-Reaction Diffusion coupling

In this Section, we consider the domain decomposition described at the beginning of Section 2.1. In Ω_1 we have a reaction diffusion equation, while in Ω_2 we have an advection reaction diffusion equation. We allow the reaction and diffusion coefficients to be different among the subdomains. The OSM reads

$$\begin{aligned} (\eta_1^2 - \nu_1\Delta)u_1^n &= f, \quad \text{in} \quad \Omega_1, \\ (\nu_1\partial_x + S_1)(u_1^n)(0,\cdot) &= (\nu_2\partial_x - \mathbf{a}\cdot(1,0)^\top + S_1)(u_2^{n-1})(0,\cdot), \\ (\eta_2^2 + \mathbf{a}\cdot\nabla - \nu_2\Delta)u_2^n &= f, \quad \text{in} \quad \Omega_2, \\ (\nu_2\partial_x - \mathbf{a}\cdot(1,0)^\top - S_2)(u_2^n)(0,\cdot) &= (\nu_1\partial_x - S_2)(u_1^{n-1})(0,\cdot), \end{aligned} \tag{2.2.1}$$

where $\mathbf{a} = (a_1, a_2)^\top$. The additional term in the transmission conditions arises from the conservation of the flux in divergence form, see Chapter 6 in [139]. We first suppose $a_2 = 0$. Then we can solve the error equations in the subdomains through separation of variables and we obtain $e_i^n = \sum_{k\in\mathcal{V}} \hat{e}_i^n \sin(ky)$, $i = 1,2$, where

$$\hat{e}_1^n(k,x) = A^n(k)e^{\sqrt{\frac{\eta_1^2}{\nu_1}+k^2}x} \qquad \hat{e}_2^n(k,x) = B^n(k)e^{\lambda_-(k)x},$$

and $\lambda_-(k) := \frac{a_1-\sqrt{a_1^2+4\nu_2^2k^2+4\nu_2\eta_2^2}}{2\nu_2}$. Inserting e_1, e_2 into the transmission conditions we get

$$\nu_1\sqrt{\frac{\eta_1^2}{\nu_1}+k^2}A^n(k)+\sigma_1(k)A^n(k) = \nu_2\lambda_-(k)B^{n-1}(k) - a_1B^{n-1}(k)+\sigma_1(k)B^{n-1}(k),$$

$$\nu_2\lambda_-(k)B^n(k) - a_1B^n(k) - \sigma_2(k)B^n(k) = \nu_1\sqrt{\frac{\eta_1^2}{\nu_1}+k^2}A^{n-1}(k) - \sigma_2(k)A^{n-1}(k).$$

The convergence factor is therefore given by

$$\rho(k,\sigma_1,\sigma_2) = \frac{\nu_2\lambda_-(k)-a_1+\sigma_1(k)}{\nu_1\sqrt{\tilde{\eta}_1^2+k^2}+\sigma_1(k)}\frac{\nu_1\sqrt{\tilde{\eta}_1^2+k^2}-\sigma_2(k)}{\nu_2\lambda_-(k)-a_1-\sigma_2(k)},$$

where $\tilde{\eta}_1^2 = \frac{\eta_1^2}{\nu_1}$. We rewrite $\lambda_-(k)$ as $\lambda_-(k) = \frac{a_1}{2\nu_2} - \sqrt{k^2+\delta^2}$ with $\delta^2 = \frac{a_1^2}{4\nu_2^2}+\frac{\eta_2^2}{\nu_2}$. Using the dependence on k, the convergence factor becomes

$$\rho(k,\sigma_1,\sigma_2) = \frac{\nu_2\sqrt{k^2+\delta^2}+\frac{a_1}{2}-\sigma_1(k)}{\nu_1\sqrt{\tilde{\eta}_1^2+k^2}+\sigma_1(k)}\frac{\nu_1\sqrt{\tilde{\eta}_1^2+k^2}-\sigma_2(k)}{\nu_2\sqrt{k^2+\delta^2}+\frac{a_1}{2}+\sigma_2(k)}.$$

We can define two optimal operators S_j^{opt} associated to the eigenvalues $\sigma_1^{\text{opt}}(k) := \nu_2\sqrt{k^2+\delta^2}+\frac{a_1}{2}$ and $\sigma_2^{\text{opt}}(k) := \nu_1\sqrt{k^2+\tilde{\eta}_1^2}$ which lead to convergence in just two iterations.

2.2.1 Zeroth order single sided optimized transmission conditions

Following the strategy of the previous section, we choose $\sigma_1(k), \sigma_2(k)$ so that they coincide with the optimal choice for the frequency $k = p$, i.e. $\sigma_1(k) = \nu_2\sqrt{p^2+\delta^2}+\frac{a_1}{2}$ and $\sigma_2(k) = \nu_1\sqrt{p^2+\tilde{\eta}_1^2}$. Defining $\lambda := \frac{\nu_1}{\nu_2}$, the convergence factor becomes

$$\rho(k,p) = \frac{\sqrt{k^2+\tilde{\eta}_1^2}-\sqrt{p^2+\tilde{\eta}_1^2}}{\frac{1}{\lambda}\left(\sqrt{k^2+\delta^2}+\frac{a_1}{2\nu_2}\right)+\sqrt{p^2+\tilde{\eta}_1^2}}\cdot\frac{\sqrt{k^2+\delta^2}-\sqrt{p^2+\delta^2}}{\lambda\sqrt{k^2+\tilde{\eta}_1^2}+\left(\sqrt{p^2+\delta^2}+\frac{a_1}{2\nu_2}\right)}. \tag{2.2.2}$$

Theorem 2.2.1. *The unique optimized Robin parameter p^* solving the min-max problem*

$$\min_{p\in\mathbb{R}}\max_{k_{\min}\le k\le k_{\max}}|\rho(k,p)|,$$

is given by the unique root of the non linear equation

$$|\rho(p^*, k_{\min})| = |\rho(p^*, k_{\max})|.$$

Proof. The proof is very similar to the proof of Theorem 2.1.3, therefore we just sketch the main steps. We start observing that $\rho(k,p)$ has only one zero located at $k=p$ and $\rho(k,p)>0 \quad \forall k,p$. Thus we may neglect the absolute value. Analysing the derivative with respect to p, we find

$$\operatorname{sign}\left(\frac{\partial \rho(k,p)}{\partial p}\right) = -\operatorname{sign}(k-p).$$

This implies that $\frac{\partial \rho(k,p)}{\partial p} > 0$ if $k<p$ and $\frac{\partial \rho(k,p)}{\partial p} < 0$ if $k>p$. We conclude that p must lie in the interval $[k_{\min}, k_{\max}]$. Similarly the derivative with respect to k satisfies $\frac{\partial \rho(k,p)}{\partial k} < 0$ if $k<p$ and $\frac{\partial \rho(k,p)}{\partial k} > 0$ if $k>p$. Hence, the local maxima with respect to k are located at the boundary points $k=k_{\min}$ and $k=k_{\max}$. Repeating the final argument of Theorem 2.1.3 we get the result. □

Since a closed form formula is again not available, we study the asymptotic behaviour for the optimal parameter p^* when taking finer and finer meshes.

Theorem 2.2.2. *If the physical parameters are fixed, $k_{\max}=\frac{\pi}{h}$ and h is small enough, then the optimized Robin parameter p^* satisfies*

$$p^* \sim C_a \cdot h^{-\frac{1}{2}},\ C_a = \frac{\sqrt{\nu_2\,(\lambda+1)\,\pi\left(2\sqrt{k_{\min}^2+\tilde{\eta}_1^2}\lambda\,\nu_2 + 2\sqrt{k_{\min}^2+\delta^2}\nu_2 - a_1\right)}}{\sqrt{2}\nu_2\,(\lambda+1)}.$$

Furthermore the asymptotic convergence factor is

$$\max_{k_{\min}\le k\le \pi/h} |\rho(k,p^*)| \sim 1 - h^{\frac{1}{2}}\left(\frac{C_a\,(\lambda+1)^2}{\lambda\,\pi}\right).$$

Proof. We insert the ansatz $p = C_a \cdot h^{-\alpha}$ into the equation (2.1.12). Expanding for small h, we get that

$$\rho(p, k_{\min}) \sim 1 - h^{\alpha}\left(\frac{C_a\,(\lambda+1)^2}{\lambda\,\pi}\right).$$

On the other hand,

$$\rho(p, k_{\max}) \sim 1 + h^{-\alpha+1}\left(\frac{1}{2}\,\frac{(\lambda+1)\left(-2\sqrt{k_{\min}^2+\tilde{\eta}_1^2}\lambda\,\nu_2 - 2\sqrt{k_{\min}^2\delta^2}\nu_2 + a_1\right)}{C_a\,\nu_2\,\lambda}\right).$$

Comparing the first two terms we get the result. □

2.2.2 Zeroth order two sided optimized transmission conditions

In this paragraph we generalize the previous transmission conditions, introducing another degree of freedom q. The operators S_j are such that their eigenvalues are

$$\sigma_1(k) = \nu_2\sqrt{q^2+\delta^2} + \frac{a_1}{2}, \qquad \sigma_2(k) = \nu_1\sqrt{p^2+\tilde{\eta}_1^2},$$

and the convergence factor becomes

$$\rho(k,p) = \frac{\sqrt{k^2+\tilde{\eta}_1^2} - \sqrt{p^2+\tilde{\eta}_1^2}}{\frac{1}{\lambda}\left(\sqrt{k^2+\delta^2}+\frac{a_1}{2\nu_2}\right)+\sqrt{p^2+\tilde{\eta}_1^2}} \cdot \frac{\sqrt{k^2+\delta^2}-\sqrt{q^2+\delta^2}}{\lambda\sqrt{k^2+\tilde{\eta}_1^2}+\left(\sqrt{q^2+\delta^2}+\frac{a_1}{2\nu_2}\right)}.$$

In order to prove a similar result as in Theorem 2.1.9, we suppose that $\tilde{\eta}_1 = 0$, i.e. only diffusion is present in Ω_1, and $a_1 > 0$, i.e. the advection flux is pointing into the subdomain Ω_2.

Theorem 2.2.3. *There are two pairs of parameters (p_1^*, q_1^*) and (p_2^*, q_2^*) such that we obtain equioscillation between all the three local maxima located at the boundary extrema $k_{\min}, k_{\max}$ and at the interior point $\widetilde{k}$,*

$$|\rho(k_{\min}, p_j^*, q_j^*)| = |\rho(k_{\max}, p_j^*, q_j^*)| = |\rho(\widetilde{k}, p_j^*, q_j^*)| \quad j = 1,2.$$

The optimal pair of parameters is the one which realizes the

$$\min_{(p_j^*,q_j^*),j=1,2} |\rho(k_{\min}, p_j^*, q_j^*)|.$$

Proof. Similarly to the proof of Theorem 2.1.9, we observe that the function admits two zeros, one located at $k = p$, the other at $k = q$ due to the choice of the transmission operators. Computing the derivatives with respect to p and q we get

$$\operatorname{sign}\left(\frac{\partial|\rho|}{\partial p}\right) = -\operatorname{sign}(\rho)\cdot\operatorname{sign}(k-q) = -\operatorname{sign}(k-p),$$
$$\operatorname{sign}\left(\frac{\partial|\rho|}{\partial q}\right) = -\operatorname{sign}(\rho)\cdot\operatorname{sign}(k-p) = -\operatorname{sign}(k-q).$$

We conclude that, at the optimum, both p and q lie in $[k_{\min}, k_{\max}]$, i.e. the function at the optimum has two zeros in the interval. Now we study the behaviour with respect to k. Computing the derivative with respect to k, we find that the potential local maxima are given by the roots of

$$\frac{\sqrt{\delta^2+k^2}-\sqrt{\delta^2+q^2}}{k(\lambda k+\sqrt{q^2+\delta^2}+\frac{a_1}{2\nu_2})} = \frac{p-k}{\sqrt{k^2+\delta^2}\left(p\lambda+\sqrt{k^2+\delta^2}+\frac{a_1}{2\nu_2}\right)}.$$

With some algebraic manipulations, we find that a sufficient condition such that $\frac{p-k}{(p\lambda+\sqrt{k^2+\delta^2}+a_1/(2\nu_2))}$ has a monotonic behaviour with respect to k is that $a_1 > 0$. Letting p, q in $[k_{\min}, k_{\max}]$, we have that the local maxima of the function are located at $k_{\min}, k_{\max}, \widetilde{k}$. Moreover we have

$$\begin{aligned} &\frac{\partial|\rho|}{\partial p}|_{k=k_{\min}} > 0, && \frac{\partial|\rho|}{\partial q}|_{k=k_{\min}} > 0, \\ &\frac{\partial|\rho|}{\partial p}|_{k=k_{\max}} < 0, && \frac{\partial|\rho|}{\partial q}|_{k=k_{\max}} < 0, \\ &\frac{\partial|\rho|}{\partial p}|_{k=\widetilde{k}} < 0, && \frac{\partial|\rho|}{\partial q}|_{k=\widetilde{k}} > 0. \end{aligned} \tag{2.2.3}$$

We can thus repeat the same arguments as in the proof of Theorem 2.1.9 since all steps are now exclusively based on the sign of the partial derivatives with respect to the parameters, see (2.2.3), and the result follows. □

Theorem 2.2.4. *Let $D := \sqrt{k_{\min}^2 + \delta^2}$. If the physical parameters $\tilde{\eta}_2^2, \nu_1, \nu_2, a_1$ are fixed, $k_{\max} = \frac{\pi}{h}$ and h goes to zero, the optimized two-sided Robin parameters are for $\lambda \geq 1$,*

$$p_1^* \sim P_1 h^{-1} + E_1 h^{-\frac{1}{2}},\ q_1^* \sim Q_1 - F_1 h^{\frac{1}{2}},\ \max_{k_{\min}\leq k\leq \frac{\pi}{h}} |\rho(k, p_1^*, q_1^*)| \sim \lambda - \tfrac{E_1\pi(\lambda+1)}{(P_1\lambda+\pi)^2} h^{\frac{1}{2}},$$

with

$$\begin{aligned} P_1 &:= \frac{\pi(\lambda-1)}{2\lambda}, \quad Q_1 := \sqrt{\frac{D + k_{\min} + \frac{a_1}{2\nu_2\lambda}}{1-\frac{1}{\lambda}} - \delta^2}, \\ E_1 &:= \frac{(2(P_1\sqrt{\delta^2+Q_1^2}+C_h^2)(\lambda+1)\nu_2 + P_1 a_1)(\lambda P_1+\pi)^2}{2\lambda^2 P_1 \nu_2 C_h \pi(\lambda+1)}, \\ F_1 &:= \frac{(2(P_1\sqrt{\delta^2+Q_1^2}+C_h^2)(\lambda+1)\nu_2 + P_1 a_1)(2\nu_2(\lambda k_{\min}+\sqrt{\delta^2+Q_1^2})+a_1)^2\sqrt{\delta^2+Q_1^2}}{4\lambda^2 P_1 \nu_2^2 C_h Q_1 (2\nu_2(\lambda k_{\min}+D)+a_1)}, \\ C_h &:= \frac{\sqrt{P_1(2\sqrt{\delta^2+Q_1^2}\nu_2(\lambda+1)+a_1)}}{\sqrt{2\nu_2(\lambda+1)}}. \end{aligned}$$

and for $\lambda < 1$,

$$p_2^* \sim P_2 - E_2 h^{\frac{1}{2}},\ q_2^* \sim Q_2 h^{-1} + F_2 h^{-\frac{1}{2}},\ \max_{k_{\min}\leq k\leq \frac{\pi}{h}} |\rho(k, p_2^*, q_2^*)| \sim \lambda - \tfrac{F_2\lambda\pi(1+\lambda)}{(\lambda\pi+Q_2)^2} h^{\frac{1}{2}}.$$

with

$$
\begin{aligned}
P_2 &:= \frac{D + k_{\min} + \frac{a_1}{2\nu_2}}{1-\lambda}, \quad Q_2 := \frac{\pi(\lambda-1)}{2}, \\
E_2 &:= \frac{((\lambda+1)(D_h^2 + P_2Q_2)\nu_2 + \frac{a_1Q_2}{2})(2\nu_2(\lambda P_2 + D) + a_1)^2}{2\nu_2^2 D_h Q_2(2k_{\min}\lambda\nu_2 + 2\nu_2 D + a_1)}, \\
F_2 &:= \frac{\sqrt{\lambda+1}\sqrt{(D+k_{\min})(\lambda+1) + \frac{a_1}{2\nu_2}}\sqrt{\pi}(3\lambda-1)^2}{\sqrt{2}(1-\lambda^2)}, \\
D_h &:= \frac{\sqrt{Q_2(2P_2\nu_2(\lambda+1)+a_1)}}{\sqrt{2\nu_2(\lambda+1)}}.
\end{aligned}
$$

Proof. The proof follows the same steps as in the proof of Theorem 2.1.10. □

2.2.3 Advection tangential to the interface

In the previous section we restricted our study to the case of advection normal to the interface. Here we consider the other relevant physical case, namely advection tangential to the interface, so that $a_1 = 0$ and $a_2 \neq 0$ in (2.2.1). For homogeneous problems, this case has been studied through Fourier transform in unbouded domains, see for instance [64]. However, in [95] we have shown that for homogeneous problems with tangential advection this procedure does not yield efficient optimized parameters. The reason behind this failure lies in the separation of variables technique which applied to the error equation,

$$
\begin{aligned}
(\eta_1^2 - \nu_1\Delta)e_1^n &= 0, \quad \text{in} \quad \Omega_1, \\
(\nu_1\partial_x + S_1)(e_1^n)(0,\cdot) &= (\nu_2\partial_x + S_1)(e_2^{n-1})(0,\cdot), \\
(\eta_2^2 + a_2\partial_y - \nu_2\Delta)e_2^n &= 0, \quad \text{in} \quad \Omega_2, \\
(\nu_2\partial_x - S_2)(e_2^n)(0,\cdot) &= (\nu_1\partial_x - S_2)(e_1^{n-1})(0,\cdot),
\end{aligned}
\tag{2.2.4}
$$

leads to

$$
e_1^n = \sum_{k\in\mathcal{V}} \hat{e}_1^n(0,k)e^{\lambda_1(k)x}\sin(ky) \quad \text{and} \quad e_2^n = \sum_{k\in\mathcal{V}} \hat{e}_2^n(0,k)e^{-\lambda_2(k)x}e^{\frac{a_2 y}{2\nu_2}}\sin(ky), \tag{2.2.5}
$$

where $\lambda_1(k) = \sqrt{k^2+\tilde{\eta_1}}$, $\lambda_2(k) = \frac{\sqrt{4\nu_2^2k^2+4\nu_2^2\tilde{\eta_2}^2+a_2^2}}{2\nu_2}$ with $\tilde{\eta}_j^2 := \frac{\eta_j^2}{\nu_j}$. Since the functions $\psi_k(y) := \sin(ky)$ and $\phi_k(y) := e^{\frac{a_2 y}{2\nu_2}}\sin(ky)$ are not orthogonal, it is not possible to get a recurrence relation which expresses $\hat{e}_j^n(0,k)$ only as a function of $\hat{e}_j^{n-2}(0,k)$ for each k and $j = 1,2$. Nevertheless, we propose a more general approach. First let us define two scalar products, the classical L^2 scalar product and the weighted scalar product

$$
\langle f,g\rangle = \frac{2}{L}\int_\Gamma fg\,dy, \quad \langle f,g\rangle_w = \frac{2}{L}\int_\Gamma fge^{-\frac{a_2 y}{\nu_2}}\,dy.
$$

It follows that $\langle\psi_k,\psi_j\rangle = \delta_{k,j}$ and $\langle\phi_k,\phi_j\rangle_w = \delta_{k,j}$. Setting $S_1 := \nu_2\lambda_2(p)I$ and $S_2 := \nu_1\lambda_1(q)I$ for $p,q\in\mathbb{R}$ and inserting the expansions (2.2.5) into the transmission conditions of (2.2.4)

we obtain

$$\begin{aligned}\sum_{i=1}^{+\infty} \hat{e}_1^n(0,i)(\nu_1\lambda_1(i)+\nu_2\lambda_2(p))\psi_i(y) &= \sum_{l=1}^{+\infty} \hat{e}_2^{n-1}(0,l)(-\nu_2\lambda_2(l)+\nu_2\lambda_2(p))\phi_l(y),\\ \sum_{l=1}^{+\infty} \hat{e}_2^n(0,l)(-\nu_2\lambda_2(l)-\nu_1\lambda_1(q))\phi_l(y) &= \sum_{i=1}^{+\infty} \hat{e}_1^{n-1}(0,i)(\nu_1\lambda_1(i)-\nu_1\lambda_1(q))\psi_i(y).\end{aligned} \tag{2.2.6}$$

We truncate the expansions for $i,l > N$, since higher frequencies are not represented by the numerical grid, and we project the first equation onto ψ_k with respect the scalar product $\langle\cdot,\cdot\rangle$ and the second one onto ϕ_j with respect to the weighted scalar product $\langle\cdot,\cdot\rangle_{\mathrm{w}}$,

$$\begin{aligned}\hat{e}_1^n(0,k)(\nu_1\lambda_1(k)+\nu_2\lambda_2(p)) &= \sum_{l=1}^{N} \hat{e}_2^{n-1}(0,l)(-\nu_2\lambda_2(l)+\nu_2\lambda_2(p))\langle\psi_k,\phi_l\rangle,\\ \hat{e}_2^n(0,j)(-\nu_2\lambda_2(j)-\nu_1\lambda_1(q)) &= \sum_{i=1}^{N} \hat{e}_1^{n-1}(0,i)(\nu_1\lambda_1(i)-\nu_1\lambda_1(q))\langle\phi_j,\psi_i\rangle_{\mathrm{w}}.\end{aligned} \tag{2.2.7}$$

Defining now the vectors $\mathbf{e}_j^n \in \mathbb{R}^N$ such that $(\mathbf{e}_j^n)_i := \hat{e}_j^n(0,i)$ for $j = 1,2$, the matrices $V_{k,l} := \langle\psi_k,\phi_l\rangle$, $W_{j,i} := \langle\phi_j,\psi_i\rangle_{\mathrm{w}}$ and the diagonal matrices $(D_1)_{l,l} := (-\nu_2\lambda_2(l)+\nu_2\lambda_2(p))$, $(\tilde{D}_1)_{k,k} := (\nu_1\lambda_1(k)+\nu_2\lambda_2(p))$, $(D_2)_{i,i} := (\nu_1\lambda_1(i)-\nu_1\lambda_1(q))$, $(\tilde{D}_2)_{j,j} := (-\nu_2\lambda_2(j)-\nu_1\lambda_1(q))$, we obtain,

$$\begin{aligned}\mathbf{e}_1^n &= \tilde{D}_1^{-1}VD_1\mathbf{e}_2^{n-1},\\ \mathbf{e}_2^n &= \tilde{D}_2^{-1}WD_2\mathbf{e}_1^{n-1},\end{aligned} \tag{2.2.8}$$

which implies

$$\mathbf{e}_1^n = \tilde{D}_1^{-1}VD_1\tilde{D}_2^{-1}WD_2\mathbf{e}_1^{n-2} \text{ and } \mathbf{e}_2^n = \tilde{D}_2^{-1}WD_2\tilde{D}_1^{-1}VD_1\mathbf{e}_2^{n-2}. \tag{2.2.9}$$

Since for two given matrices A, B the spectral radius satisfies $\rho(AB) = \rho(BA)$, we conclude that $\rho(\tilde{D}_1^{-1}VD_1\tilde{D}_2^{-1}WD_2) = \rho(\tilde{D}_2^{-1}WD_2\tilde{D}_1^{-1}VD_1)$ and therefore, in order to accelerate the method, we are interested in the minimization problem

$$\min_{p,q\in\mathbb{R}} \rho((\tilde{D}_1^{-1}VD_1\tilde{D}_2^{-1}WD_2)(p,q)). \tag{2.2.10}$$

Remark 2.2.5. Problem (2.2.10) does not have a closed formula solution. However in the next subsection we show its efficiency by solving numerically the minimization problem. For these cases where the theoretical analysis falls short without providing any good estimates, in subsection 2.5 we discuss a cheap and reliable numerical procedure to find optimized transmission conditions.

Remark 2.2.6. Equation (2.2.10) is a straight generalization of the min-max problem (2.1.8). Indeed, assuming that the functions ψ_k and ϕ_j are orthogonal, the matrices V and W are the identity matrix. Therefore equation (2.2.9) simplifies to $\mathbf{e}_1^n = \bar{D}\mathbf{e}_1^{n-2}$ and $\mathbf{e}_2^n = \bar{D}\mathbf{e}_2^{n-2}$, where the diagonal matrix $\bar{D}$ satisfies $(\bar{D})_{k,k} = \frac{\nu_2\lambda_2(k)-\nu_2\lambda_2(p)}{\nu_1\lambda_1(k)+\nu_2\lambda_2(p)}\frac{\nu_1\lambda_1(k)-\nu_1\lambda_1(q)}{\nu_2\lambda_2(k)+\nu_1\lambda_1(q)}$. Since the eigenvalues of a diagonal matrix are its diagonal entries we get that if $W = V = I$,

$$\min_{p,q\in\mathbb{R}} \rho((\tilde{D}_1^{-1}VD_1\tilde{D}_2^{-1}WD_2)(p,q)) = \min_{p,q}\max_k \left|\frac{\nu_2\lambda_2(k)-\nu_2\lambda_2(p)}{\nu_1\lambda_1(k)+\nu_2\lambda_2(p)}\frac{\nu_1\lambda_1(k)-\nu_1\lambda_1(q)}{\nu_2\lambda_2(k)+\nu_1\lambda_1(q)}\right|.$$

Remark 2.2.7. The case of an arbitrary advection, i.e. $a_1 \neq 0$ and $a_2 \neq 0$ has been recently treated in [95] for homogeneous problems. Considering a heterogeneous problem with advection fields $\mathbf{a}_j = (a_{1j}, a_{2j})^\top$ in domain Ω_j, $j = 1,2$, a separation of variables approach would lead to non orthogonal functions $\psi_k(y) = e^{\frac{a_{21}y}{2\nu_1}}\sin(ky)$ and $\phi_k(y) = e^{\frac{a_{22}y}{2\nu_2}}\sin(ky)$ unless $\frac{a_{21}}{2\nu_1} = \frac{a_{22}}{2\nu_2}$, and thus it is not possible to obtain a recurrence relation as shown in (2.2.5). However the approach developed in this section can be readily applied. The subdomain solutions are

$$e_1^n(x,y) = \sum_{k\in\mathcal{V}} \hat{e}_{1,k}^n e^{\frac{a_{21}y}{2\nu_1}}\sin(ky)e^{\lambda_1(k)x}, \quad e_2^n(x,y) = \sum_{k\in\mathcal{V}} \hat{e}_{2,k}^n e^{\frac{a_{22}y}{2\nu_2}}\sin(ky)e^{-\lambda_2(k)x},$$

with $\lambda_1(k) = \frac{a_{11}+\sqrt{4\nu_1^2k^2+4\nu_1^2\tilde{\eta}_1{}^2+a_{11}^2+a_{21}^2}}{2\nu_1}$ and $\lambda_2(k) = \frac{-a_{12}+\sqrt{4\nu_2^2k^2+4\nu_2^2\tilde{\eta}_2{}^2+a_{12}^2+a_{22}^2}}{2\nu_2}$. Defining $S_1 = \nu_2\lambda_2(p) + a_{12}$, $S_2 = \nu_1\lambda_1(p) - a_{11}$, the two scalar products $\langle f,g\rangle_{w_1} = \frac{2}{L}\int_\Gamma fge^{-\frac{a_{21}y}{\nu_1}}dy$ and $\langle f,g\rangle_{w_2} = \frac{2}{L}\int_\Gamma fge^{-\frac{a_{22}y}{\nu_2}}dy$ and repeating the calculations (2.2.6)-(2.2.8), one finds the recurrence relation (2.2.9), with $V_{k,l} := \langle\psi_k,\phi_l\rangle_{w_1}$, $W_{j,i} := \langle\phi_j,\psi_i\rangle_{w_2}$ and the diagonal matrices $(D_1)_{l,l} := (-\nu_2\lambda_2(l) + \nu_2\lambda_2(p))$, $(\tilde{D}_1)_{k,k} := (\nu_1\lambda_1(k) + \nu_2\lambda_2(p) - a_{11} + a_{12})$, $(D_2)_{i,i} := (\nu_1\lambda_1(i) - \nu_1\lambda_1(q))$, $(\tilde{D}_2)_{j,j} := (-\nu_2\lambda_2(j) - \nu_1\lambda_1(q) - a_{12} + a_{11})$.

2.3 Numerical results

The numerical experiments are performed using the subdomains $\Omega_1 = (-1,0)\times(0,1)$, $\Omega_2 = (0,1)\times(0,1)$. We use a classical five point finite difference scheme for the interior points and treat the normal derivatives with second order discretization using a ghost point formulation.

2.3.1 Reaction Diffusion-Diffusion coupling

We first consider the reaction diffusion-diffusion coupling analyzed in Section 2.1. Tables 2.1 and 2.2 show the values of the convergence factor in two different asymptotic regimes, when $h \to 0$, and for strong heterogeneity. As the asymptotic Theorem 2.1.10 and Remark 2.1.5 state, a strong heterogeneity improves the performance of the algorithm. In the single sided optimized case, the value of the convergence factor $|\rho(k)|$ tends to 1, while in the double sided case, $|\rho(k)|$ is bounded either by λ or by $1/\lambda$. Fig. 2.3 shows the number of iterations required to reach convergence with a tolerance of 10^{-6} as function of the optimized parameters in both the single and double sided cases. We see that the analysis predicts the optimized parameter very well.

2.3.2 Advection Reaction Diffusion-Diffusion coupling

Next we consider the advection reaction diffusion-diffusion coupling with advection normal to the interface. Table 2.3 summarizes the behaviour of $\rho(k)$ as $h \to 0$ and for strong heterogeneity. Similarly Fig 2.4 shows the number of iterations required to reach convergence with the tolerance of 10^{-6}. Figure 2.5 shows the number of iterations to reach convergence for the tangential advection case. The minimization problem (2.2.10) is solved

h	ρ single sided	ρ double sided
1/50	0.7035	0.4052
1/100	0.7801	0.4748
1/500	0.8950	0.6160
1/1000	0.9245	0.6672
1/5000	0.9655	0.7650

h	ρ single sided	ρ double sided
1/50	0.1721	0.0337
1/100	0.2625	0.0456
1/500	0.4868	0.0685
1/1000	0.5823	0.0760
1/5000	0.7662	0.0872

Table 2.1: Asymptotic behaviour as $h \to 0$ for the reaction diffusion-diffusion coupling. Physical parameters: left table $\tilde{\eta}^2 = \lambda = 1$, right table $\tilde{\eta}^2 = \lambda = 10$.

λ	ρ single sided	ρ double sided
0.001	0.0125	$7.8 \cdot 10^{-4}$
0.01	0.1075	0.0078
0.1	0.4453	0.0757
1	0.5851	0.4748
10	0.2625	0.076
100	0.0389	0.0078
1000	0.0040	$7.8 \cdot 10^{-4}$

Table 2.2: Asymptotic behaviour as $\lambda \to 0$ and $\lambda \to \infty$, with $h = 0.05$ for the reaction diffusion-diffusion coupling. Physical parameter: $\tilde{\eta}^2 = 1$.

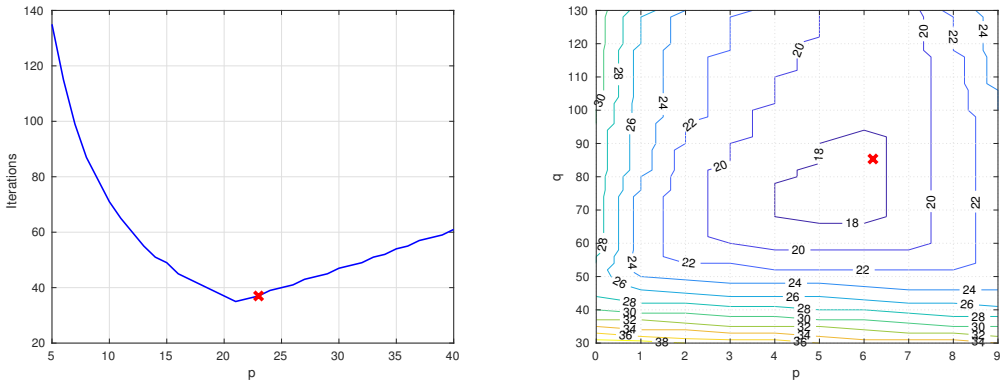


Figure 2.3: Number of iterations required to reach convergence with a tolerance of 10^{-6} as function of the optimized parameters for the reaction diffusion-diffusion coupling. The left panel shows the single sided case while the right panel shows the double sided case. Physical parameters : $\nu_1 = 2$, $\nu_2 = 1$, $\eta^2 = 10$, mesh size $h = 0.02$.

h	ρ single sided	ρ double sided
1/50	0.4766	0.1835
1/100	0.5910	0.2306
1/500	0.7889	0.3274
1/1000	0.8452	0.3618
1/5000	0.9273	0.4228

λ	ρ single sided	ρ double sided
0.001	0.0031	$4.89 \cdot 10^{-4}$
0.01	0.0297	0.0049
0.1	0.2101	0.0458
1	0.4865	0.2552
10	0.2786	0.0517
100	0.0459	0.0056
1000	0.0049	$5.6 \cdot 10^{-4}$

Table 2.3: For the advection reaction diffusion-diffusion coupling, the left table shows the asymptotic behaviour when $h \to 0$ while the right table shows the values of the convergence factor for strong heterogeneity when $h = 1/50$. Physical parameters: $\eta_1^2 = 1, \eta_2^2 = 2, \nu_1 = 2, \nu_2 = 1, a_2 = 0,\ a_1 = 5$, mesh size $h = 0.02$.

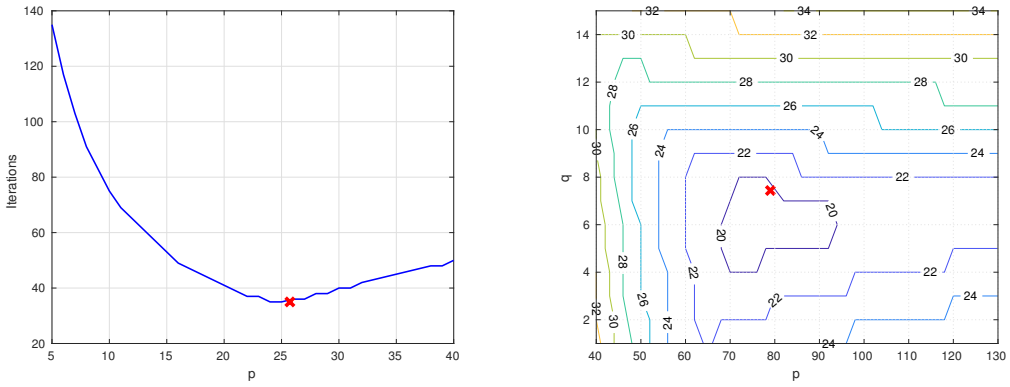


Figure 2.4: Number of iterations required to reach convergence with a tolerance of 10^{-6} as function of the optimized parameters for the advection reaction diffusion-diffusion coupling with normal advection. Physical parameters: $\nu_1 = 2$, $\nu_2 = 1$, $\eta_1^2 = 1$, $\eta_2^2 = 2$, $a_1 = 5$, mesh size $h = 0.02$.

numerically to find the optimal parameters p and q using the Nelder-Mead algorithm. We have solved the minimization problem with different initial couples (p, q) and we have noticed that the optimal solution satisfies an ordering relation between p and q depending on λ as in Theorem 2.1.10 and 2.2.4.

2.3.3 Application to the contaminant transport problem

Contaminant transport in underground media is a topic of great interest in the last thirty years due, for instance, to the increasing threat of contamination of groundwater supplies by waste treatments and landfill sites or to the disposal of nuclear radioactive waste [10]. We refer to [5] for a reference regarding modeling issues of contaminant transport. Our model assumes that the computational domain $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4$, represented in Fig-

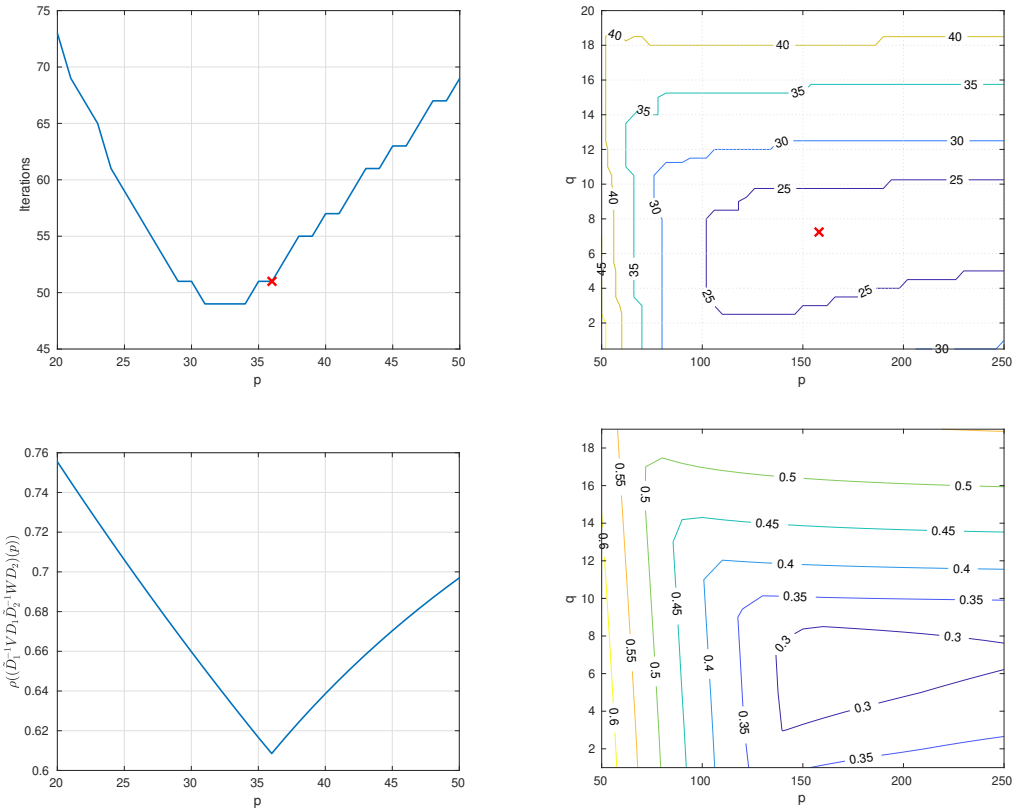


Figure 2.5: In the top row, we show the number of iterations required to reach convergence with a tolerance of 10^{-6} as function of the optimized parameters for the advection reaction diffusion-diffusion coupling with tangential advection. In the bottom row, we show the dependence on p and the level curves of the objective function in the min-max problem (2.2.10). Physical parameters: $\nu_1 = 1$, $\nu_2 = 2$, $\eta_1^2 = 1$, $\eta_2^2 = 2, a_2 = 15$, mesh size $h = 0.01$.

ure 2.6, can be partitioned into four layers. In the first one, the contaminant, whose concentration is described through the unknown u, penetrates mainly thanks to rainfalls and therefore an advection towards the negative y direction is present. The next two layers are formed by porous media so that the contaminant spreads in a diffusive regime described by the Laplace equation. We furthermore suppose that in the second layer, some chemical reactions may take place which are synthesized in the reaction term. Finally in the last layer, an underground flow transports the contaminant in the x direction towards a groundwater supply which is connected to a water well. The problem belongs to the heterogeneous class, since in different parts of the domain we have different physical phenomena, and thus in the last paragraph we use the results discussed in this manuscript to design an efficient domain decomposition method to compute the stationary and time dependent distribution of the contaminant.

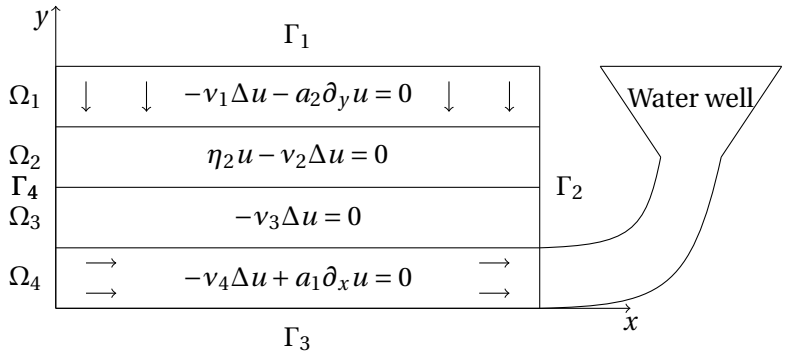


Figure 2.6: Geometry for the contaminant transport problem.

The computational domain Ω is set equal to $\Omega = (0,8) \times (-4,0)$, with $\Omega_j = (0,8) \times (1-j,-j)$, $j = 1...4$. On the top boundary Γ_1, we impose a condition on the incoming contaminant flow, i.e. $\frac{\partial u}{\partial y} - a_2 u = 1$ while on the bottom edge Γ_3 we impose a zero Neumann boundary condition $\frac{\partial u}{\partial y} = 0$. On the vertical edges Γ_2 and Γ_4 we set absorbing boundary conditions so that

$$\begin{aligned} \tfrac{\partial u}{\partial \mathbf{n}} + pu = 0 \quad &\text{on} \quad \{0\} \times [-3;0] \text{ and } \{8\} \times [-3;0], \\ \tfrac{\partial u}{\partial \mathbf{n}} - a_1 u + pu = 0 \quad &\text{on} \quad \{0\} \times [-4;-3] \text{ and } \{8\} \times [-4;-3], \end{aligned}$$

where $\mathbf{n}$ is the outgoing normal vector. The parameter p is chosen equal to $p = \sqrt{\pi \frac{\pi}{h}}$, being $k_{\min} = \pi$ and $k_{\max} = \frac{\pi}{h}$. This choice derives from the observation that imposing $\frac{\partial u}{\partial n} + \mathscr{S} u = 0$, where $\mathscr{S}$ is the Steklov-Poincaré, is an exact transparent boundary condition, see [136, 135]. Thus we replace the expensive exact transparent boundary condition with an approximation of the Steklov-Poincaré operator. We know from [74] that $p = \sqrt{\pi \frac{\pi}{h}}$ is indeed a zero order approximation of $\mathscr{S}$. To solve the system of PDEs, we consider the OSM:

$$\begin{aligned}
-\nu_1 \Delta u_1^n - a_2 \partial_y u_1^n &= 0 && \text{in } \Omega_1, && \mathscr{B}_1(u_1^n) = 0 \text{ on } \partial\Omega_1 \setminus \widetilde{\Gamma}_1, \\
\partial_{n_{1,2}} u_1^n + p_{12} u_1^n &= \partial_{n_{1,2}} u_2^{n-1} + p_{12} u_2^{n-1} && \text{on } \widetilde{\Gamma}_1, && \\
\eta_2^2 u_2^n - \nu_2 \Delta u_2^n &= 0 && \text{in } \Omega_2, && \mathscr{B}_2(u_2^n) = 0 \text{ on } \partial\Omega_2 \setminus \{\widetilde{\Gamma}_1, \widetilde{\Gamma}_2\}, \\
\partial_{n_{1,1}} u_2^n + p_{21} u_2^n &= \partial_{n_{1,1}} u_1^{n-1} + p_{21} u_1^{n-1} && \text{on } \widetilde{\Gamma}_1, && \\
\partial_{n_{2,3}} u_2^n + p_{23} u_2^n &= \partial_{n_{2,3}} u_3^{n-1} + p_{23} u_3^{n-1} && \text{on } \widetilde{\Gamma}_2, && \\
-\nu_3 \Delta u_3^n &= 0 && \text{in } \Omega_3, && \mathscr{B}_3(u_3^n) = 0 \text{ on } \partial\Omega_3 \setminus \{\widetilde{\Gamma}_2, \widetilde{\Gamma}_3\}, \\
\partial_{n_{2,2}} u_3^n + p_{32} u_3^n &= \partial_{n_{2,2}} u_2^{n-1} + p_{32} u_2^{n-1} && \text{on } \widetilde{\Gamma}_2, && \\
\partial_{n_{3,4}} u_3^n + p_{34} u_3^n &= \partial_{n_{3,4}} u_4^{n-1} + p_{34} u_4^{n-1} && \text{on } \widetilde{\Gamma}_3, && \\
-\nu_4 \Delta u_4^n + a_1 \partial_x u_4^n &= 0 && \text{in } \Omega_4, && \mathscr{B}_4(u_4^n) = 0 \text{ on } \partial\Omega_4 \setminus \widetilde{\Gamma}_3, \\
\partial_{n_{3,3}} u_4^n + p_{43} u_4^n &= \partial_{n_{3,3}} u_3^{n-1} + p_{43} u_3^{n-1} && \text{on } \widetilde{\Gamma}_3, &&
\end{aligned} \tag{2.3.1}$$

where $\widetilde{\Gamma}_i$ are the shared interfaces $\widetilde{\Gamma}_i = \partial\Omega_i \cap \partial\Omega_{i+1}$, $i = 1,2,3$, the vectors $\mathbf{n}_{i,j}$ are the normal vectors on the interface $\widetilde{\Gamma}_i$ pointing towards the interior of the domain Ω_j and the operators $\mathscr{B}_i(u_i)$ represent the boundary conditions to impose on the boundary excluding

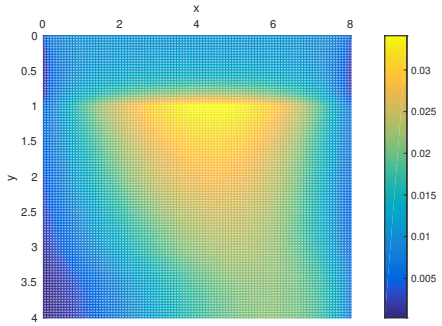


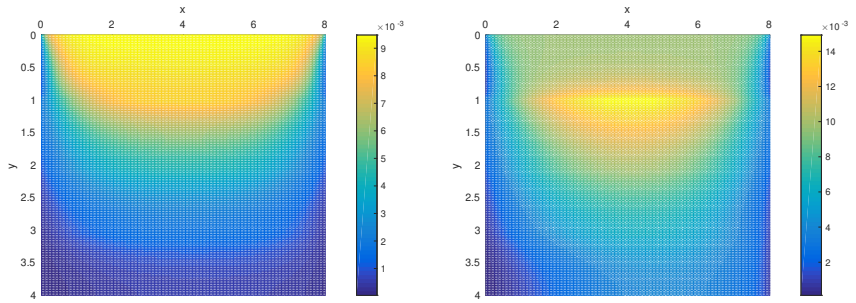
Figure 2.7
Stationary distribution of the contaminant. Physical parameters:
$\nu_1 = 0.5, \nu_2 = 3, \nu_3 = 3, \nu_4 = 1, \eta_2^2 = 0.01, a_2 = 2, a_1 = 2.$

the shared interfaces. Regarding the Robin parameters $p_{i,j}$, we choose them according to the two subdomain analysis carried out in this Chapter. Due to the exponential decay of the error away from the interface, see eq. (2.1.3), if the subdomains are not too narrow in the y direction, the information transmitted from each subdomain to the neighbouring one does not change significantly and therefore the $p_{i,j}$ from a two subdomain analysis are still a good choice. We remark that this argument does not hold for the Helmholtz equation, for which there are resonant modes for frequencies $k \leq \omega$, where ω is the wave number, which travel along the domains and they do not decay away from the interface. Figure 2.7 shows the stationary distribution of the contaminant. We observe that due to the advection in the y direction in Ω_1, the contaminant accumulates on the interface with Ω_2, representing the porous medium, and here we have the highest concentration. Then the contaminant diffuses into the layers below and already in the porous media region it feels the presence of the tangential advection in Ω_4. Next we also consider the transient version of equations (2.3.1). We discretize the time derivative with an implicit Euler scheme, so that each equation has a further reaction term equal to $\eta_{j,tran}^2 = \eta_{j,stat}^2 + \frac{1}{\Delta t}$. Figure 2.8 shows the time dependent evolution of the concentration u over 400 integration steps. The initial condition is set equal to zero on the whole domain Ω.

Table 2.4 shows the number of iterations to reach a tolerance of 10^{-6} for the algorithm (2.3.1) both used as iterative method and as a preconditioner for GMRES for the substructured system, see [80] for an introduction to the substructured version of (2.3.1). We consider both single and double sided optimizations for the parameters $p_{i,j}$ at each interface. For the time evolution problem, the stopping criterion is

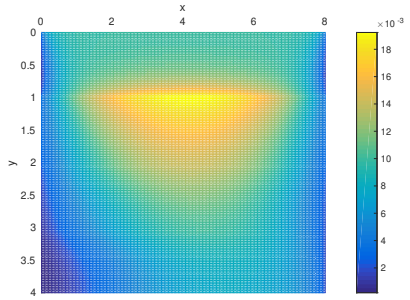
$$\max\left\{ \frac{\| u_{1,\tilde{\Gamma}_1}^{n,k} - u_{2,\tilde{\Gamma}_1}^{n,k} \|}{\| u_{1,\tilde{\Gamma}_1}^{n,k} \|}, \frac{\| u_{2,\tilde{\Gamma}_2}^{n,k} - u_{3,\tilde{\Gamma}_2}^{n,k} \|}{\| u_{2,\tilde{\Gamma}_2}^{n,k} \|}, \frac{\| u_{3,\tilde{\Gamma}_1}^{n,k} - u_{4,\tilde{\Gamma}_3}^{n,k} \|}{\| u_{3,\tilde{\Gamma}_3}^{n,k} \|} \right\} \leq 10^{-6}. \tag{2.3.2}$$

From Figures 2.7 and 2.8, we note that this physical configuration would represent a safe situation since a very small concentration of contaminant manages to get through the ver-

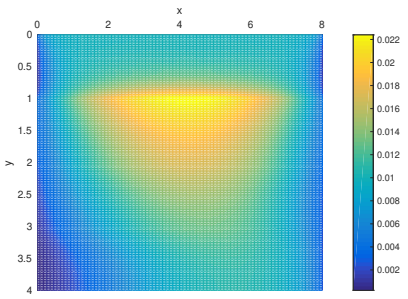


(a) 100 time steps.

(b) 200 time steps.



(c) 300 time steps.



(d) 400 time steps.

Figure 2.8: Evolution of the contaminant concentration u.

	Iterative	GMRES
Single sided	270	33
Double sided	55	25

	Iterative	GMRES
Single sided	11.5	5.7
Double sided	9.6	4.3

Table 2.4
Number of iterations to reach a tolerance of 10^{-6} for the OSM (2.3.1) used as an iterative method and as a preconditioner. The left side refers to the stationary case while the right side to the transient one where we consider the number of iterations needed to satisfy the stopping criterion (2.3.2) averaged over 400 time steps.

tical diffusive layers and to reach the right-bottom of the domain, where it could pollute the water well.

2.4 Coupling Helmholtz and Laplace Equations

In this Section, we introduce and analyze heterogeneous OSMs with zeroth order optimized transmission conditions for the coupling between the hard to solve Helmholtz equation [71] and the Laplace equation. It is a simplified instance of the coupling of parabolic and hyperbolic operators, which might arise in Maxwell equations. The Helmholtz

equation is used in the time harmonic regime of a wave equation and the Laplace operator represents the parabolic part.

2.4.1 Well-posedness analysis

We consider the nonoverlapping decomposition described at the beginning of Section 1.3. Our model problem is

$$\begin{aligned} (-\Delta - q\omega^2)u &= f && \text{in } \Omega, \\ \frac{\partial u}{\partial n} + i\omega u &= 0 && \text{on } \Gamma_1 := \partial\Omega_1 \setminus \Gamma, \\ u &= 0 && \text{on } \Gamma_2 := \partial\Omega_2 \setminus \Gamma, \end{aligned} \tag{2.4.1}$$

where $\omega > 0$ is the Helmholtz frequency, and $q \in L^\infty(\Omega)$ satisfies $q = 1$ in Ω_1 and $q = 0$ in Ω_2. Since the well-posedness of the problem is not straightforward due to the indefinite nature of the Helmholtz part, we first analyze it in more detail adapting arguments presented by Després in his PhD thesis [50].

Lemma 2.4.1. *The norm* $||u||^2 = \int_\Omega |\nabla u|^2 + \omega \int_{\Gamma_1} |u|^2$ *is equivalent to the canonical norm on* $H^1(\Omega)$ *if* $|\Gamma_1| > 0$.

Proof. We first observe that $H^1(\Omega)$ is the direct sum of $\overline{V} = \left\{v \in H^1(\Omega) : \int_\Omega v = 0\right\}$ and $\widetilde{V} = \left\{v \in H^1(\Omega) : v \text{ is constant in } \Omega\right\}$, $H^1(\Omega) = \widetilde{V} \oplus \overline{V}$. Then, on the one hand, it easy to see that for all $v \in \widetilde{V}$, there exist a constant $\widetilde{C} = \sqrt{\frac{\omega|\Gamma_1|}{|\Omega|}}$ such that

$$\widetilde{C}||v||_{H^1(\Omega)} \leq ||v|| \leq \widetilde{C}||v||_{H^1(\Omega)}. \tag{2.4.2}$$

On the other hand, for every $v \in \overline{V}$, we first use the Poincaré inequality with constant C to get

$$||v||^2_{H^1(\Omega)} \leq (1+C)\int_\Omega |\nabla v|^2 \leq (1+C)\left(\int_\Omega |\nabla v|^2 + \omega\int_{\Gamma_1} |v|^2\right) = (1+C)||v||^2. \tag{2.4.3}$$

Using the continuity of the trace operator, we obtain

$$||v||^2 = \int_\Omega |\nabla v|^2 + \omega\int_{\Gamma_1} |v|^2 \leq \int_\Omega |\nabla v|^2 + \omega\int_{\partial\Omega} |v|^2 \leq \max(1, C_{\partial\Omega}\omega)\left(\int_\Omega |\nabla v|^2 + \int_\Omega |v|^2\right). \tag{2.4.4}$$

Having proved that the two norms are equivalent on the subspaces $\overline{V}$ and $\widetilde{V}$ with $\widetilde{V} \oplus \overline{V} = H^1(\Omega)$, the two norms are also equivalent on $H^1(\Omega)$. □

Let us define $V := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_2\}$, with $||\cdot||_V = ||\cdot||_{H^1(\Omega)}$, and consider problem (2.4.1) in the variational form

$$\text{Find } u \in V : \; a(u,v) - b(u,v) =_{V^{-1}} \langle f, v\rangle_V \quad \forall v \in V, \tag{2.4.5}$$

where $a(u,v) = \int_\Omega \nabla u \nabla \bar{v} + i\omega\int_{\Gamma_1} u\bar{v}$, $b(u,v) = \omega^2\int_{\Omega_2} u\bar{v}$ and $f \in V^{-1}$. To use Fredholm theory, we now show that the bilinear form b is a compact pertubation of a.

Lemma 2.4.2. *Let $\mathscr{B}$ be an operator from V to V such that*

$$a(\mathscr{B}u, v) = b(u, v) \quad \forall v \in V, \tag{2.4.6}$$

then $\mathscr{B}$ is a continuous compact operator.

Proof. We first prove continuity, i.e. $\exists C > 0 : \forall u \in V, ||\mathscr{B}u||_V \leq C||u||_V$. From the definition of $\mathscr{B}$, and applying Lax-Milgram to (2.4.6), we have $||\mathscr{B}u||_V \leq \frac{1}{\alpha}||b(u)||_{V^{-1}}$, where $b(u) : V \to \mathbb{C}$ is the functional defined by ${}_{V^{-1}}\langle b(u), v\rangle_V := b(u, v)$. Then we have $\forall v \in V$

$$|{}_{V^{-1}}\langle b(u), v\rangle_V| := |b(u, v)| = \omega^2 \left| \int_{\Omega_2} u\bar{v} \right| \leq \omega^2 ||u||_{L^2(\Omega_2)} ||v||_{L^2(\Omega_2)} \leq \omega^2 ||u||_{L^2(\Omega_2)} ||v||_V.$$

We thus conclude that $||b(u)||_{V^{-1}} \leq \omega^2 ||u||_{L^2(\Omega_2)}$, and hence we have the bound

$$||\mathscr{B}u||_V \leq \frac{1}{\alpha}\omega^2 ||u||_V.$$

To prove compactness, let u_n be a bounded sequence in V, i.e $\exists C > 0 : \forall n, ||u_n||_V < C$. From weak compactness of V it follows that there exists a subsequence u_{n_j} such that $u_{n_j} \rightharpoonup u$ for some u. Hence u_{n_j} converge strongly to u in $L^2(\Omega)$. Considering $a(\mathscr{B}u_{n_j} - \mathscr{B}u, \mathscr{B}u_{n_j} - \mathscr{B}u) = b(u_{n_j} - u, \mathscr{B}u_{n_j} - \mathscr{B}u)$ we have letting $n \to \infty$ and using the Cauchy-Schwarz inequality

$$\left| \int_\Omega |\nabla(\mathscr{B}u_{n_j} - \mathscr{B}u)|^2 + i\omega \int_{\Gamma_1} |\mathscr{B}u_{n_j} - \mathscr{B}u|^2 \right| \leq \omega^2 ||u_{n_j} - u||_{L^2(\Omega_2)} ||\mathscr{B}u_{n_j} - \mathscr{B}u||_{L^2(\Omega_2)}. \tag{2.4.7}$$

We observe that $\mathscr{B}u_{n_j} \rightharpoonup \mathscr{B}u$ in V because $u_{n_j} \rightharpoonup u$ in V and $\mathscr{B}$ is a continuous operator [44]. Hence, both u_{n_j} and $\mathscr{B}u_{n_j}$ converge strongly in $L^2(\Omega)$. In particular we have that $a(\mathscr{B}u_{n_j} - \mathscr{B}u, \mathscr{B}u_{n_j} - \mathscr{B}u) \to 0$ which implies $||\mathscr{B}u_{n_j} - \mathscr{B}u|| \to 0$. With Lemma 2.4.1, we have that $\mathscr{B}u_{n_j} \to \mathscr{B}u$ in V and thus $\mathscr{B}$ is a compact operator. □

Since $\mathscr{B}$ is a compact operator, due to the Fredholm alternative, existence of the solution of problem (2.4.5) follows from uniqueness. We need two further Lemmas to prove uniqueness. We denote with $\gamma_j u$ and $\mathscr{N}_j u$ the trace of u and the trace of the normal derivative on the j-th interface and we introduce the space $E(\Omega, \Delta) := \{u \in H^1(\Omega) : -\Delta u \in L^2(\Omega)\}$.

Lemma 2.4.3 (Grisvard, Theorem 1.5.3.11, page 61, [106])**.** *Let Ω be an open bounded subset of $\mathbb{R}^2$ whose boundary is a curvilinear polygon of class $C^{1,1}$ with interfaces Γ_j, $j = 1, .., N$. The mappings $u \to \gamma_j u$ and $u \to \mathscr{N}_j u$ have a unique continuous extension from $E(\Omega, \Delta)$ to respectively $H^{\frac{1}{2}}(\Gamma_j)$ and $H^{-\frac{1}{2}}(\Gamma_j)$. Moreover for every $u \in E(\Omega, \Delta)$ and $v \in H^1(\Omega)$ with $\gamma_j v \in H^{\frac{1}{2}}(\Gamma_j)$ $\forall j$, the Green's formula holds:*

$$(-\Delta u, v) = (\nabla u, \nabla v) - \sum_{j=1}^{N} \langle \mathscr{N}_j u, \overline{\gamma_j v}\rangle. \tag{2.4.8}$$

Lemma 2.4.4 (Després, Corollary 2.1, page 22, [50])**.** *Let Ω be an open bounded arc-connected subset of $\mathbb{R}^2$ and assume that Γ is a nonempty open subset of $\partial\Omega$ of class C^2 and $q \in L^\infty(\Omega)$. If $u \in H^2(\Omega)$ satisfies*

$$(-\Delta - q\omega^2)u = 0 \text{ on } \Omega, \quad u|_\Gamma = \partial_n u|_\Gamma = 0, \tag{2.4.9}$$

then u=0 in Ω.

Theorem 2.4.5. *Under the hypotheses of Lemmas 2.4.3 and 2.4.4, $u \equiv 0$ is the only solution of the boundary value problem (2.4.1) with $f = 0$.*

Proof. Choosing $v \in D(\Omega)$, the space of $C^\infty(\Omega)$ functions with compact support, in the weak formulation of eq. (2.4.1) we obtain $-\Delta u - q\omega^2 u = 0$. Hence, since $u \in V$, $\Delta u \in L^2(\Omega)$ and $u \in E(\Omega, \Delta)$. Using Green's formula and choosing $v = u$ we get

$$\int_\Omega |\nabla u|^2 - \omega^2 \int_{\Omega_1} |u|^2 + i\omega \int_{\Gamma_1} |u|^2 = 0. \tag{2.4.10}$$

Considering the imaginary part we have $\int_{\Gamma_1} |u|^2 = 0$, which implies $u = 0$ on Γ_1. We now have homogeneous Dirichlet data on the whole domain $\partial\Omega = \Gamma_1 \cup \Gamma_2$. Regularity results for Dirichlet problems in smooth domains state that $u \in H^2(\Omega)$. Using again the Green's formula and $-\Delta u - q\omega^2 u = 0$ in Ω , we obtain

$$_{H^{-\frac{1}{2}}(\Gamma_1)}\left\langle \frac{\partial u}{\partial n}, v \right\rangle_{H^{\frac{1}{2}}(\Gamma_1)} + iw \int_{\Gamma_1} uv = 0, \quad \forall v \in V. \tag{2.4.11}$$

Since $u = 0$ on Γ_1, we can conclude that $\partial_n u = 0$ on Γ_1 and by the unique continuation principle in Lemma 2.4.4, the result follows. □

2.4.2 Zeroth order single sided optimized transmission conditions

In order to make analytical calculations, we simplify the analysis and set $\Omega = \mathbb{R}^2$, with Ω_1 being the left half plane and Ω_2 the right half plane. The heterogeneous OSM is

$$\begin{array}{rclcrcl} (-\omega^2 - \Delta)u_1 & = & f & \text{in } \Omega_1, & (\partial_x + S_1)(u_1^n)(0,\cdot) & = & (\partial_x + S_1)(u_2^{n-1})(0,\cdot), \\ -\Delta u_2 & = & f & \text{in } \Omega_2, & (\partial_x + S_2)(u_2^n)(0,\cdot) & = & (\partial_x + S_2)(u_1^{n-1})(0,\cdot), \end{array} \tag{2.4.12}$$

where the S_j, $j = 1,2$ are linear operators along the interface in the y direction. The system is closed by the Sommerfeld radiation condition $\lim_{x\to-\infty} \sqrt{|x|}\frac{x}{|x|}(\partial_x u_1^n - i\omega u_1^n) = 0$ and by the boundedness condition $\lim_{x\to+\infty} u_2^n = 0$. The goal is to find which operators lead to the fastest convergence. We define the errors $e_j := u - u_j$, and taking the Fourier transform of the error equations in the y direction, we obtain

$$\begin{array}{rclcl} (-\omega^2 - \partial_{xx} + k^2)(\hat{e}_1^n) & = & 0 & & k \in \mathbb{R},\ x < 0, \\ (\partial_x + \sigma_1(k))(\hat{e}_1^n)(0,k) & = & (\partial_x + \sigma_1(k))(\hat{e}_2^{n-1})(0,k), & & k \in \mathbb{R}, \\ (-\partial_{xx} + k^2)(\hat{e}_2^n) & = & 0 & & k \in \mathbb{R},\ x > 0, \\ (\partial_x + \sigma_2(k))(\hat{e}_2^n)(0,k) & = & (\partial_x + \sigma_2(k))(\hat{e}_1^{n-1})(0,k), & & k \in \mathbb{R}, \end{array} \tag{2.4.13}$$

where $\sigma_j(k)$ are the Fourier symbols of the operators S_j. Solving the equations in (2.4.13) and imposing the radiation/boundedness conditions, we get

$$\hat{e}_1^n = \hat{e}_1^n(0,k)e^{\lambda(k)x}, \quad \hat{e}_2^n = \hat{e}_2^n(0,k)e^{-|k|x},$$

where $\lambda(k) := i\sqrt{\omega^2 - k^2}$ if $k < \omega$ and $\lambda(k) := \sqrt{k^2 - \omega^2}$ if $k \geq \omega$. Applying the transmission conditions, it follows that

$$\hat{e}_1^n = \rho(k)\hat{e}_1^{n-2}, \qquad \hat{e}_2^n = \rho(k)\hat{e}_2^{n-2},$$

where

$$\rho(k) = \frac{-|k| + \sigma_1(k)}{\lambda(k) + \sigma_1(k)} \frac{\lambda(k) + \sigma_2(k)}{-|k| + \sigma_2(k)}.$$

Next, to approximate the optimal choice for $\sigma_1(k)$ and $\sigma_2(k)$ which would require non local operators, we set $\sigma_1 = -\sigma_2 = p(1+i)$. This choice is motivated by [70] where the single and double sided optimizations were studied and compared for the time harmonic Maxwell equations. Since both σ_j and $\lambda(k)$ contain complex numbers, we have to study the modulus of the convergence factor,

$$|\rho(k,p)|^2 = \begin{cases} \dfrac{((k-p)^2+p^2)}{((k+p)^2+p^2)} \dfrac{((\sqrt{k^2-\omega^2}-p)^2+p^2)}{((\sqrt{k^2-\omega^2}+p)^2+p^2)} & k \geq \omega, \\ \dfrac{((k-p)^2+p^2)}{((k+p)^2+p^2)} \dfrac{((\sqrt{\omega^2-k^2}-p)^2+p^2)}{((\sqrt{\omega^2-k^2}+p)^2+p^2)} & k < \omega. \end{cases} \tag{2.4.14}$$

Since we are interested in minimizing the convergence factor over all relevant numerically represented frequencies, we study now the minimax problem

$$\min_{p \geq 0} \max_{k \in [k_{\min}, k_{\max}]} |\rho(k,p)|^2, \tag{2.4.15}$$

where $k_{\min}$ is the minimum frequency and $k_{\max}$ is the maximum frequency supported by the numerical grid.

Theorem 2.4.6. *Assuming that $k_{\max} > 2\omega$, the solution of the minimax problem (2.4.15) is given by $p^* = \frac{\omega}{\sqrt{2}}$ if $|\rho(k_{\max}, p^* = \frac{\omega}{\sqrt{2}})|^2 \leq \frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1}$, and otherwise it is given by the unique p^* such that $|\rho(k=\omega, p^*)|^2 = |\rho(k_{\max}, p^*)|^2$.*

Proof. We consider $p > 0$, because for $p = 0$ the convergence factor is equal to 1, and for $p < 0$ it is greater than one, while for values of $p > 0$, the convergence factor is always less than 1. We introduce a change of variables which will be useful in the computations, namely $x = \sqrt{k^2 - \omega^2}$ if $k \geq \omega$ and $x = \sqrt{\omega^2 - k^2}$ for $k < \omega$. Problem (2.4.15) then becomes

$$\min_{p>0} \max \left(\max_{[0, \sqrt{\omega^2 - k_{\min}^2}]} G(x,p), \max_{[0, \sqrt{k_{\max}^2 - \omega^2}]} F(x,p) \right), \tag{2.4.16}$$

where

$$G(x,p) = \frac{((x-p)^2+p^2)}{((x+p)^2+p^2)} \frac{((\sqrt{\omega^2-x^2}-p)^2+p^2)}{((\sqrt{\omega^2-x^2}+p)^2+p^2)},$$

$$F(x,p) = \frac{((x-p)^2+p^2)}{((x+p)^2+p^2)} \frac{((\sqrt{x^2+\omega^2}-p)^2+p^2)}{((\sqrt{x^2+\omega^2}+p)^2+p^2)}.$$

First, we observe that $\frac{\partial G}{\partial x}|_{x=0} = \frac{\partial F}{\partial x}|_{x=0} = -\frac{(2((\omega-p)^2+p^2))}{(p((\omega+p)^2+p^2))} < 0$ for all $p > 0$ and $G(0,p) = F(0,p)$. Indeed, $x = 0$ ($k = \omega$) is a cusp for $\rho^2(k,p)$ and hence it is a local maximum which needs to be minimized. The minimum of $G(0,p)$ with respect to the variable p is given by $\bar{p} = \frac{\omega}{\sqrt{2}}$ and $G(x=0, p=\frac{\omega}{\sqrt{2}}) = \frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1} \approx 0.176$. We thus have found a lower bound for the value of the minimax problem. Next, we study how $G(x,p)$ behaves in the rest of the interval, and start by restricting our attention to the case $p \geq \bar{p}$. Computing the partial derivative with respect to x of $G(x,p)$, we find that it has a unique zero x_1 given by the root of the non linear equation

$$x(4p^4+x^4)(2p^2+x^2-\omega^2) = ((\omega^2-x^2)^2+4p^2)(2p^2-x^2)\sqrt{\omega^2-x^2}. \tag{2.4.17}$$

To proof uniqueness, it is enough to notice that the LHS is zero for $x = 0$ and strictly increasing in x, if $p \geq \bar{p}$, while the RHS is greater than zero for $x = 0$ and strictly decreasing in x. Therefore $G(x,p)$ decreases until $x < x_1$ and then increases monotonically. If $x_1 > \sqrt{\omega^2-k_{\min}^2}$ then the $\max_{[0,\sqrt{\omega^2-k_{\min}^2}]} G(x,p) = G(0,p)$, otherwise if $x_1 \leq \sqrt{\omega^2-k_{\min}^2}$ it is sufficient to notice that $G(\sqrt{\omega^2-k_{\min}^2}, p) < G(\omega,p) = G(0,p)$, to conclude that it holds again $\max_{[0,\sqrt{\omega^2-k_{\min}^2}]} G(x,p) = G(0,p)$. Next we focus on the second interval, considering the function $F(x,p)$. The zeros of the derivative $\frac{\partial F}{\partial x}$ are given by the zeros of the equation

$$x(4p^2+x^4)(2^2+x^2-2p^2) = (2p^2-x^2)((\omega^2+x^2)^2+4p^2)\sqrt{\omega^2+x^2}.$$

Repeating an argument similar to the one above, we find that again there is a unique zero x_2, in this case $\forall\, p > 0$, which again might or might not belong to the interval $[0, \sqrt{k_{\max}^2-\omega^2}]$. If x_2 is outside the interval or $F(\sqrt{k_{\max}^2-\omega^2}, \bar{p}) \leq F(0,\bar{p})$, then we can conclude that the optimal value p^* is given by $p^* = \bar{p}$, i.e. the value which minimizes the convergence factor for the frequency $k = \omega$. Otherwise the local maxima are located at $x = 0$ and $x = \sqrt{k_{\max}^2-\omega^2}$. We compute the partial derivative w.r.t the variable p, which satisfies $\frac{\partial F}{\partial p}|_{x=\sqrt{k_{\max}^2-\omega^2}} < 0$ for $p \in I = [0, \sqrt{\frac{k_{\max}^2-\omega^2}{2}}]$, and under the non restrictive hypothesis $k_{\max} > 2\omega$, we have that $\bar{p} \in I$. Analyzing the sign of the derivative shows that it is not useful to look for p^* in $[0, \frac{\omega}{\sqrt{2}}]$, since both local maxima would increase. This justifies why we studied G only for $p \geq \bar{p}$. Since $\frac{\partial F}{\partial p}|_{x=0} > 0$ for $p > \frac{\omega}{\sqrt{2}}$ and because

$$F\left(\sqrt{k_{\max}^2-\omega^2}, \sqrt{\frac{k_{\max}^2-\omega^2}{2}}\right) = \left(\frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1}\right)^2 < F\left(0, \frac{\omega}{\sqrt{2}}\right) < F\left(0, \sqrt{\frac{k_{\max}^2-\omega^2}{2}}\right), \tag{2.4.18}$$

we conclude that there exists a unique value p^* such that $F(0,p^*) = F(\sqrt{k_{\max}^2 - \omega^2}, p^*)$, which concludes the proof. □

Remark 2.4.7. It is interesting to note that this problem is different from the ones already studied in the literature. For instance we showed immediately that the convergence factor is bounded from below, i.e. it is not possible to get a better convergence factor than $\rho^2(k,p) = \frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1}$. We also did not have to exclude the resonance frequency $k = \omega$ by introducing ω_- and ω_+, as in the Helmholtz case [91]; the OSM can benefit from the heterogeneity, leading to $|\rho(k=\omega,p)|^2 < 1$.

We now present two asymptotic results. First we let $h \to 0$, h being the mesh size, and suppose that the maximum frequency supported by the numerical grid scales like $k_{\max} = \pi/h \to \infty$.

Theorem 2.4.8. *When the physical parameters ω and $k_{\min}$ are fixed, $k_{\max} = \frac{\pi}{h}$ and $h \to 0$, then the solution of problem (2.4.15) is given by*

$$p^* = \frac{\sqrt{\omega\pi}}{2} \cdot h^{-1/2} + o(h^{-1/2}), \quad |\rho(k,p^*)|^2 = 1 - \frac{4\sqrt{\omega}}{\sqrt{\pi}} h^{\frac{1}{2}} + o(h^{1/2}). \tag{2.4.19}$$

Proof. For $k_{\max} \to \infty$, $\rho(k_{\max},p) \to 1$, and hence the solution of the minimax problem is given by equioscillation. Inserting the ansatz $p \approx C_p h^{-\alpha}$ into $|\rho(k=\omega,p)|^2 = |\rho(k = k_{\max},p)|^2$ and comparing the leading order terms then gives the result. □

The second asymptotic limit is typical of the Helmholtz equation. As ω increases, in order to control the so called pollution effect [3], we need to decrease significantly h in order to have a good approximation of the solution. Generally, the scaling relation used is $h = \frac{C_h}{\omega^\gamma}$, with $\gamma > 1$. Common values are $\gamma = \frac{3}{2}$, or $\gamma = 2$.

Theorem 2.4.9. *If $k_{\min}$ is fixed, $k_{\max} = \frac{\pi}{h}$, ω goes to infinity and $h = \frac{C_h}{\omega^\gamma}$, with $\gamma > 1$, then the solution of problem (2.4.15) is given by*

$$p^* = \frac{\sqrt{\pi}}{2\sqrt{C_h}} \cdot \omega^{\frac{1+\gamma}{2}} + o(\omega^{\frac{1+\gamma}{2}}), \quad |\rho(k,p^*)|^2 = 1 - \frac{4\sqrt{C_h}}{\sqrt{\pi}} \omega^{\frac{1-\gamma}{2}} + o(\omega^{\frac{1-\gamma}{2}}).$$

Proof. A direct calculation shows that $|\rho(k = k_{\max}, \frac{\omega}{\sqrt{2}})|^2 \to 1$ for $\omega \to \infty$, and thus again the solution is given by equioscillation. Expanding equation $|\rho(k=\omega,p)|^2 = |\rho(k = k_{\max},p)|^2$, with the ansatz $p = C_p \omega^\alpha$ then leads to the desired result. □

2.4.3 Numerical results

We implemented our heterogeneous OSM on a square domain $\Omega := (-1,1) \times (-1,1)$, with $\Omega_1 := (-1,0) \times (-1,1)$ and $\Omega_2 := (0,1) \times (-1,1)$. We used second order centered finite differences for the interior points and first order approximations for the boundary terms. In

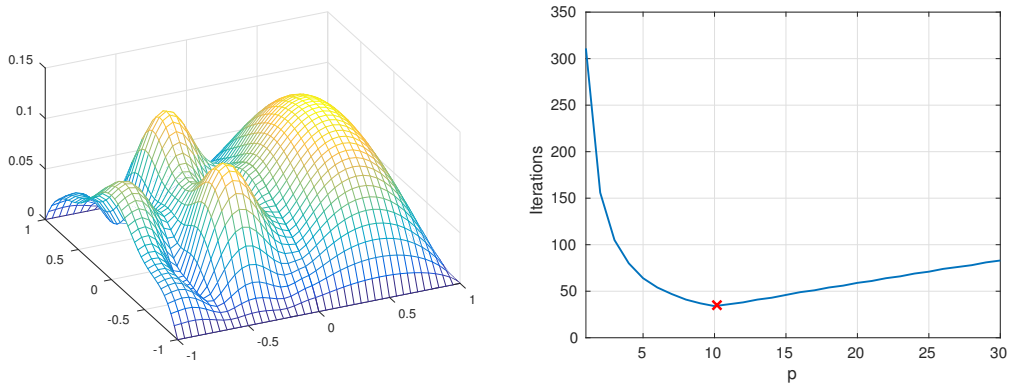


Figure 2.9: Parameters $\omega^2 = 50$, $h = 0.05$. Left: Modulus of $u(x, y)$. Right: Parameter p vs number of iterations. The optimal p given by equioscillation is indicated by a star.

h	Optimal p^*	$\max_k \lvert\rho^2(p^*, k)\rvert$	iterations
$\frac{1}{50}$	16.52	0.4225	53 (810)
$\frac{1}{100}$	23.53	0.55043	73 (1614)
$\frac{1}{200}$	33.37	0.6543	104 (3284)
$\frac{1}{400}$	47.27	0.7403	148 (6554)
ω	Optimal p^*	$\max_k \lvert\rho^2(p^*, k)\rvert$	iterations
10π	34.8451	0.2119	31 (839)
20π	84.7084	0.2622	38 (2954)
40π	205.0570	0.3167	46 (8096)
60π	342.6739	0.3506	48 (>10000)

Table 2.5: The two tables show the behaviour of the heterogeneous OSM under mesh refinement and when ω increases with $h\omega^{\frac{3}{2}}$ held constant.

Figure 2.9 on the left, we show the modulus of the solution of problem (2.4.1) for $\omega^2 = 50$ and $f = 1$. On the right in Figure 2.9, we show a comparison between the optimal numerical value p and the theoretical estimation provided by Theorem 2.4.6. We see that our simplified analysis on unbounded domains is able to give a good approximation of the optimal parameter in the bounded domain context. Finally, we show in Table 1 the behavior of the algorithm when the mesh size h decreases and for large values of ω, with $h\omega^{\frac{3}{2}} = \text{const}$. In brackets, we show the number of iterations required for a non-optimized case, i.e. using $p = 1$. We clearly see that the optimization leads to a much better algorithm, which deteriorates much more slowly when the mesh is refined, and ω increases.

2.5 Probing the Steklov-Poincaré operator

Over the last decades, several theoretical results have been developed to establish optimized transmission conditions for many different PDEs. We have already cited several of these contributions and we proposed a new analysis for heterogeneous PDEs in the previous sections. Despite this large effort and their better convergence with respect to other classical domain decomposition methods, optimized transmission conditions are not so widely. This is mainly due to the strict hypotheses used in the theoretical analysis, that are not always satisfied in practice, and they can discourage the potential user from implementing optimized transmission conditions. This is especially true if the PDEs needs to be solved few times and thus one relies on other domain decomposition methods. Nevertheless, there are cases where one really needs to use OSMs and to have good estimates of the optimized parameters. Heterogeneous problems are instances, since a decoupling approach is often preferable compared to a monolithic one. Then, OSMs provide a robust and simple decoupling framework. Other nonoverlapping domain decomposition methods, like the Dirichlet-Neumann method, are not so robust and they cannot be adapted to physical parameters of different scales between the subdomains. Concerning homogeneous problems, time dependent PDEs or parametric PDEs require the solution of the same PDEs several times, and thus the better convergence rate of OSM can represent a significant advantage. In this paragraph, we discuss numerical procedures to find optimized transmission conditions in those cases where the theoretical analysis falls short.

From (1.3.17) and setting $\mu = 0$, we observe that an OSM can be seen as a fixed point iteration $\lambda_2^{k+1} = \mathscr{T}(s_1, s_2)\lambda_2^k$, where the iteration operator $\mathscr{T} : \Lambda \to \Lambda$ is

$$\mathscr{T}(s_1, s_2) := (s_2 I + \mathscr{S}_2)^{-1}(s_2 I - \mathscr{S}_1)(s_1 I + \mathscr{S}_1)^{-1}(s_1 I - \mathscr{S}_2). \tag{2.5.1}$$

It is possible to show that if $s_1 = s_2 = s$, then $\|\mathscr{T}(s, s)\| < 1$, see [52, Theorem 5.4.2], and thus the sequence $\{\lambda_2^k\}_{k\geq 1}$ converges in Λ thanks to the Banach fixed point theorem. Given the equivalence between (1.3.17) and (1.3.15), this result is another proof of the convergence of OSMs, different from the one originally proposed by Lions in [126]. The discrete counterpart of (2.5.1) is

$$T(s_1, s_2) = (s_2 I + \Sigma_2)^{-1}(s_2 I - \Sigma_1)(s_1 I + \Sigma_1)^{-1}(s_1 I - \Sigma_2), \tag{2.5.2}$$

where I is the identity matrix which derives from Robin transmission conditions. In order to speed up the convergence of the fixed point iteration, we can rely on the wide literature about ADI methods, see [2, Chapter 7.7] and [153, Chapter 7] for a general introduction, and [130, 52] for an application to OSMs. We here summarize these approaches and we distinguish two cases. If Σ_1 and Σ_2 commute, then they share a common eigenbasis $\{\mathbf{v}_j\}_j$ [2, Lemma 7.18]. Defining μ_j^i the eigenvalues of Σ_i associated to the eigenvector $\mathbf{v}_j$, a

direct calculation shows that

$$T(s_1,s_2)\mathbf{v}_j = (s_2I+\Sigma_2)^{-1}(s_2I-\Sigma_1)(s_1I+\Sigma_1)^{-1}(s_1-\mu_j^2)\mathbf{v}_j = (s_2I+\Sigma_2)^{-1}(s_2I-\Sigma_1)\frac{s_1-\mu_j^2}{s_1+\mu_j^1}\mathbf{v}_j$$
$$= (s_2I+\Sigma_2)^{-1}\frac{(s_2-\mu_j^i)(s_1-\mu_j^2)}{s_1+\mu_j^1}\mathbf{v}_j = \frac{s_1-\mu_j^2}{s_1+\mu_j^1}\frac{s_2-\mu_j^1}{s_2+\mu_j^2}\mathbf{v}_j.$$

We thus consider the problem

$$\min_{s_1,s_2}\max\rho(T(s_1,s_2)) = \min_{s_1,s_2}\max_j\left|\frac{s_1-\mu_j^2}{s_1+\mu_j^1}\frac{s_2-\mu_j^1}{s_2+\mu_j^2}\right|. \tag{2.5.3}$$

Supposing $s_1 = s_2 = s$, and introducing the upper and lower bounds α,β, $0<\alpha<\mu_j^i<\beta$, $i=1,2$, $\forall\, j$, we focus on the simpler problem

$$\min_s\max_{x\in[\alpha,\beta]}\left(\frac{s-x}{s+x}\right)^2. \tag{2.5.4}$$

The solution of (2.5.4) is provided by $s^{\text{opt}} = \sqrt{\alpha\beta}$, see for instance [74]. Hence, we have found an optimized choice for the transmission conditions if Σ_1 and Σ_2 commute. Let us briefly note that one could also consider a sequence of parameter $\{s_k\}_k$. For q steps, this choice leads to an iteration operator

$$T(\{s_k\}_k) = \prod_{k=1}^{q}(s_kI+\Sigma_2)^{-1}(s_kI-\Sigma_1)(s_kI+\Sigma_1)^{-1}(s_kI-\Sigma_2).$$

This possibility has been studied under the commutative hypothesis by Wachspress in [154] and [155]. See also [79] for an application to OSMs.
If the matrices do not commute and setting $s_1 = s_2 = s$, one relies on the estimate

$$\rho(T(s)) = \rho((sI+\Sigma_2)^{-1}(sI-\Sigma_1)(sI+\Sigma_1)^{-1}(sI-\Sigma_2))$$
$$= \rho((sI-\Sigma_2)(sI+\Sigma_2)^{-1}(sI-\Sigma_1)(sI+\Sigma_1)^{-1}))$$
$$\leq \|(sI-\Sigma_2)(sI+\Sigma_2)^{-1}\|_2\|(sI-\Sigma_1)(sI+\Sigma_1)^{-1})\|_2 = \max_j\left|\frac{s-\mu_j^2}{s+\mu_j^2}\right|\max_j\left|\frac{s-\mu_j^1}{s+\mu_j^1}\right|. \tag{2.5.5}$$

Supposing that $0<\alpha<\mu_j^i<\beta$, $i=1,2$, $\forall\, j$, we have

$$\max_j\left|\frac{s-\mu_j^2}{s+\mu_j^2}\right|\max_j\left|\frac{s-\mu_j^1}{s+\mu_j^1}\right| \leq \max\left\{\left|\frac{s-\alpha}{s+\alpha}\right|,\left|\frac{s-\beta}{s+\beta}\right|\right\}.$$

The two terms are simultaneously minimized when $s=\sqrt{\alpha\beta}$.
From this analysis we obtain that, in both the commutative and noncommutative case and supposing $s_1 = s_2 = s$, it is reasonably to set $s=\sqrt{\alpha\beta}$. One could estimate α and β

either through variational estimates or using the power method. Both these tecniques are discussed in [52]. From [130] we know that $\alpha \sim O(1)$ while $\beta \sim O(h^{-1})$, thus we obtain $s = O(h^{-\frac{1}{2}})$ which is in agreement with several results in literature, see [74, 160].
Let us remark that this analysis does not lead to any good insights in case we use higher order transmission conditions, since instead of the identity matrix we would have a general sparse matrix. Finally the estimates (2.5.5) split the iteration operator into two parts, one depending on Σ_1, the other on Σ_2, thus neglecting completely the interaction between the Σ_i. It follows that for non commutative matrices, (2.5.5) is not always a sharp upper bound.

In this paragraph, we consider a new numerical procedure to find optimized parameters based on probing which has been already studied in the context of domain decomposition methods. In [143, Section 4.2.6], the authors proposed to probe the matrix Σ to find an approximate matrix $\widetilde{\Sigma}$ to use directly as preconditioner for the Schur complement system (1.3.5). In the frequency filtering method proposed in [156, 157], probing is used to replace inverse matrices with cheap approximations in a block LU decomposition. We refer the interested reader to [92, 93] for a discussion about the relations between frequency filters, analytical factorizations and OSMs. Our renewed interest in probing is mainly motivated by [10], where the authors estimated an optimized Robin parameter by probing the Steklov-Poincaré operator.
With the word 'probing' we mean the numerical procedure through which we estimate a generic matrix G by testing this matrix over a set of vectors. In mathematical terms, given a set of vectors $\mathbf{x}_i$ and $\mathbf{y}_i := G\mathbf{x}_i$, $i \in \mathscr{I} := \{1, ..., M\}$, we study the problem

$$\text{Find } \widetilde{G} \quad \text{such that} \quad \widetilde{G}\mathbf{x}_i = \mathbf{y}_i, \forall i \in \mathscr{I}. \tag{2.5.6}$$

In general we look for a matrix $\widetilde{G}$ with some nice properties(diagonal, tridiagonal, sparse...), so that problem (2.5.6) does not always have a solution. Calling D the set of admissible matrices, it is better to consider the problem

$$\min_{\widetilde{G}\in D} \max_{i\in\mathscr{I}} \|\mathbf{y}_i - \widetilde{G}\mathbf{x}_i\|, \tag{2.5.7}$$

We now observe that if one uses more general transmission conditions described by matrices $\widetilde{\Sigma}_1$ and $\widetilde{\Sigma}_2$, (2.5.2) becomes

$$T(\widetilde{\Sigma}_1, \widetilde{\Sigma}_2) = (\widetilde{\Sigma}_2 + \Sigma_2)^{-1}(\widetilde{\Sigma}_2 - \Sigma_1)(\widetilde{\Sigma}_1 + \Sigma_1)^{-1}(\widetilde{\Sigma}_1 - \Sigma_2). \tag{2.5.8}$$

It is sufficient to assume that $\widetilde{\Sigma}_j$ are positive definite matrices to guarantee the invertibility of $(\widetilde{\Sigma}_j + \Sigma_j)$, $j = 1, 2$. Choosing $\widetilde{\Sigma}_1 = \Sigma_2$ or $\widetilde{\Sigma}_2 = \Sigma_1$, we have $T = 0$, and hence the method is nilpotent. We have simply reobtained that the Steklov-Poincaré operators are optimal transmission operators [32, 136, 88]. Since the assembly of matrices Σ_i, $i = 1, 2$ is too expensive, we propose to approximate them through probing. Unfortunately, this very natural idea turns out to be inefficient.
To see this, let us carry out a continous analysis on a infinite strip $\Omega = \Omega_1 \cup \Omega_2$, with $\Omega_1 = (-\infty, 0) \times (0, 1)$ and $\Omega_2 = (0, \infty) \times (0, 1)$. We consider the Laplace equation and thanks

to symmetry we have $\mathscr{S}_1 = \mathscr{S}_2 =: \mathscr{S}_e$. It is well known that the eigenvectors of $\mathscr{S}_e$ are $v_k = \sin(k\pi y)$, $k \in \mathbb{N}^+$ with eigenvalues $\mu_k = k\pi$ so that $\mathscr{S}_e v_k = \mu_k v_k =: y_k$, see [74]. Suppose now that we look for an operator $S = sI$ with $s \in \mathbb{R}^+$, which corresponds to a Robin transmission condition with parameter s. Choosing as probing functions the normalized functions v_k with $k = 1, \dots N_h$ [2], (2.5.7) becomes

$$\begin{aligned} &\min_{S:S=sI,\ s\in\mathbb{R}^+} \max_{k\in[1,N_h]} \|y_k - Sv_k\| = \min_{s} \max_{k\in[1,N_h]} \|y_k - sv_k\| \\ &= \min_{s} \max_{k\in[1,N_h]} \|\mu_k v_k - sv_k\| = \min_{s} \max_{k\in[1,N_h]} |k\pi - s|. \end{aligned} \tag{2.5.9}$$

The solution of (2.5.9) is $s^* = \frac{N_h\pi}{2}$, while, since $\mathscr{S}_1$ and $\mathscr{S}_2$ commute, the parameter which minimizes the spectral radius of T is $s^{\text{opt}} = \sqrt{N_h\pi}$. The difference between the two estimates can be explained as follows. Probing the optimal Steklov-Poincaré corresponds to minimize the numerator of (2.5.3), neglecting completely that the iteration operator T involves also the inverse of $(\Sigma_i + \widetilde{\Sigma}_i)$, $i = 1,2$. This observation suggests us to consider the minimization problem

$$\min_{\widetilde{\Sigma}_1,\widetilde{\Sigma}_2\in D} \max_{i\in\mathscr{I}} \frac{\|\Sigma_2 x_i - \widetilde{\Sigma}_1 x_i\|}{\|\Sigma_1 x_i + \widetilde{\Sigma}_1 x_i\|} \frac{\|\Sigma_1 x_i - \widetilde{\Sigma}_2 x_i\|}{\|\Sigma_2 x_i + \widetilde{\Sigma}_2 x_i\|}. \tag{2.5.10}$$

We say that this problem is 'consistent' in that sense that if Σ_1, Σ_2 share a common eigenbasis $\{\mathbf{v}_j\}_j$ with eigenvalues $\left\{\mu_j^i\right\}$, $\widetilde{\Sigma}_i = s_i I$, $i = 1,2$, then choosing $\mathbf{x}_j = \mathbf{v}_j$,

$$\min_{\widetilde{\Sigma}_1,\widetilde{\Sigma}_2\in D} \max_{j\in\mathscr{I}} \frac{\|\Sigma_2 \mathbf{x}_j - \widetilde{\Sigma}_1 \mathbf{x}_j\|}{\|\Sigma_1 \mathbf{x}_j + \widetilde{\Sigma}_1 \mathbf{x}_j\|} \frac{\|\Sigma_1 \mathbf{x}_j - \widetilde{\Sigma}_2 \mathbf{x}_j\|}{\|\Sigma_2 \mathbf{x}_j + \widetilde{\Sigma}_2 \mathbf{x}_j\|} = \min_{s_1,s_2} \max_{j\in\mathscr{I}} \in \left| \frac{s_1 - \mu_j^2}{s_1 + \mu_j^1} \frac{s_2 - \mu_j^1}{s_2 + \mu_j^2} \right| = \min_{s_1,s_2\in\mathbb{R}^+} \rho(T(s_1,s_2)). \tag{2.5.11}$$

We stress that the (2.5.10) has to be solved numerically. In our experiments we used the function fminsearch in MATLAB which is based on the Nelder-Mead algorithm. Furthermore, the application of the Schur complements Σ_i on a vector $\mathbf{x}_j$ requires a subdomain solve. These evaluations can be done in parallel, however it is desirable to keep the number of probing vectors small. The choice of $\mathbf{x}_j$ plays a key role to obtain good estimates and it is usually driven by insights provided by theoretical analysis.

2.5.1 Numerical results

In this subsection we present numerical results to validate the approach of (2.5.10) in different cases. We also discuss how to choose properly the probing vectors $\mathbf{x}_i$.

2.5.1.1 Laplace equation in a square box

We start with a sanity check considering a Laplace equation in a domain $\Omega = \Omega_1 \cup \Omega_2$ with $\Omega_1 = (-1,0) \times (0,1)$, $\Omega_2 = (0,1) \times (0,1)$ and $\Gamma = \{0\} \times (0,1)$. Given a discretization of the

[2] We approximate the maximum frequency on the numerical grid with $k_{\max} = N_h$, where N_h is the size of the Schur complement matrix, i.e. the number of degrees of freedom on the interface.

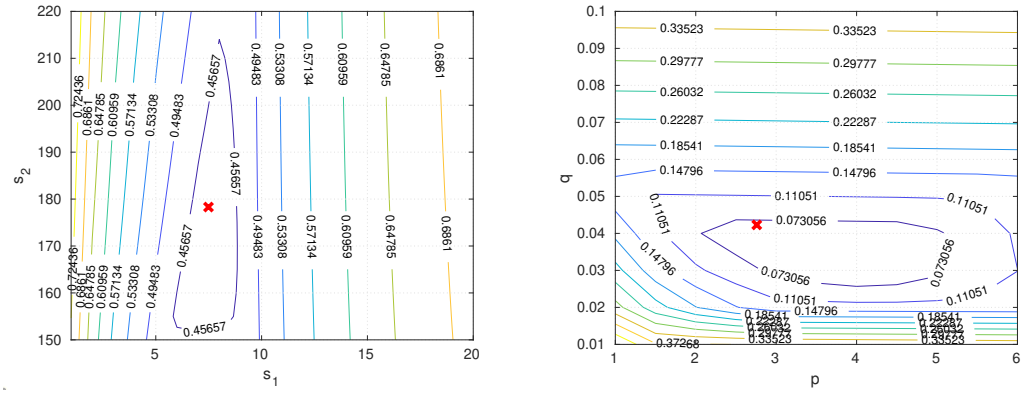


Figure 2.10: Contour plot of the spectral radius of the iteration matrix $T(\widetilde{\Sigma}_1, \widetilde{\Sigma}_2)$ with $\widetilde{\Sigma}_i = s_i I$ (left) and of $T(\widehat{\Sigma}_1, \widehat{\Sigma}_2)$ with $\widehat{\Sigma}_i = pI + qH$ (right). The red crosses are the parameters estimated solving (2.5.10).

interface Γ with N points, we choose as probing vectors the discretization of the following functions

$$x_1 = \sin(\pi y), \quad x_2 = \sin(\sqrt{N_h}\pi y), \quad x_3 = \sin(N_h \pi y). \tag{2.5.12}$$

This choice is motivated by the theoretical analysis in [74] and Theorem 2.1.9, which show that the optimal parameters s_i satisfy equioscillation between the minimum, the maximum and a medium frequency which scales as $\sqrt{N_h}$. We first look for matrices $\widetilde{\Sigma}_i = s_i I$ representing zeroth order double sided optimized transmission conditions. Then, we look for matrices $\widehat{\Sigma}_i = pI + qH$, where H is a tridiagonal matrix $H := \text{Diag}(\frac{2}{h^2}) - \text{Diag}(\frac{1}{h^2}, -1) - \text{Diag}(\frac{1}{h^2}, +1)$, where h is the mesh size. At the continuous level, $\widehat{\Sigma}_i$ represents second order transmission conditions, so that $(1.3.14)_{2,3}$ become

$$\begin{aligned} \frac{\partial u_1^{n+1}}{\partial \mathbf{n}_1} + p u_1^{n+1} - q\frac{\partial^2 u_1^{n+1}}{\partial \tau^2} &= \frac{\partial u_2^{n}}{\partial \mathbf{n}_1} + p u_2^{n} - q\frac{\partial^2 u_2^{n}}{\partial \tau^2}, \qquad \text{on } \Gamma, \\ \frac{\partial u_2^{n+1}}{\partial \mathbf{n}_2} + p u_2^{n+1} - q\frac{\partial^2 u_2^{n+1}}{\partial \tau^2} &= \frac{\partial u_1^{n}}{\partial \mathbf{n}_2} + p u_1^{n} - q\frac{\partial^2 u_1^{n}}{\partial \tau^2}, \qquad \text{on } \Gamma. \end{aligned} \tag{2.5.13}$$

Fig 2.10 shows that solving (2.5.10) with just three probing vectors permits to obtain good estimates in both cases.

2.5.1.2 Second order PDE with curved interface

We now look at a more challenging problem. We solve a second order PDE

$$-\nabla \cdot \nu(\mathbf{x}) \nabla u + \mathbf{a}(\mathbf{x})^\top \cdot \nabla u + \eta(\mathbf{x})\, u = f \quad \text{in } \Omega, \tag{2.5.14}$$

where Ω is represented in Fig 2.11. The interface Γ is a parametric curve $\gamma(t) : [0,1] \to (0, r\sin(\widehat{k}\pi t))$, with $r \in \mathbb{R}^+$. In our first example, we choose $\nu(\mathbf{x}) = 10$, $\mathbf{a}(\mathbf{x}) = (8,0)^\top$, $\eta(\mathbf{x}) = 0$ in Ω_1, $\nu(\mathbf{x}) = 1$, $\mathbf{a}(\mathbf{x}) = (2,0)^\top$, $\eta(\mathbf{x}) = 5$ in Ω_2 , $r = 0.1$ and $\widehat{k} = 4$. Driven by the theoretical

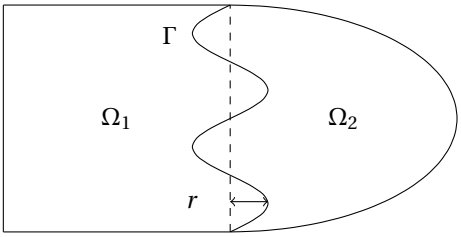


Figure 2.11: Representation of the domain Ω and its decomposition into Ω_1 and Ω_2.

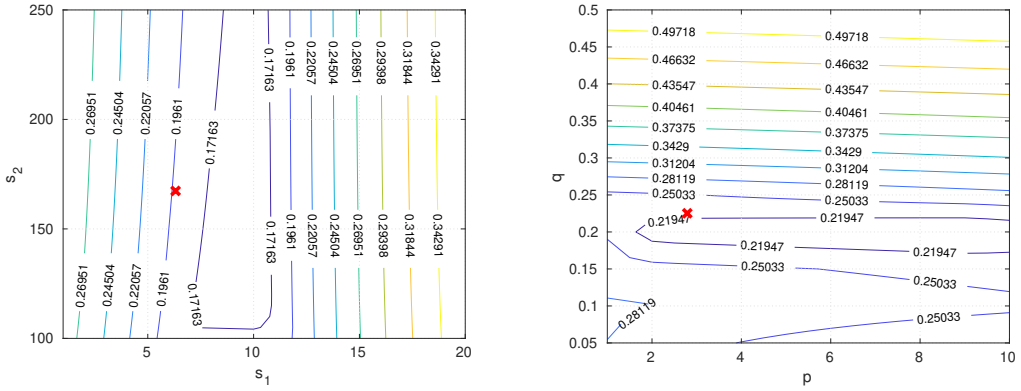


Figure 2.12: Contour plot of the spectral radius of the iteration matrix $T(\widetilde{\Sigma}_1, \widetilde{\Sigma}_2)$ with $\widetilde{\Sigma}_i = s_i I$ (left) and $\widetilde{\Sigma}_i = pI + qH$ (right). The red crosses are the parameters estimated solving (2.5.10).

analysis of Section 2.2, we rescale the transmission conditions according to the physical parameters setting $S_i := f_i(s_i)I$, where $f_i := \nu_i \sqrt{s_i^2 + \frac{a_{i1}^2}{4\nu_i^2} + \frac{a_{i2}^2}{4\nu_i^2} + \frac{\eta_i}{\nu_i}} - \frac{a_{i1}}{2}$, for the zeroth order transmission conditions, and $S_i := f_i(s_i)I + qH$ for the second order ones. In Fig 2.12, we show that (2.5.10) still leads to a good estimate of the optimized parameters both for zeroth and second order transmission conditions.

If we increase the parameter r and $\widehat{k}$, then (2.5.10) loses its accuracy. The reasons are two-fold. On the one hand, the matrices Σ_i becomes strongly non commutative and their eigenvectors are significantly different as Fig 2.13 shows. On the other hand, there is no reason to choose the probing vectors according to (2.5.12). In the case of strong heterogeneity or strong asymmetry of the decomposition into subdomains, we suggest to choose the probing vectors as the eigenvectors of the Steklov-Poincaré operators Σ_i. To approximate the eigenvectors associated to the largest eigenvalues, we select some initial vectors(we can choose for instance x_3 of (2.5.12)) and we compute few iterations, e.g. 2 to 3, of the power method applied to the operators Σ_i. To approximate the eigenvectors associated to the smallest eigenvalues, we apply the power method to Σ_i^{-1}, which is known as the Neumann to Dirichlet operator. All these computations require the solution of a

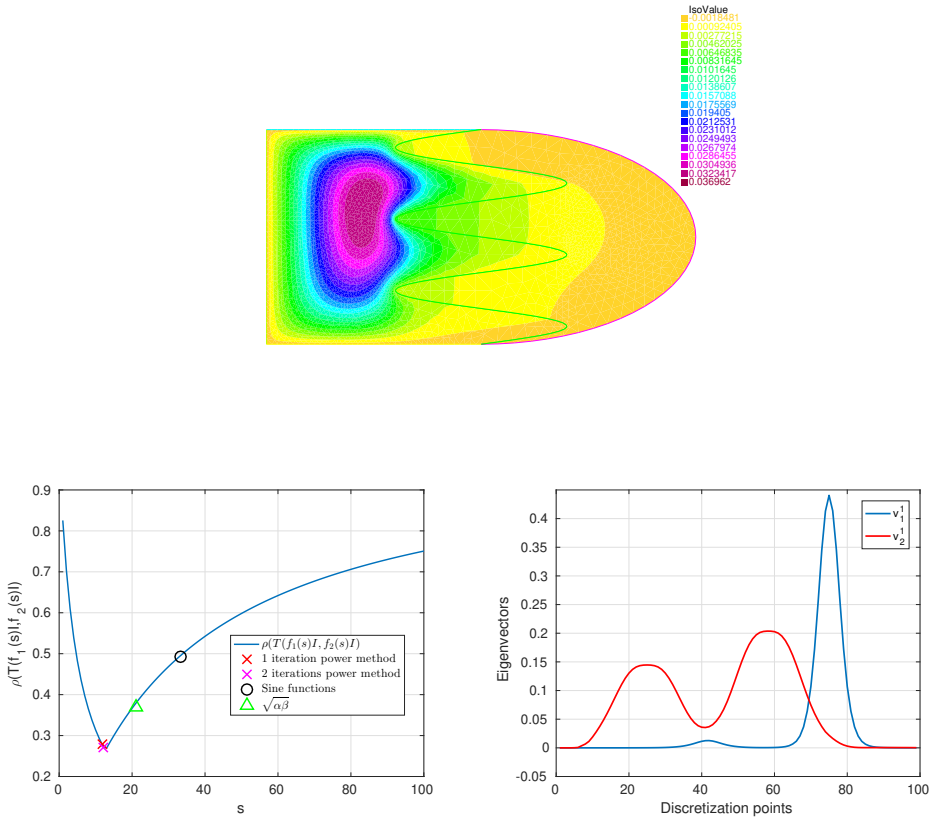


Figure 2.13: The top panel shows the solution of (2.5.14) with $\nu(\mathbf{x}) = 1$, $\mathbf{a}(\mathbf{x}) = (10(y + x^2), 0)^\top$, $\eta(\mathbf{x}) = 0.1(x^2+y^2)$ in Ω_1, $\nu(\mathbf{x}) = 100$, $\mathbf{a}(\mathbf{x}) = (10(1-x), x)^\top$, $\eta(\mathbf{x}) = 0$ in Ω_2, $f(\mathbf{x}) = x^2 + y^2$ in Ω, $r = 0.4$ and $\widehat{k} = 8$. In the bottom-left panel, the crosses represent the optimized parameters obtained solving (2.5.10) with probing vectors computed with k iterations of the power method. The black circle represents the solution of (2.5.10) with the probing vectors in (2.5.12). The green upward-pointing triangle corresponds to $\sqrt{\alpha\beta}$ where $\alpha = \min\{\mu_k^1, \mu_k^2\}$ and $\beta = \max\{\mu_k^1, \mu_k^2\}$. On the right, we plot the eigenvectors associated to the smallest eigenvalues of Σ_j, $j = 1, 2$ for this test case.

Dirichlet or Neumann problem, but they can be performed in parallel. Fig. 2.13 shows that solving (2.5.10) with probing vectors (2.5.12) does not lead to a good estimate of the optimized parameter s in the case $\widetilde{\Sigma}_i = f_i(s)I$. Solving (2.5.10) using the approximation of the eigenvectors of Σ_i instead leads to a perfect estimate. Let us remark that already with only one iteration of the power method we obtain a very satisfactory estimate.

CHAPTER 3

Multilevel optimized Schwarz methods

"The third idea to improve the efficiency of Schwarz's methods is not to use multigrid as a solver in an overall DD context, but to use the DD idea for smoothing in an overall multigrid context."

— U. Trottenberg, C. Oosterlee, A. Schüller, Multigrid, Chapter 6.5.1.

We have seen in Figure 1.2 that one-level domain decomposition methods are in general not weakly scalable, that is, their rate of convergence deteriorates when the number of subdomains grows. Some exceptions are possible in specific geometries, and we have discussed examples in Section 1.4. Two-level methods are needed to achieve scalability. A two-level domain decomposition method consists of two components: a domain decomposition iteration and a coarse correction which can be either additive or multiplicative. For several decades, the main goal of the second level was just to make the subdomains communicate among them to obtain a convergence independent of the possibly large number of subdomains. An example is the Nicolaides coarse space [61, Chapter 4]. Under this perspective, a coarse grid correction for OSM was first studied numerically in [66], where the authors proposed to consider a coarse mesh defined by a single mesh point for each subdomain, see also [34] Variants of this idea were also discussed in [67] where a convergence analysis is carried out.

However, coarse corrections can do much more than just providing scalability; indeed for every domain decomposition method there is a coarse correction, called optimal coarse correction, that makes the domain decomposition method a direct solver, that is the iteration becomes nilpotent [89, 94, 83]. The main idea behind these articles is to identify a space which contains the optimal coarse correction. This space is called complete coarse space and then one aims to find the smallest space possible containing the optimal coarse correction which is denoted as optimal coarse space. Finally, the optimal coarse space, which is still to large, is approximated leading to a so called optimized coarse space. This approach shares similarities with the ideal coarse correction in Algebraic Multigrid, see

for instance [152, Section A.2.3] or [148]. In this Chapter, we follow the same path. We analyze the one-level OSM and we identify in which coarse space lies the optimal coarse correction. However, computing the optimal coarse correction would be as expensive as solving directly the original problem. We therefore approximate the optimal coarse space geometrically, solving on a coarse grid the equation for the optimal coarse correction. These steps lead to the definition of a two-level optimized Schwarz method.

Our method can also be interpreted in another perspective. Following multigrid literature, our method can also be thought as a two-level method consisting of a smoother, the OSM iteration, and a coarse grid solver. Generalizing, we define a multilevel optimized Schwarz methods where on each level we use the OSM as a smoother and we call it multilevel optimized Schwarz method (MOSM). The idea of using domain decomposition methods as smoothers inside a multigrid scheme is not new. Some remarks in this direction are available for instance in Section 15.3.3 of [109], Section 3.4 of [143] and Section 6.5 of [152]. There is also a wide literature regarding multilevel domain decomposition methods which traces back to the 90s and an introduction is available in Chapter 3 of [143]. However, in this research area the authors framed several multilevel preconditioners, into the so-called Schwarz abstract theory, see Chapter 1.2.1, and they provided condition number estimates for the resulting preconditioned systems in very general settings [159, 14]. In this Chapter, we do not consider condition number estimates but we focus on the properties of the iterative method, carrying out a Fourier analysis which, although under more restrictive hypotheses, permits to have a complete description of the method through the derivation of an iteration matrix which acts on the Fourier modes. In addition, the use of Fourier techniques is motivated by our interest in understanding the dependence of the multilevel method on the optimized transmission conditions.

We analyze the OSM because it has a very good smoothing property: in case of overlap, it inherits the smoothing property from the classical Schwarz method which converges exponentially fast for high frequencies. This property can even be enhanced by an adapted choice of the transmission conditions in the OSM. However the potentiality of OSM is remarkable in the case without overlap, essential for heterogeneous problems, in which the classical Schwarz method simply would not work, while the transmission conditions in OSM allow us to tune at will the OSM as a smoother or as a rougher. Thus, even though for homogeneous problems, the classical parallel Schwarz method would do its job, for heterogeneous problems only the OSM has the desired properties. Furthermore, we show that there is no need to develop a complete new theory for the optimized transmission conditions in a multilevel setting. Indeed we show that one can just choose the optimized parameters using the already available literature for the one-level OSM by just changing the range of frequencies in the min-max problems in order to optimize the smoothing property of OSM. We show that these two approaches are asymptotically equivalent as $h \to 0$. We also prove mesh independent convergence for the two-level OSM, recovering the well-known properties of multigrid schemes, see Chapter 2 of [152]. This is a significant improvement over the one-level Schwarz methods which have a mesh dependent convergence and therefore their convergence deteriorates as the number of unknowns

increases [74].

The strength of the approach presented in this chapter lies in its generality and flexibility. In fact, even though the development of efficient smoothers has reached a certain maturity in the multigrid literature, see for instance Chapters 5-8 of [152], one may have to use ad hoc solutions according to the specific equation under study which might be very sophisticated and difficult to implement. Our approach is instead very general since the smoother, being a domain decomposition method, does not change according to the equation and it is straightforward to implement as long as one has a routine for the one-level domain decomposition method.

This Chapter is based extensively on [99]. In Section 3.1, we present the two-level OSM for a nonoverlapping decomposition and in Section 3.2 we propose a convergence analysis based on Fourier expansion. Section 3.3 defines the method for overlapping decompositions and studies its convergence properties. In Section 3.4, we generalize the two-level OSM to a multilevel framework, discussing implementation details on how to modify the residual while moving from one grid to the other to assure we add the right correction on the fine grid. Numerical results are presented in Section 3.5.

3.1 Two-level OSM for a nonoverlapping decomposition

In this section we introduce the two-level OSM for a nonoverlapping decomposition. We consider a second order elliptic PDE,

$$\mathscr{L}u = f \quad \text{on} \quad \Omega, \tag{3.1.1}$$

in the geometrical setting described at the beginning of Chapter 1.3. Given two initial guesses u_1^0, u_2^0, the one-level parallel OSM reads for $n \geq 1$,

$$\begin{array}{llll} \mathscr{L}u_1^n = f & \text{in} \quad \Omega_1, & \partial_x u_1^n + p u_1^n = \partial_x u_2^{n-1} + p u_2^{n-1} & \text{on } \Gamma, \\ \mathscr{L}u_2^n = f & \text{in} \quad \Omega_2, & -\partial_x u_2^n + p u_2^n = -\partial_x u_1^{n-1} + p u_1^{n-1} & \text{on } \Gamma. \end{array} \tag{3.1.2}$$

If we define two functions on Γ as

$$r_1 := -\partial_x u_1^n - p u_1^n + \partial_x u_2^n + p u_2^n \quad \text{and} \quad r_2 := -\partial_x u_1^n + p u_1^n + \partial_x u_2^n - p u_2^n, \tag{3.1.3}$$

and then we solve the coupled system

$$\begin{array}{llll} \mathscr{L}e_1 = 0 & \text{in} \quad \Omega_1, & \partial_x e_1 + p e_1 - \partial_x e_2 - p e_2 = r_1 & \text{on } \Gamma, \\ \mathscr{L}e_2 = 0 & \text{in} \quad \Omega_2, & -\partial_x e_2 + p e_2 + \partial_x e_1 - p e_1 = r_2 & \text{on } \Gamma, \end{array} \tag{3.1.4}$$

we have that $\widetilde{u}_1 := u_1^n + e_1$ and $\widetilde{u}_2 := u_2^n + e_2$ are solutions of problem (3.1.1). Indeed $\widetilde{u}_1, \widetilde{u}_2$ satisfy the PDE in the interior of the subdomains, and from (3.1.4) we have that $\partial_x(e_1 - e_2) = -\partial_x(u_1^n - u_2^n)$ and $(e_1 - e_2) = -(u_1^n - u_2^n)$. Thus

$$\partial_x \widetilde{u}_1 = \partial_x u_1^n + \partial_x e_1 = \partial_x u_1^n - \partial_x(u_1^n - u_2^n) + \partial_x e_2 = \partial_x u_2^n + \partial_x e_2 = \partial_x \widetilde{u}_2^n.$$

Similarly we have that $\widetilde{u}_1 = \widetilde{u}_2$ on Γ. Clearly at the continous level we have that e, such that $e_{|\Omega_j} = e_j$, lies in the complete infinite dimensional coarse space [94]

$$\mathscr{A} := \left\{ v \in H^{1,\text{disc}}(\Omega) : \mathscr{L}(v_{|\Omega_j}) = 0, j = 1,2 \right\},$$

with $H^{1,\text{disc}}(\Omega) := \left\{ v \in L^2(\Omega) : v_{|\Omega_j} \in H^1(\Omega_j), j = 1,2 \right\}$. With this observation, it has been proposed to construct a discrete coarse space $\mathscr{V}_h \subset \mathscr{A}$, a restriction matrix R_c to the coarse space, and to solve the linear system

$$(R_c A R_c^\top)^{-1} R_c \mathbf{e} = R_c(\mathbf{f} - A\mathbf{u}^n).$$

Following this strategy, it is possible to define a complete discrete coarse space $\mathscr{A}_h$ which leads to a direct method. However the complete coarse space is too expensive to use and therefore it is usually approximated obtaining optimized coarse spaces which are subspaces of $\mathscr{A}_h$, see Refs [94, 83, 89] for an overview.

In this manuscript we define a two-level OSM which, inspired by the multigrid method, solves equation (3.1.4) on a coarse mesh. Our two-level method can be summarized as follows: we iterate algorithm (3.1.2) on a fine grid, we define after n_1 iterations the functions r_1 and r_2, we restrict them to a coarse grid where we solve directly system (3.1.4), we interpolate the corrections e_1 and e_2 to the fine grid and we add them to the iterates. To analyze the discrete version of this algorithm we set $\Omega := (-\frac{1}{2}, \frac{1}{2}) \times (0,1)$, $\Omega_1 := (-\frac{1}{2}, 0) \times (0,1)$, $\Omega_2 := (0, \frac{1}{2}) \times (0,1)$, with an interface $\Gamma := \{0\} \times [0,1]$. We discretize equations (3.1.2) on a fine mesh with mesh size $h := \frac{1}{2^\ell}$ and equations (3.1.4) on a coarser mesh with mesh size $H := \frac{1}{2^{\ell-1}}$. Thus the fine mesh has $N_y := 2^\ell - 1$ degrees of freedom in the y direction and $N_x := \frac{N_y+1}{2}$ in the x direction for each subdomain, while the coarser mesh has $N_y^c := \frac{N_y+1}{2} - 1$ and $N_x^c := \frac{N_y^c+1}{2}$. Therefore each subdomain has $N := N_y N_x$ degrees of freedom on the fine mesh and $N^c := N_y^c N_x^c$ on the coarse one. In the following we use the index ℓ to indicate which mesh we are considering. In [146] the authors introduced the augmented system $\widetilde{A}_\ell \widetilde{\mathbf{u}}_\ell = \widetilde{\mathbf{f}}_\ell$, which contains the variables at the interface Γ twice, with

$$\widetilde{A}_\ell = \begin{pmatrix} A_{1,\ell} & -B_{12,\ell} \\ -B_{21,\ell} & A_{2,\ell} \end{pmatrix} \in \mathbb{R}^{2N,2N},$$

where $A_{j,\ell} \in \mathbb{R}^{N,N}$ is the discrete Laplacian in the domain Ω_j with Robin boundary conditions on Γ, $B_{ji,\ell}$ are interface operators, $\widetilde{\mathbf{f}}_\ell = [\mathbf{f}_{1,\ell}, \mathbf{f}_{2,\ell}]$ is the force vector and $\widetilde{\mathbf{u}}_\ell = [\mathbf{u}_{1,\ell}, \mathbf{u}_{2,\ell}] \in \mathbb{R}^{2N}$ is the vector of the degrees of freedom on the mesh indexed by ℓ. They showed that the discrete version of eq. (3.1.2), i.e.

$$A_{j,\ell}\mathbf{u}_{j,\ell}^{n+1} = \mathbf{f}_{j,\ell} + \sum_{k \neq j} B_{jk,\ell}\mathbf{u}_{k,\ell}^n, \quad j = 1,2,$$

is equivalent to the algebraic iterative method ORAS, which in the correction form reads

$$\widetilde{\mathbf{u}}_\ell^{n+1} = S_\ell(\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^{n-1}, \widetilde{f}_\ell) := \widetilde{\mathbf{u}}^n + \sum_{j=1}^{2} R_{j,\ell}^T A_{j,\ell}^{-1} R_{j,\ell}(\widetilde{\mathbf{f}}_\ell - \widetilde{A}_\ell \widetilde{\mathbf{u}}_\ell^n), \tag{3.1.5}$$

where $R_{j,\ell} \in \mathbb{R}^{N,2N}$ are restriction operators on the domain Ω_j. We emphasize that the residual $\mathbf{r}_\ell^n = \widetilde{\mathbf{f}}_\ell - \widetilde{A}_\ell \widetilde{\mathbf{u}}_\ell^n$ is a vector with zero entries except for the degrees of freedom associated to the interface, where it represents a discretization of the functions in (3.1.3). Thus, we construct the restriction operator R_ℓ so that it acts as the full weighting restriction operator $R_{1D} \in \mathbb{R}^{N_y^c, N_y}$ for the points on the interface, and it has zero blocks corresponding to the interior degrees of freedom,

$$R_\ell \mathbf{r}_\ell = \begin{pmatrix} 0 & & & \\ & R_{1D} & & \\ & & R_{1D} & \\ & & & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \mathbf{r}_{1\Gamma,\ell} \\ \mathbf{r}_{2\Gamma,\ell} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ R_{1D}\mathbf{r}_{1\Gamma,\ell} \\ R_{1D}\mathbf{r}_{2\Gamma,\ell} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{r}_{1\Gamma,\ell-1} \\ \mathbf{r}_{2\Gamma,\ell-1} \\ 0 \end{pmatrix} = \mathbf{r}_{\ell-1}.$$

Other possible choices are available: one could replace the zero blocks with 2D full weighting restriction operators R_{2D}, or with straight injection operators. This change would not affect the method, since they all map a zero function on the fine mesh to a zero function on the coarse mesh. Therefore the properties of the restriction operators are uniquely defined once we characterized the action of the restriction operator on the interface. This is an advantage of the two-level OSM, and of a large class of two-level domain decomposition methods: they do not require to restrict on the whole volume but only on the interfaces, which are 1-D curves for two dimensional problems, or 2-D surfaces for three dimensional problems. In Chapter 4 we will further introduce a new framework of two-level and multilevel domain decomposition methods defined directly on the interfaces. On the coarse mesh, we solve the restricted residual equation inverting the operator $\widetilde{A}_{\ell-1}$, which corresponds to a direct discretization of the original problem on the mesh indexed by $\ell - 1$. Finally concerning the interpolation operator, we define $I_\ell = \text{diag}(I_{2D,\ell}, I_{2D,\ell})$, with $I_{2D,\ell} \in \mathbb{R}^{N,N^c}$ being the standard linear interpolation operator from the coarse to the fine grid. Another possible choice is to define I_ℓ^A which interpolates on the interface and then extends harmonically on the fine grid. With all these ingredients, the algorithm we have described previously at the continuous level to solve the continous problem (3.1.1) can be rewritten in the discrete form as

Algorithm 1: Function two-level OSM($\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^0, \widetilde{\mathbf{f}}_\ell$).

- For $n = 1 : n_1$, $\widetilde{\mathbf{u}}_\ell^n \leftarrow S_\ell(\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^{n-1}, \widetilde{\mathbf{f}}_\ell)$.
- $\mathbf{r}_\ell \leftarrow \widetilde{\mathbf{f}}_\ell - \widetilde{A}_\ell \widetilde{\mathbf{u}}_\ell^{n_1}$.
- $\mathbf{r}_{\ell-1} \leftarrow R_\ell \mathbf{r}_\ell$.
- $\widetilde{\mathbf{e}}_{\ell-1} \leftarrow \widetilde{A}_{\ell-1}^{-1} \mathbf{r}_{\ell-1}$.
- $\widetilde{\mathbf{u}}_\ell^{n_1} \leftarrow \widetilde{\mathbf{u}}_\ell^{n_1} + I_\ell \widetilde{\mathbf{e}}_{\ell-1}$.
- For $n = n_1 + 1 : n_2$, $\widetilde{\mathbf{u}}_\ell^n \leftarrow S_\ell(\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^{n-1}, \widetilde{\mathbf{f}}_\ell)$.
- Return $\widetilde{\mathbf{u}}_\ell^{n_2}$.

Considering the error equation, i.e. $\widetilde{f}_\ell = 0$, Algorithm 1 leads to the classical iteration matrix of a two-level method,

$$S_{2LOSM} \widetilde{\mathbf{u}}_\ell = S_\ell^{n_2} \left(I - I_\ell \widetilde{A}_{\ell-1}^{-1} R_\ell \widetilde{A}_\ell \right) S_\ell^{n_1} \widetilde{\mathbf{u}}_\ell, \tag{3.1.6}$$

where $I \in \mathbb{R}^{2N,2N}$ is the identity matrix and S_ℓ^n corresponds to n iterations of the iterative method defined in (3.1.5). The convergence properties of the two-level method depend on an effective team play between the smoother and the coarse correction. As in the multigrid literature, n_1 and n_2 are mainly chosen heuristically. The stronger is the smoothing property of the smoother, the smaller n_1 and n_2 can be. We show in Sections 3.2.2 and 3.3.1 that OSMs are very efficient smoothers, so that common choices for n_1 and n_2 are 1 or 2. In the numerical experiments section, we set $n_1 = n_2 = 2$. We close this section emphasizing that the exact correction given by $\widetilde{\mathbf{e}}_\ell = \widetilde{A}_\ell^{-1}\mathbf{r}_\ell$ is a discrete harmonic function (in the PDE sense) in the interior of the subdomains. The coarse correction, being $\widetilde{\mathbf{e}}_{\ell-1} = \widetilde{A}_{\ell-1}^{-1} R_\ell \mathbf{r}_\ell$ is still harmonic on the coarse mesh, but the interpolated correction $I_\ell \widetilde{\mathbf{e}}_{\ell-1}$ is not harmonic on the fine grid. In other words, the linear interpolator destroys the harmonicity of the correction and thus we conclude that with the linear interpolator I_ℓ we cannot have a direct method! The interpolator I_ℓ^A should therefore be preferred since it adds a correction which lies in the complete discrete coarse space. However its use is more expensive since it requires to solve subdomain problems. In the rest of the manuscript we will always use the geometric interpolator I_ℓ if not explicitly stated otherwise.

3.2 Convergence analysis for the two-level OSM

Our analysis is based on a semi-discrete study of Algorithm 1. We take into account the mesh properties in the y direction, while we consider a continuous problem in the x direction. We carry out the calculations supposing that $\mathscr{L} = -\Delta$, but at the end of the subsection we discuss how the analysis adapts to general second order operators. Furthermore we assume n_1 and n_2 to be even numbers. Accordingly, the error equation of (3.1.1) can be written as

$$-\partial_{xx} u - \partial_{yy,h} u = 0 \quad \text{on} \quad \Omega_j,\ j = 1,2,$$

where, using separation of variables, u is semi discrete as well, i.e. $u = \phi(x)\psi(jh)$, where $j = 1,\dots,N_y$. Inserting this ansatz we obtain the eigenvalue equation in the y direction

$$-\partial_{yy,h}\psi(jh) = \gamma^2 \psi(jh),$$

whose solutions are given by $\psi_k(jh) := \sin(k\pi jh)$, $j = 1,\dots,N_y$, $k = 1,\dots,N_y$ and $\gamma^2(k) := \frac{4}{h^2}\sin^2(k\pi\frac{h}{2})$. Solving the equation in x we obtain $\phi_k(x) = A(k)e^{\lambda(k)x} + B(k)e^{-\lambda(k)x}$, with $\lambda(k) = \sqrt{\gamma^2(k)}$. To simplify further the problem we suppose that the domain is unbounded in the x direction so that the general solution is given by

$$u_1 = \sum_{k=1}^{N_y} A(k)\psi_k e^{\lambda(k)x} \quad \text{and} \quad u_2 = \sum_{k=1}^{N_y} B(k)\psi_k e^{-\lambda(k)x}. \tag{3.2.1}$$

The initial guesses u_1^0 and u_2^0 can be written in the general form of eq. (3.2.1) for a proper choice of $A(k)$ and $B(k)$. After an even number n_1 of iterations of the smoother, standard

calculations, see for instance [74], show that

$$u_1^{n_1} = \sum_{k=1}^{N_y} \rho(k,p)^{n_1} A(k)\psi_k e^{\lambda(k)x} \text{ and } u_2^{n_1} = \sum_{k=1}^{N_y} \rho(k,p)^{n_1} B(k)\psi_k e^{-\lambda(k)x},$$

where $\rho(k,p) = \left(\frac{\lambda(k)-p}{\lambda(k)+p}\right)$. If n_1 is not even then the role of $A(k)$ and $B(k)$ is flipped, and the analysis follows the same calculations, but the notation to keep track of both cases becomes cumbersome. Thus we prefer to assume that n_1 is even for the sake of clarity. Then we compute the residuals r_1 and r_2 in (3.1.3),

$$\begin{aligned} r_1 &= \textstyle\sum_{k=1}^{N_y} g_-(k)A(k)\rho(k,p)^{n_1}\psi_k + \sum_{k=1}^{N_y} g_+(k)B(k)\rho(k,p)^{n_1}\psi_k, \\ r_2 &= \textstyle\sum_{k=1}^{N_y} g_+(k)A(k)\rho(k,p)^{n_1}\psi_k + \sum_{k=1}^{N_y} g_-(k)B(k)\rho(k,p)^{n_1}\psi_k, \end{aligned} \tag{3.2.2}$$

where $g_-(k) := -\lambda(k) - p$ and $g_+(k) := -\lambda(k) + p$. We observe that r_1 and r_2 are one dimensional functions in the variable y, which are sums of eigenfunctions of the discrete Laplacian. Well known results are available for the action of the full weighted restriction operator R_{1D} and the linear interpolation operator $I_{1D} := 2R_{1D}^\top$ on these functions, see for instance Chapter 2 of [109]. In particular, defining $\widetilde{k} := N_y + 1 - k$, we have

$$R_{1D}(e_k\psi_k + e_{\widetilde{k}}\psi_{\widetilde{k}}) = (e_k c_k^2 - e_{\widetilde{k}} s_k^2)\phi_k, \tag{3.2.3}$$

where $c_k := \cos(k\pi\frac{h}{2})$, $s_k := \sin(k\pi\frac{h}{2})$ and $\phi_{k,j} := \sin(k\pi jH)$ with $k, j \in \mathcal{V} := \{1,2,...,N_y^c\}$ are the eigenvectors of the 1D discrete Laplacian on the coarse grid. The eigenfunction $\psi_{\frac{N_y+1}{2}}$ is actually mapped to zero by the restriction operator, that is $R_{1D}\psi_{\frac{N_y+1}{2}} = 0$, and thus this frequency is not represented on the coarse level. Using these results we obtain

$$\begin{aligned} R_{1D}r_1 &= \sum_{k=1}^{N_y^c} \phi_k \left[\rho(k)^{n_1}\left(g_-(k)A(k) + g_+(k)B(k)\right)c_k^2 - \rho(\widetilde{k})^{n_1}\left(g_-(\widetilde{k})A(\widetilde{k}) + g_+(\widetilde{k})B(\widetilde{k})\right)s_k^2\right], \\ R_{1D}r_2 &= \sum_{k=1}^{N_y^c} \phi_k \left[\rho(k)^{n_1}\left(g_+(k)A(k) + g_-(k)B(k)\right)c_k^2 - \rho(\widetilde{k})^{n_1}\left(g_+(\widetilde{k})A(\widetilde{k}) + g_-(\widetilde{k})B(\widetilde{k})\right)s_k^2\right], \end{aligned} \tag{3.2.4}$$

where for the sake of brevity we omitted the dependence of $\rho(k,p)$ on p. On the coarse mesh the general solution of the semi-discrete Laplace equation is again given by a formula similar to (3.2.1),

$$e_1 = \sum_{k=1}^{N_y^c} \overline{A}(k)\phi_k e^{\lambda_c(k)x} \quad \text{and} \quad e_2 = \sum_{k=1}^{N_y^c} \overline{B}(k)\phi_k e^{-\lambda_c(k)x}, \tag{3.2.5}$$

where $\lambda_c^2(k) := \frac{4}{H^2}\sin^2(k\pi\frac{H}{2})$ are the eigenvalues of the 1D Laplacian on the coarse mesh. Imposing the boundary conditions to solve the residual system (3.1.4), we obtain

$$\begin{aligned} (\lambda_c(k) + p)\bar{A}(k) + (\lambda_c(k) - p)\bar{B}(k) &= R_{1D}r_1(k), \\ (\lambda_c(k) - p)\bar{A}(k) + (\lambda_c(k) + p)\bar{B}(k) &= R_{1D}r_2(k), \end{aligned} \tag{3.2.6}$$

which leads to

$$\begin{aligned} \bar{A}(k) &= \frac{R_{1D}r_1(k) + R_{1D}r_2(k)}{4\lambda_c(k)} + \frac{R_{1D}r_1(k) - R_{1D}r_2(k)}{4p}, \\ \bar{B}(k) &= \frac{R_{1D}r_1(k) + R_{1D}r_2(k)}{4\lambda_c(k)} + \frac{R_{1D}r_2(k) - R_{1D}r_1(k)}{4p}. \end{aligned}$$

The last step is to interpolate the correction to the fine grid. Since we deal with a semi-discrete analysis, we can use the results on the interpolation of the eigenvectors of the Laplace operator [109]. In particular we have that, $\forall k \in \mathcal{V}$, $I_{1D}\phi_k = c_k^2\psi_k - s_k^2\psi_{\widetilde{k}}$. It follows that

$$\begin{aligned} u_1^{n_1} + I_\ell e_1 &= \sum_{k=1}^{N_y} \left(\rho(k,p)^{n_1} A(k) + d_k^2 \bar{A}(k)\right) e^{\lambda(k)x}\psi_k, \\ u_2^{n_1} + I_\ell e_2 &= \sum_{k=1}^{N_y} \left(\rho(k,p)^{n_1} B(k) + d_k^2 \bar{B}(k)\right) e^{-\lambda(k)x}\psi_k, \end{aligned}$$

where $d_k^2 = c_k^2$ if $k \le N_y^c$, $d_k^2 = -s_k^2$ if $k \ge N_y^c + 2$ and $d_k^2 = 0$ if $k = N_y^c + 1$. Algebraic calculations allow us to write a linear relation which maps the coefficients $A(k), B(k), A(\widetilde{k}), B(\widetilde{k})$ after one step of this two-level method. Denoting with $\rho = \rho(k,p)$, $\widetilde{\rho} = \rho(\widetilde{k},p)$ and $\mathbf{v}_k^n = \left(A^n(k), B^n(k), A^n(\widetilde{k}), B^n(\widetilde{k})\right)^\top$, we obtain

$$\mathbf{v}_k^n = G_k^{n_2}\widetilde{D}_k G_k^{n_1}\mathbf{v}_k^{n-1} \quad \forall k \in \mathcal{V}, \tag{3.2.7}$$

where

$$\widetilde{D}_k := \begin{pmatrix} \left(1-\frac{c_k^4}{2}\left(1+\frac{\lambda(k)}{\lambda_c(k)}\right)\right) & \frac{c_k^4}{2}\left(1-\frac{\lambda(k)}{\lambda_c(k)}\right) & \frac{c_k^2 s_k^2}{2}\left(1+\frac{\lambda(\widetilde{k})}{\lambda_c(k)}\right) & \frac{c_k^2 s_k^2}{2}\left(\frac{\lambda(\widetilde{k})}{\lambda_c(k)}-1\right) \\ \frac{c_k^4}{2}\left(1-\frac{\lambda(k)}{\lambda_c(k)}\right) & \left(1-\frac{c_k^4}{2}\left(1+\frac{\lambda(k)}{\lambda_c(k)}\right)\right) & \frac{c_k^2 s_k^2}{2}\left(\frac{\lambda(\widetilde{k})}{\lambda_c(k)}-1\right) & \frac{c_k^2 s_k^2}{2}\left(1+\frac{\lambda(\widetilde{k})}{\lambda_c(k)}\right) \\ \frac{c_k^2 s_k^2}{2}\left(1+\frac{\lambda(k)}{\lambda_c(k)}\right) & \frac{c_k^2 s_k^2}{2}\left(\frac{\lambda(k)}{\lambda_c(k)}-1\right) & \left(1-\frac{s_k^4}{2}\left(1+\frac{\lambda(\widetilde{k})}{\lambda_c(k)}\right)\right) & \frac{s_k^4}{2}\left(1-\frac{\lambda(\widetilde{k})}{\lambda_c(k)}\right) \\ \frac{c_k^2 s_k^2}{2}\left(\frac{\lambda(k)}{\lambda_c(k)}-1\right) & \frac{c_k^2 s_k^2}{2}\left(1+\frac{\lambda(k)}{\lambda_c(k)}\right) & \frac{s_k^4}{2}\left(1-\frac{\lambda(\widetilde{k})}{\lambda_c(k)}\right) & \left(1-\frac{s_k^4}{2}\left(1+\frac{\lambda(\widetilde{k})}{\lambda_c(k)}\right)\right) \end{pmatrix}, \tag{3.2.8}$$

$$G_k^n := \begin{pmatrix} \rho(k)^n & & & \\ & \rho(k)^n & & \\ & & \rho(\widetilde{k})^n & \\ & & & \rho(\widetilde{k})^n \end{pmatrix}. \tag{3.2.9}$$

The smoother is described by the matrix G_k^n while $\widetilde{D}_k$ takes into account the coarse correction. Denoting with $\mathbf{e}^n = \left(\mathbf{v}_1^n, ..., \mathbf{v}_{N_y^c}^n, A\left(\frac{N_y+1}{2}\right), B\left(\frac{N_y+1}{2}\right)\right)^\top$, we conclude that $\mathbf{e}^n = T\mathbf{e}^{n-1}$, where

$$T = \begin{pmatrix} G_1^{n_2}\widetilde{D}_1 G_1^{n_1} & & & & \\ & \ddots & & & \\ & & G_{N_y^c}^{n_2}\widetilde{D}_{N_y^c} G_{N_y^c}^{n_1} & & \\ & & & \rho\left(\frac{N_y+1}{2},p\right)^{n_1+n_2} & \\ & & & & \rho\left(\frac{N_y+1}{2},p\right)^{n_1+n_2} \end{pmatrix}. \tag{3.2.10}$$

Remark 3.2.1 (Extension to more general differential operators). Equation (3.2.10) has been obtained supposing $\mathcal{L} = -\Delta$, but it can be readily extended to more general operators. The necessary hypothesis for the calculations are the assumptions on the geometry of the problem, on the use of a uniform mesh along the interface and that $\psi_k(jh) = \sin(k\pi jh)$, so that we can characterize the action of the restriction and prolongation operators. As long as these assumptions are verified, one can consider a general equation

$(-\nu\Delta + a_1\partial_x u + a_2\partial_y + c)u = 0$. If $a_1 = a_2 = 0$ and $c \neq 0$, then equation (3.2.10) is still valid replacing $\lambda(k)$ with $\lambda(k) = \sqrt{\gamma(k)^2 + \frac{c}{\nu}}$ and using the corresponding convergence factor [74]. If only $a_2 = 0$, then using the expansions

$$u_1 = \sum_{k=1}^{N_y} A(k)\psi_k e^{\lambda_+(k)x} \quad \text{and} \quad u_2 = \sum_{k=1}^{N_y} B(k)\psi_k e^{-\lambda_-(k)x}, \tag{3.2.11}$$

with $\lambda(k)_{+,-} = \frac{a_1 \pm \sqrt{a_1^2 + 4\nu(\frac{4}{h^2}\sin^2(k\pi\frac{h}{2})) + 4\nu c}}{2\nu}$, and carrying out the same calculations, one can derive a similar iteration matrix. The case $a_2 \neq 0$ cannot be treated in this framework because it leads to $\psi_k(jh) \neq \sin(k\pi jh)$, see Chapter 2.2.3 for more details on tangential advection.

3.2.1 Optimization of the semidiscrete nonoverlapping two-level OSM

To optimize the parameter of the two-level method one would have to solve the minimization problem $\min_p \rho(T)$, which, T being block diagonal, is equivalent to minimizing the spectral radii of the matrices $G_k^{n_2}\widetilde{D}_k G_k^{n_1}$ and $\rho\left(\frac{N+1}{2}, p\right)^{n_1+n_2}$. However the eigenvalues of the matrices $G_k^{n_2}\widetilde{D}_k G_k^{n_1}$ are lengthy expressions. Thus, we look for a sharp upper bound of $\rho(T)$. We first prove the following Lemma.

Lemma 3.2.2. *Defining* $\Gamma(k,p) := 3\rho(k,p)^{n_1+n_2}s_k^2$, *we have*

$$\rho(T) \leq \max_{k\in[1,N_y]} \Gamma(k,p). \tag{3.2.12}$$

Proof. We define the matrix $\widetilde{T}$ which is obtained from T replacing the blocks $G_k^{n_2}\widetilde{D}_k G_k^{n_1}$ with $D_k := \widetilde{D}_k G_k^{n_1} G_k^{n_2} = \widetilde{D}_k G_k^n$, where $n := n_1 + n_2$. A classical property of the spectral radius states that $\rho(G_k^{n_2}\widetilde{D}_k G_k^{n_1}) = \rho(\widetilde{D}_k G_k^{n_1} G_k^{n_2}) = \rho(D_k)$. Therefore we have $\rho(T) = \rho(\widetilde{T}) \leq \|\widetilde{T}\|_1$. We note that due to the diagonal structure of the matrix $\widetilde{T}$,

$$\|\widetilde{T}\|_1 = \max\left\{\max_{k\in\mathcal{V}}\|D_k\|_1, \rho\left(\frac{N_y+1}{2}, p\right)^n\right\}.$$

Thus we focus on the term $\|D_k\|_1$. Using the trigonometric formula $\sin(2x) = 2\sin(x)\cos(x)$, we obtain

$$\frac{\lambda(k)}{\lambda_c(k)} = \frac{\frac{2}{h}\sin(k\pi\frac{h}{2})}{\frac{2}{H}\sin(k\pi\frac{H}{2})} = \frac{H\sin(k\pi\frac{h}{2})}{h\sin(k\pi h)} = \frac{1}{\cos(k\pi\frac{h}{2})} = \frac{1}{c_k} > 1, \quad \forall k \in \mathcal{V},$$
$$\frac{\lambda(\widetilde{k})}{\lambda_c(k)} = \frac{\frac{2}{h}\sin(\frac{\pi}{2} - k\pi\frac{h}{2})}{\frac{2}{H}\sin(k\pi\frac{H}{2})} = \frac{H\cos(k\pi\frac{h}{2})}{h\sin(k\pi h)} = \frac{1}{\sin(k\pi\frac{h}{2})} = \frac{1}{s_k} > 1, \quad \forall k \in \mathcal{V}.$$

Substituting these expressions into (3.2.8), direct calculations yield

$$\|D_k\|_1 = \max\left\{\rho^n\left(1 - c_k^4 + c_k s_k^2\right), \widetilde{\rho}^n\left(1 - s_k^4 + c_k^2 s_k\right)\right\}.$$

Exchanging the order of the max operations over a finite set we have

$$\begin{aligned}\max_{k\in\mathcal{V}}\|D_k\| &= \max_{k\in\mathcal{V}}\max\left\{\rho^n\left(1-c_k^4+c_k s_k^2\right),\widetilde{\rho}^n\left(1-s_k^4+c_k^2 s_k\right)\right\}\\ &= \max\left\{\max_{k\in\mathcal{V}}\rho^n\left(1-c_k^4+c_k s_k^2\right),\max_{k\in\mathcal{V}}\widetilde{\rho}^n\left(1-s_k^4+c_k^2 s_k\right)\right\}.\end{aligned}$$

Now we proceed with a change of variable in the second term in the curly brackets. Due to our hypothesis on the mesh, we have that $h=\frac{1}{N_y+1}$, so that $\rho(\widetilde{k},p)=\left(\frac{\frac{2}{h}\sin((N_y+1-k)\pi\frac{h}{2})-p}{\frac{2}{h}\sin((N_y+1-k)\pi\frac{h}{2})+p}\right)=\left(\frac{\frac{2}{h}c_k-p}{\frac{2}{h}c_k+p}\right)$, where we used the trigonometric identity $\sin(\frac{\pi}{2}-x)=\cos(x)$. Using again this relation and denoting with $\mathcal{Z}:=\left\{\frac{N_y+1}{2}+1,...,N_y\right\}$, we conclude that

$$\begin{aligned}\max_{k\in\mathcal{V}}\left(\frac{\frac{2}{h}c_k-p}{\frac{2}{h}c_k+p}\right)^n\left(1-s_k^4+c_k^2 s_k\right) &= \max_{k\in\mathcal{Z}}\left(\frac{\frac{2}{h}s_k-p}{\frac{2}{h}s_k+p}\right)^n\left(1-c_k^4+s_k^2 c_k\right)\\ &= \max_{k\in\mathcal{Z}}\left(\rho(k,p)^n\left(1-c_k^4+s_k^2 c_k\right)\right).\end{aligned}$$

Thus we obtain the equality,

$$\|\widetilde{T}\|_1=\max\left\{\max_{k\in\mathcal{V}\cup\mathcal{Z}}\left(\rho(k,p)^n\left(1-c_k^4+s_k^2 c_k\right)\right),\rho\left(\frac{N+1}{2},p\right)^n\right\}.$$

Now we relax the discrete constraint and we consider the continuous frequencies $k\in[1,\frac{N_y+1}{2})\cup(\frac{N_y+1}{2},N_y]$. Clearly it holds that

$$\begin{aligned}&\max\left\{\max_{k\in\mathcal{V}\cup\mathcal{Z}}\left(\rho(k,p)^n\left(1-c_k^4+s_k c_k^2\right)\right),\rho\left(\frac{N_y+1}{2},p\right)^n\right\}\leq\\ &\max\left\{\max_{k\in[1,\frac{N_y+1}{2})\cup(\frac{N_y+1}{2},N_y]}\left(\rho(k,p)^n\left(1-c_k^4+s_k c_k^2\right)\right),\rho\left(\frac{N_y+1}{2},p\right)^n\right\}.\end{aligned}$$

We now use the key observation that

$$\begin{aligned}\lim_{k\to\frac{N_y+1}{2}}\rho(k,p)^n\left(1-c_k^4+s_k^2 c_k\right) &= \rho\left(\frac{N_y+1}{2},p\right)^n\left(1-\left(\frac{\sqrt{2}}{2}\right)^4+\left(\frac{\sqrt{2}}{2}\right)^3\right) &(3.2.13)\\ &> \rho\left(\frac{N_y+1}{2},p\right)^n, &(3.2.14)\end{aligned}$$

and hence we can bound $\|\widetilde{T}\|_1$ as

$$\|\widetilde{T}\|_1\leq\max_{k\in[1,N_y]}\rho(k,p)^n\left(1-c_k^4+s_k^2 c_k\right).$$

To simply further the right hand side of this inequality we use the relation

$$(1-c_k^4+s_k^2 c_k)=s_k^2(1+c_k^2+c_k)\leq s_k^2(3-s_k^2)\leq s_k^2 3,\quad\forall k\in[1,N_y],$$

and defining $\Gamma(k,p):=3\rho(k,p)^n s_k^2$ we obtain the desired bound. □

We now consider the problem $\min_p \|\widetilde{T}\|_1$. From now on we restrict our analysis to the case $n_1 + n_2 = 2$ and due to Lemma 3.2.2, we study the simpler problem $\min_p \max_{k\in[0,N_y]} \Gamma(k,p)$, where we expanded the range of frequencies to $k \in [0, N_y]$.

Theorem 3.2.3. *Assuming that $n_1 + n_2 = 2$, the solution of the min-max problem*

$$\min_p \max_{k\in[0,N_y]} \Gamma(k,p) \tag{3.2.15}$$

is given by

$$p^* = \frac{(2\sqrt{6}+2\sqrt{3}-2\sqrt{2}-4)\sin(\frac{1}{2}hN_y\pi)}{h}, \tag{3.2.16}$$

which is the unique root of the non linear equation

$$\Gamma(\widetilde{k},p) = \Gamma(N_y,p),$$

where $\widetilde{k}$ is the unique interior maximum of $\Gamma(k,p)$ in the interval $k \in [0,N_y]$.

Proof. First we observe that $\Gamma(k,p) \geq 0$, $\forall k,p$ and $\Gamma(k,p) = 0$ if and only if $k = \frac{2\arcsin(\frac{hp}{2})}{h\pi}$ or $k = 0$. Second, we compute the derivative of $\Gamma(k,p)$ with respect to p,

$$\operatorname{sign}\left(\frac{\partial\Gamma(k,p)}{\partial p}\right) = \operatorname{sign}(hp - 2s_k).$$

Therefore, at the optimum, p must lie inside the interval $[0, \frac{2}{h}s_{N_y}]$. We then look for the maximum with respect to k. We have that $\frac{\partial\Gamma(k,p)}{\partial k} = 0$ if and only if $k_1 = \frac{2\arcsin(\frac{hp}{2})}{h\pi}$ which therefore is a minimum and zero, and for $k_2 = 0$, $k_3 = \frac{1}{h}$ and $\widetilde{k} = \frac{2\arcsin(\frac{1}{2}(\sqrt{2}-1)ph)}{\pi h} < k_1$. We can conclude that the function for $p \in [0, \frac{2s_{N_y}}{h}]$ starts from zero at $k = 0$, it increases until it reaches an interior maximum at $\widetilde{k}$, then it decreases until the zero k_1 and then it is strictly increasing until k_3. We observe that $k_3 = \frac{1}{h} > N_y$, therefore the function has two local maxima, one located at $\widetilde{k}$ and the other one at $k = N_y$. Moreover varying $p \in [0, s_{N_y}]$, the zero $k_1(p)$ is mapped into the interval $[0,N_y]$. Suppose now that $\Gamma(\widetilde{k},p) > \Gamma(N_y,p)$, the other case is treated similarly. Since $\operatorname{sign}(\partial_p\Gamma) = \operatorname{sign}(k_1 - k)$, we have $\partial_p\Gamma(\widetilde{k},p) > 0$ and $\partial_p\Gamma(N_y,p) < 0$, therefore increasing p decreases the maximum of $\Gamma(k,p)$ until $\Gamma(\widetilde{k},p^*) = \Gamma(N_y,p^*)$. This is the optimal solution since varying the parameter p would increase the value of Γ either at $k = \widetilde{k}$ or $k = N_y$. The uniqueness follows from the strict monotonicity. Finally, solving the equation $\Gamma(\widetilde{k},p^*) = \Gamma(N_y,p^*)$ we get the expression for p^*. □

Theorem 3.2.4 (Mesh independent convergence)**.** *Assuming that $n_1 + n_2 = 2$ and choosing p as in Theorem 3.2.3, the spectral radius of the two-level OSM iteration matrix T is bounded below 1 uniformly with respect to h,*

$$\rho(T(p^*)) \leq C < 1 \quad \text{as } h \to 0, \quad \text{with} \quad C = 0.0520. \tag{3.2.17}$$

Proof. Based on Lemma 3.2.2, we have

$$\rho(T(p)) \leq \|T(p)\|_1 \leq \max_{k\in[0,N_y]} \Gamma(k,p).$$

Taking the minimum with respect to p, the inequality still holds, thus

$$\min_p \rho(T(p)) \leq \min_p \max_{k\in[0,N_y]} \Gamma(k,p).$$

We denote with p^* the solution of the min-max problem studied in Theorem 3.2.3. Clearly there is no reason why p^* would still be the solution of the min-max problem $\min_p \rho(T(p))$. Nevertheless we have that

$$\min_p \rho(T(p)) \leq \rho(T(p^*)) \leq \Gamma(N_y,p^*) = \min_p \max_k \Gamma(k,p).$$

Substituting the expression of p^* we get that

$$\Gamma(N_y,p^*) = \frac{3(\sqrt{2}+3-\sqrt{6}-\sqrt{3})^2 \sin\left(\frac{hN_y\pi}{2}\right)^2}{(\sqrt{2}+1-\sqrt{6}-\sqrt{3})^2}.$$

We now observe that $\Gamma(N_y,p^*)$ is a strictly decreasing function of h, therefore it has its maximum for $h \to 0$. Computing the limit $\lim_{h\to 0} \Gamma(N_y,p^*) = 0.0520 =: C$. Hence we conclude that $\min_p \rho(T(p)) \leq \rho(T(p^*)) \leq C < 1$, as $h \to 0$. □

Remark 3.2.5. The asymptotic performance of the one-level OSM has been the subject of intensive study. For straight interfaces, in [74] it has been shown that for zero order transmission conditions the spectral radius is bounded from above by $1 - O(h^{\frac{1}{2}})$ in a nonoverlapping decomposition and by $1-O(h^{\frac{1}{3}})$ in the overlapping case with overlap proportional to the mesh size. See also [130] for a generalization to arbitrary interfaces. For second order transmission conditions [74], we have respectively $1 - O(h^{\frac{1}{4}})$ and $1 - O(h^{\frac{1}{5}})$. Theorem 3.2.4 shows that the two-level OSM gains the same property of the multigrid scheme with a convergence independent of the mesh size because of the presence of the coarse correction. We emphasize that the same conclusion holds if one uses the classical parallel Schwarz method instead of OSM as smoother.

3.2.2 How to choose the optimized parameter in the nonoverlapping case

As we emphasized in the proof of Theorem 3.2.4, in general p^* is not a solution of the minimization problem $\min_p \rho(T(p))$. Thus we study numerically the behaviour of the spectral radius and of the other bounds as functions of p. On the left of Figure 3.1, we plot the behaviour of different quantities as p varies. From the right panel, we observe that the solutions of none of the min-max problems involving the different bounds or even $\rho(T(p))$ provides an optimized convergence. The reasons of this discrepancy lie in the several simplifications used in the literature for the derivation of the convergence factors for the one-level OSMs which is mainly based on a continuous analysis. It therefore neglects the computation of the discrete derivative and it approximates the eigenvalues of

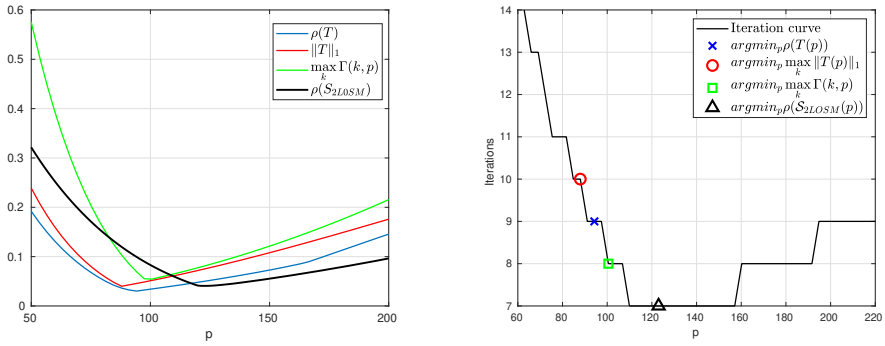


Figure 3.1: On the left comparison between the spectral radius of S_{2LOSM} and T and various upper bounds. On the right, number of iterations required to reach convergence as function of p and comparison between the predicted p obtained by solving different min-max problems involving the quantities presented in the left panel. The fine mesh corresponds to $\ell = 6$.

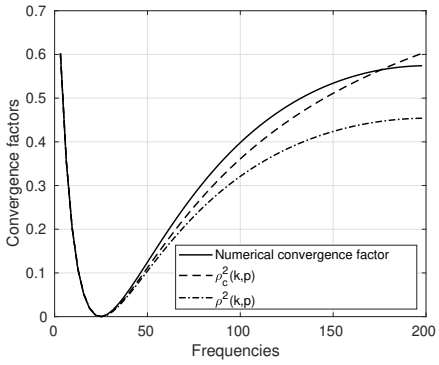


Figure 3.2: The continous line corresponds to the numerical convergence factor, the dashed line corresponds to $\rho_c^2(k,p)$ and the dash-dotted line to $\rho^2(k,p)$.

the discrete Laplacian with the ones of the continuous Laplacian[1]. In our analysis, we indeed take into account the eigenvalues of the discrete Laplacian, but we did not include the discrete derivative. We show that our small theoretical improvement actually worsens the approximation of the numerical convergence factor in the high frequencies regime. In Figure 3.2, we plot, for a fixed p, the numerical convergence factor, $\rho^2(k,p)$ and also the continuous analogue of $\rho^2(k,p)$, i.e. $\rho_c^2(k,p) = \left(\frac{\pi k - p}{\pi k + p}\right)^2$, which involves the continuous eigenvalues of the Laplace operator in 1D. It is evident that actually $\rho_c^2(k,p)$ is a better approximation of the numerical convergence factor. On the other hand, $\rho^2(k,p)$ is wrongly faster for high frequencies and that is why our estimates for p are constantly lower than

[1]We remind that usually the unbounded hypothesis is also made, but it has no significant impact in this case.

the optimal ones.

Guided by these observations we now consider an analysis which is less precise from the theoretical point of view than the one proposed in Section 3.1, but that will provide a better approximation of the optimal parameter p. We carry out a complete continuous analysis by replacing the expansions (3.2.1) with

$$u_1 = \sum_{k=1}^{N_y} A(k)\psi_k e^{\pi k x} \quad \text{and} \quad u_2 = \sum_{k=1}^{N_y} B(k)\psi_k e^{-\pi k x}. \tag{3.2.18}$$

We insert this anstanz into the iterative method and when dealing with the restriction and prolongation operators we assume[2] that the same results as in the discrete case hold, see for instance eq (3.2.3). Repeating the same calculations we obtain the recurrence relation $\mathbf{v}_k^n = \widetilde{D}_k^c \mathbf{v}_k^{n-1}$ similar to (3.2.8) where

$$\widetilde{D}_k^c := \begin{pmatrix} \left(1-c_k^4\right) & 0 & \frac{c_k^2 s_k^2}{2}\left(1+\frac{\widetilde{k}}{k}\right) & \frac{c_k^2 s_k^2}{2}\left(\frac{\widetilde{k}}{k}-1\right) \\ 0 & \left(1-c_k^4\right) & \frac{c_k^2 s_k^2}{2}\left(\frac{\widetilde{k}}{k}-1\right) & \frac{c_k^2 s_k^2}{2}\left(1+\frac{\widetilde{k}}{k}\right) \\ c_k^2 s_k^2 & 0 & 1-\frac{s_k^4\left(1+\frac{\widetilde{k}}{k}\right)}{2} & \frac{s_k^4\left(1-\frac{\widetilde{k}}{k}\right)}{2} \\ 0 & c_k^2 s_k^2 & \frac{s_k^4\left(1-\frac{\widetilde{k}}{k}\right)}{2} & 1-\frac{s_k^4\left(1+\frac{\widetilde{k}}{k}\right)}{2} \end{pmatrix}. \tag{3.2.19}$$

Defining $G_k^{c,n} = \mathrm{diag}(\rho_c^n(k,p), \rho_c^n(k,p), \rho_c^n(\widetilde{k},p), \rho_c^n(\widetilde{k},p))$, $\overline{D}_k := \widetilde{D}_k^c G_k^{c,n_1}$ and recalling $\mathbf{e}^n := \left(\mathbf{v}_1^n, ..., \mathbf{v}_{N_c}^n, A\left(\frac{N+1}{2}\right), B\left(\frac{N+1}{2}\right)\right)$, we conclude that $\mathbf{e}^n = \overline{T}\mathbf{e}^{n-1}$, where

$$\overline{T} = \begin{pmatrix} \overline{D}_1 & & & & & & \\ & \overline{D}_2 & & & & & \\ & & \ddots & & & & \\ & & & \overline{D}_{N_y^c} & & & \\ & & & & \rho_c^n\left(\frac{N_y+1}{2}, p\right) & & \\ & & & & & \rho_c^n\left(\frac{N_y+1}{2}, p\right) & \end{pmatrix}.$$

Lemma 3.2.6. *Defining* $\overline{\Gamma}(k,p) := 3\rho_c^n(k,p)s_k^2$, *we have*

$$\rho(\overline{T}) \leq \|\overline{T}\|_1 \leq \max_{k\in[1,N_y]} \overline{\Gamma}(k,p).$$

Proof. The proof follows the step of the proof of Lemma 3.2.2. Direct calculations show that

$$\|\overline{T}\|_1 = \max\left\{\max_{k\in\mathcal{V}} \rho_c^n(k,p)s_k^2(1+2c_k^2), \max_{k\in\mathcal{V}} \rho_c^n(\widetilde{k},p)c_k^2\left(1+s_k^2\frac{Ny+1}{k}\right), \rho_c^n\left(\frac{N_y+1}{2},p\right)\right\}.$$

[2] It is a slight abuse of notation since under our hypothesis on the mesh, the eigenvectors of the discrete Laplacian correspond to the discretization on the mesh points of the eigenvectors of the continuous Laplacian.

Studying the second term in the brackets we conclude that $1 + s_k^2 \frac{Ny+1}{k} \leq 2$ for $k \in \mathcal{V}$, so that we can consider the upper bound

$$\|\overline{T}\|_1 \leq \max\left\{\max_{k\in\mathcal{V}} 3\rho_c^n(k,p)s_k^2, \max_{k\in\mathcal{V}} 3\rho_c^n(\widetilde{k},p)c_k^2, \rho_c^n\left(\frac{N_y+1}{2},p\right)\right\}.$$

Similarly to Lemma 3.2.2, we introduce the set $\mathcal{Z} := \left\{\frac{Ny+1}{2}+1,\ldots,Ny\right\}$, so that $\|T\|_1 \leq \max\left\{\max_{k\in\mathcal{V}\cup\mathcal{Z}} 3\rho_c^n(k,p)s_k^2, \rho_c^n\left(\frac{N_y+1}{2},p\right)\right\}$.

Finally observing that $3\rho_c^n(\frac{Ny+1}{2},p)s^2_{\frac{Ny+1}{2}} \geq \rho_c^n\left(\frac{N_y+1}{2},p\right)$ and considering a continous set of frequencies $k \in [1,N_y]$, we get the desired bound. □

We are ready to prove the following theorem.

Theorem 3.2.7. *Assuming $n = 2$, the solution of the min-max problem*

$$\min_p \max_{k\in[0,N_y]} \overline{\Gamma}(k,p) \tag{3.2.20}$$

is given by $\overline{p}$ which is the unique solution of the nonlinear equation

$$\overline{\Gamma}(\widehat{k},p) = \overline{\Gamma}(N,p), \tag{3.2.21}$$

where $\widehat{k}$ is the unique interior maximum of $\overline{\Gamma}(k,p)$.

Proof. The function has two zeros, one located at $k = 0$, the other at $k = \frac{p}{\pi}$. Analyzing the sign of the derivative with respect to p we obtain that

$$\operatorname{sign}\left(\frac{\partial\overline{\Gamma}}{\partial k}\right) = \operatorname{sign}(p - k\pi),$$

and we conclude that at the optimum $p \in [0, N_y\pi]$. The derivative with respect to k is given by

$$\frac{\partial\overline{\Gamma}}{\partial k} = 3\,\frac{(\pi k - p)\sin\left(\frac{1}{2}hk\pi\right)\pi\left(\cos\left(\frac{1}{2}hk\pi\right)hk^2\pi^2 - \cos\left(\frac{1}{2}hk\pi\right)hp^2 + 4\sin\left(\frac{1}{2}hk\pi\right)p\right)}{(\pi k + p)^3}.$$

Therefore the stationary points are located at $k = 0$, $k = \frac{p}{\pi}$, $k = \frac{1}{h}$, which is actually outside the interval $[0, N_y]$ and at $k = \widehat{k}$, which is the unique root of the equation $\cos\left(\frac{1}{2}hk\pi\right)hk^2\pi^2 - \cos\left(\frac{1}{2}hk\pi\right)hp^2 + 4\sin\left(\frac{1}{2}hk\pi\right) = 0$. Indeed dividing by $\cos\left(\frac{1}{2}hk\pi\right) \neq 0,\ \forall k \in [0, N_y]$, we get the equation

$$hk^2\pi - hp^2 + 4\tan\left(\frac{1}{2}\pi hk\right)p = 0,$$

which is a strictly increasing function of k which for $k = 0$ is negative and for $k = N_y$ is positive. Moreover we have that $\widehat{k} \leq \frac{p}{\pi}$. The function therefore has the following behaviour:

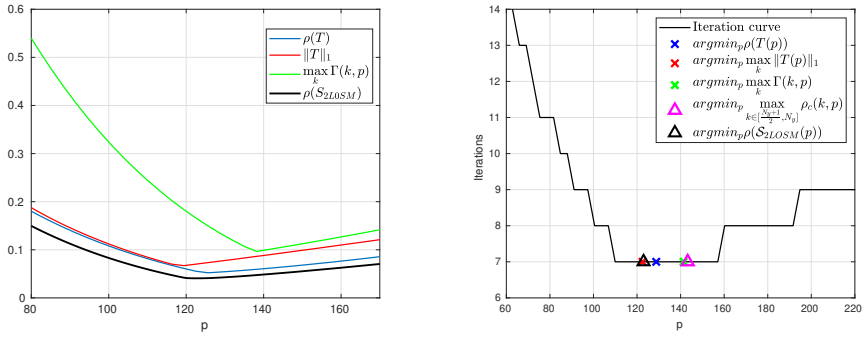


Figure 3.3: On the left comparison between the spectral radius of S_{2LOSM} and of $\overline{T}$ and various upper bounds. On the right, number of iterations required to reach convergence for different values of p obtained by solving different min-max problems involving the quantities presented in the left panel. We also add a magenta triangle which represents the solution of $\min_p \max_{k\in[\frac{N_y+1}{2},N_y]} \rho_c^2(k,p)$.

it starts from 0 at $k=0$ and it is strictly increasing until it reaches its local maximum at $k=\widehat{k}$. Then it decreases and it reaches zero at $k=\frac{p}{\pi}$ and eventually it increases until the local maximum located on the boundary at $k=N_y$. Using the classical arguments of Theorem 3.2.3 we conclude that the solution is indeed given by equioscillation between the two local maxima. □

In conclusion we show in Figure 3.3 a comparison of the different optimized parameters that can be obtained minimizing the spectral radius, the 1-norm or the upper bound $\overline{\Gamma}(k,p)$. We see that the unique solution of equation (3.2.21) leads to an optimized convergence.

Remark 3.2.8. In a one-level setting, one chooses the optimized parameter solving the min-max problem

$$\min_p \max_{[1,N_y]} \rho_c^2(k,p). \tag{3.2.22}$$

In the case without overlap, the parameter p solution of (3.2.22) does not lead to a smoother, since it tries to balance the convergence factor for low and high frequencies. However, similarly to the Jacobi smoother in a multigrid setting [109], one can choose p such that the OSM eliminates the high frequencies, see Figure 3.4, while the low ones are corrected on the coarse mesh. One obvious choice would then be to solve the min-max problem

$$\min_p \max_{k\in[\frac{N_y+1}{2},N_y]} \rho_c^2(k,p). \tag{3.2.23}$$

Figure 3.3 shows that this heuristic idea indeed leads to an excellent optimized parameter so that, instead of the min-max problems involving the new quantites $\Gamma,\overline{\Gamma}$, one could just use the same min-max solution involving the one-level convergence factor $\rho_c(k,p)$

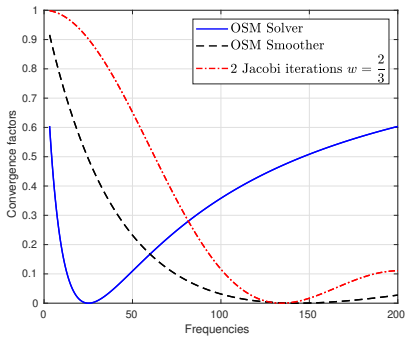


Figure 3.4: Comparison of the smoothing property among the Jacobi method with damping parameter $w = \frac{2}{3}$, the OSM tuned as a solver, and the OSM tuned as a smoother.

by changing the interval for the variable k. The analytical solution of (3.2.23) is given by $p^* = \sqrt{\frac{N_y+1}{2} N_y} \approx Ch^{-1}$ as $h \to 0$, so that the asymptotic behaviour of the optimized parameter obtained from (3.2.23) and (3.2.16) is the same.

3.3 Two-level OSM analysis for an overlapping decomposition

In this section, we present an analogous analysis for the overlapping two-level OSM. Given two initial guesses u_1^0, u_2^0 and an overlapping decomposition of Ω into Ω_1, Ω_2 with $\Gamma_j := \partial\Omega_j \setminus \partial\Omega$, $j = 1,2$, the one-level parallel overlapping OSM reads for $n \geq 1$

$$\begin{aligned} \mathscr{L} u_1^n &= f \quad \text{in} \quad \Omega_1, \quad \partial_x u_1^n + p u_1^n = \partial_x u_2^{n-1} + p u_2^{n-1} \quad \text{on } \Gamma_1, \\ \mathscr{L} u_2^n &= f \quad \text{in} \quad \Omega_2, \quad -\partial_x u_2^n + p u_2^n = -\partial_x u_1^{n-1} + p u_1^{n-1} \quad \text{on } \Gamma_2. \end{aligned} \tag{3.3.1}$$

Defining two functions on the interfaces as

$$r_1 := -\partial_x u_1^n - p u_1^n + \partial_x u_2^n + p u_2^n \quad \text{on } \Gamma_1 \quad \text{and} \quad r_2 := -\partial_x u_1^n + p u_1^n + \partial_x u_2^n - p u_2^n \quad \text{on } \Gamma_2, \tag{3.3.2}$$

and then solving

$$\begin{aligned} \mathscr{L} e_1 &= 0 \quad \text{in} \quad \Omega_1, \quad \partial_x e_1 + p e_1 - \partial_x e_2 - p e_2 = r_1 \quad \text{on } \Gamma_1, \\ \mathscr{L} e_2 &= 0 \quad \text{in} \quad \Omega_2, \quad -\partial_x e_2 + p e_2 + \partial_x e_1 - p e_1 = r_2 \quad \text{on } \Gamma_2, \end{aligned} \tag{3.3.3}$$

we have that $\widetilde{u}_1 := u_1^n + e_1$ and $\widetilde{u}_2 := u_2^n + e_2$ are solution of problem (3.1.1) in the sense that $\widetilde{u}_1 = u|_{\Omega_1}$ and $\widetilde{u}_2 = u|_{\Omega_2}$.

To analyze the method, we suppose that the two subdomains are $\Omega_1 = (-\frac{1}{2}, a) \times (0,1)$ and $\Omega_2 = (-a, \frac{1}{2}) \times (0,1)$, with two interfaces $\Gamma_1 = [a] \times [0,1]$ and $\Gamma_2 = [-a] \times [0,1]$. Inserting the expansions (3.2.18) into the residual definition (3.3.2) we get

$$\begin{aligned} r_1 &= \sum_{k=1}^{N_y} g_-(k) A(k) \rho_o^{n_1}(k,p) \psi_k e^{k\pi a} + \sum_{k=1}^{N_y} g_+(k) B(k) \rho_o^{n_1}(k,p) \psi_k e^{-k\pi a} \text{ on } \Gamma_1, \\ r_2 &= \sum_{k=1}^{N_y} g_+(k) A(k) \rho_o^{n_1}(k,p) \psi_k e^{-k\pi a} + \sum_{k=1}^{N_y} g_-(k) B(k) \rho_o^{n_1}(k,p) \psi_k e^{k\pi a} \text{ on } \Gamma_2, \end{aligned} \tag{3.3.4}$$

where $\rho_o(k,p) := \left(\frac{\pi k - p}{\pi k + p}\right) e^{-2a\pi k}$, see [74] for a derivation. Solving the corresponding version of eq (3.2.6), we obtain

$$\overline{A}(k) = \frac{R_{1D}r_1(k)+R_{1D}r_2(k)}{2F(k)} + \frac{R_{1D}r_1(k)-R_{1D}r_2(k)}{2G(k)},$$
$$\overline{B}(k) = \frac{R_{1D}r_1(k)+R_{1D}r_2(k)}{2F(k)} + \frac{R_{1D}r_2(k)-R_{1D}r_1(k)}{2G(k)},$$

where $F(k) := 2k\pi\cosh(\pi ka)+2p\sinh(\pi ka)$, $G(k) := 2k\pi\sinh(\pi ka)+2p\cosh(\pi ka)$. Then computing the updated approximations $u_1^{n_1} + Ie_1$ and $u_2^{n_1} + Ie_2$ we obtain that the vector $\mathbf{v}_k^n = \left(A^n(k), B^n(k), A^n(\widetilde{k}), B^n(\widetilde{k})\right)^t$ satisfies the recurrence relation $\mathbf{v}_k^n = \widetilde{D}_k^O \mathbf{v}_k^{n-1}$, with

$$\widetilde{D}_k^O := \begin{pmatrix} \rho_o^{n_1}\left(1-c_k^4\right) & 0 & \frac{\widetilde{\rho}_o^{n_1} c_k^2 s_k^2}{2}\left(\frac{\widetilde{F}}{F}+\frac{\widetilde{G}}{G}\right) & \frac{\widetilde{\rho}_o^{n_1} c_k^2 s_k^2}{2}\left(\frac{\widetilde{F}}{F}-\frac{\widetilde{G}}{G}\right) \\ 0 & \rho_o^{n_1}\left(1-c_k^4\right) & \frac{\widetilde{\rho}_o^{n_1} c_k^2 s_k^2}{2}\left(\frac{\widetilde{F}}{F}-\frac{\widetilde{G}}{G}\right) & \frac{\widetilde{\rho}_o^{n_1} c_k^2 s_k^2}{2}\left(\frac{\widetilde{F}}{F}+\frac{\widetilde{G}}{G}\right) \\ \rho_o^{n_1} c_k^2 s_k^2 & 0 & \widetilde{\rho}_o^{n_1} - \frac{\widetilde{\rho}_o^{n_1} s_k^4\left(\frac{\widetilde{F}}{F}+\frac{\widetilde{G}}{G}\right)}{2} & \frac{\widetilde{\rho}_o^{n_1} s_k^4\left(\frac{\widetilde{F}}{F}-\frac{\widetilde{G}}{G}\right)}{2} \\ 0 & \rho_o^{n_1} c_k^2 s_k^2 & \frac{\widetilde{\rho}_o^{n_1} s_k^4\left(\frac{\widetilde{F}}{F}-\frac{\widetilde{G}}{G}\right)}{2} & \widetilde{\rho}_o^{n_1} - \frac{\widetilde{\rho}_o^{n_1} 2 s_k^4\left(\frac{\widetilde{F}}{F}+\frac{\widetilde{G}}{G}\right)}{2} \end{pmatrix}, \tag{3.3.5}$$

where $F := F(k,p)$, $\widetilde{F} := F(\widetilde{k},p)$ and similarly for G and $\widetilde{G}$.

Remark 3.3.1. We note that the same calculations can be adapted to obtain an iteration matrix for a two-level method which uses the parallel Schwarz method as a smoother. We need to replace Equation (3.3.1) with the classical parallel Schwarz method, the residuals are $r_1 = -u_1^n + u_2^n = -r_2$ and in the residual problem (3.3.3) we impose $e_1 - e_2 = r_1$ on Γ_1 and $e_2 - e_1 = r_2$ on Γ_2. Finally, we use the properties of the interpolation and restriction operators and the convergence factor $\rho_{PSM}(k) := e^{-2a\pi k}$.

Computing the 1-norm of $\widetilde{D}_k^O$ is delicate because the sign of the terms $\frac{\widetilde{F}}{F} - \frac{\widetilde{G}}{G}$ depends on p, and therefore many possible cases arise. Therefore assuming $n_1 = 2$, we look for a proxy quantity to analyze, and inspired by Section 3.2 we define $\overline{\Gamma}_{over}(k,p) := 3s_k\rho_o^2$. We consider the problem analogous to (3.2.20) for the overlapping case.

Theorem 3.3.2. *The solution of the min-max problem*

$$\min_p \max_{k\in[0,+\infty]} \overline{\Gamma}_{over}(k,p) \tag{3.3.6}$$

is given by $\overline{p}$ which is the unique solution of the nonlinear equation

$$\overline{\Gamma}_{over}(\widehat{k},p) = \overline{\Gamma}_{over}(\widetilde{k},p),$$

where $\widehat{k}$ and $\widetilde{k}$ are the interior maxima of $\overline{\Gamma}_{over}(k,p)$ for $k \in [0,\infty]$.

Proof. We first observe that $\overline{\Gamma}_{over}(k,p) \geq 0, \forall k,p$ and $\overline{\Gamma}_{over}(k,p) = 0$ if and only if $k = 0$ or $k = \frac{p}{\pi}$. The sign of the derivative of $\overline{\Gamma}_{over}$ with respect to p is

$$\operatorname{sign}\left(\frac{\partial\overline{\Gamma}_{over}(k,p)}{\partial p}\right) = \operatorname{sign}(p - k\pi),$$

therefore we conclude that at the optimum $p \geq 0$. The zeros of the derivative with respect to k are located at $k = 0, k = \frac{p}{\pi}$ and at the only two zeros $\widetilde{k}, \widehat{k}$ of the non linear equation

$$\tan\left(\frac{hk\pi}{2}\right) = \frac{h\pi(p^2 - k^2\pi^2)}{4p\pi + 2\delta p^2 - 2\delta k^2\pi^2}. \tag{3.3.7}$$

Indeed $g(k) := \tan\left(\frac{hk\pi}{2}\right)$ is a strictly increasing function in k, which is equal to zero for $k = 0$ and goes to infinity as $k \to +\infty$. The function $l(k,p,\delta) := \frac{h\pi(p^2-k^2\pi^2)}{4p\pi+2\delta p^2-2\delta k^2\pi^2}$ is positive for $k = 0$, it is stricly decreasing for every k and equal to zero at $k = \frac{p}{\pi}$. Therefore there exists a $\widehat{k}$ in $[0, \frac{p}{\pi}]$ solution of (3.3.7). On the other hand $l(k,p,\delta)$ has a vertical asymptote at $k_1 = \frac{\sqrt{\delta p(\delta p+2\pi)}}{\delta\pi} > \frac{p}{\pi}$ and we have $\lim_{k\to k_1^+} l(k,p,\delta) = +\infty$ and $\lim_{k\to+\infty} l(k,p,\delta) = \frac{\pi h}{2\delta}$. Therefore we conclude that there exists a $\widetilde{k} > \frac{p}{\pi}$ solution of (3.3.7). $\overline{\Gamma}_{over}(k,p)$ has therefore two local maxima $\widehat{k} \leq \frac{p}{\pi} \leq \widetilde{k}$, and repeating the final argument of Theorem 3.2.3 we obtain that the solution of (3.3.6) is given by equioscillation. □

3.3.1 How to choose the optimized parameter in the overlapping case

The nonoverlapping OSM is not a natural smoother and therefore the tuning of the transmission conditions is essential to have an efficient two-level method. In constrast, the overlapping OSM is a perfect smoother since it is exponentially fast for high frequencies and thus we expect the tuning to be less important. Nevertheless, we want to study how close the solution of the optimization problem (3.3.6) involving $\overline{\Gamma}_{over}(k,p)$ is to the solution of $\min_p \rho(S_{2LOSM})$. On the left panel of Figure 3.5 we plot the behavior of the iteration matrix (3.1.6) and of $\overline{\Gamma}_{over}(k,p)$ as function of p. We denote with S_{2LOSM} the iteration matrix where we use the linear interpolator I_ℓ and with S^A_{2LOSM} the one which uses I^A_ℓ. Concerning the choice of the optimized parameter, we deduce from Figure 3.5 that if we use the harmonic extension operator, then Theorem 3.3.6 provides a perfect choice for the optimized parameter. However, we observe that there is a significant difference in the spectral properties of S_{2LOSM} and S^A_{2LOSM}. The explanation for this behavior lies in the different choice of the interpolation operator: if we use the linear interpolator I_ℓ, the corrections which we add to the iterates are not harmonic anymore. Hence, we cannot assume that the expansions (3.2.18) hold in the interior of the subdomains and especially on Γ_j, $j = 1,2$, where the smoother S_ℓ acts. In Fig. 3.6 we plot the first eigenvector of the iteration matrices S_{2LOSM} and S^A_{2LOSM} in the overlapping case. We can clearly observe that the eigenvector of S_{2LOSM} does not behave as an exponential along the x direction, as required by expansion (3.2.18).

On the right panel of Fig. 3.5, we plot the spectral radius of S_{2LOSM} and S^A_{2LOSM} in the nonoverlapping case where the discrepancy is negligible, the $\min_p S_{2LOSM}$ being attained at $p \approx 124$ and $\min_p S^A_{2LOSM}$ at $p \approx 118$. Hence, in the nonoverlapping case the use of I_ℓ or of I^A_ℓ does not influence the method significantly. We remark that using I_ℓ, the correction we add is not harmonic on the fine grid, but the nonoverlapping smoother takes values next to the interface and thus is less affected by the non harmonicity inside the domain.

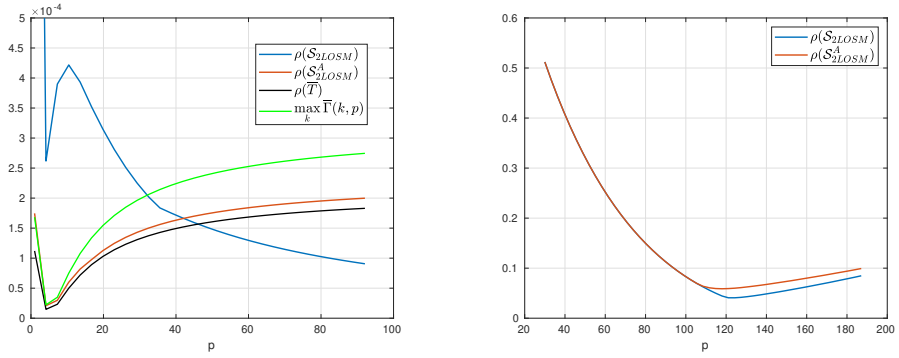


Figure 3.5: On the left, we plot the behavior of S_{2LOSM}, S^A_{2LOSM}, $\rho(\overline{T})$ and $\max_k \overline{\Gamma}_{over}(k,p)$ with respect to p. We remark that S_{2LOSM} does not behave as our analysis predicts. On the right we plot S_{2LOSM}, S^A_{2LOSM} in the nonoverlapping case in which the discrepancy is negligible. The fine mesh corresponds to $\ell = 6$.

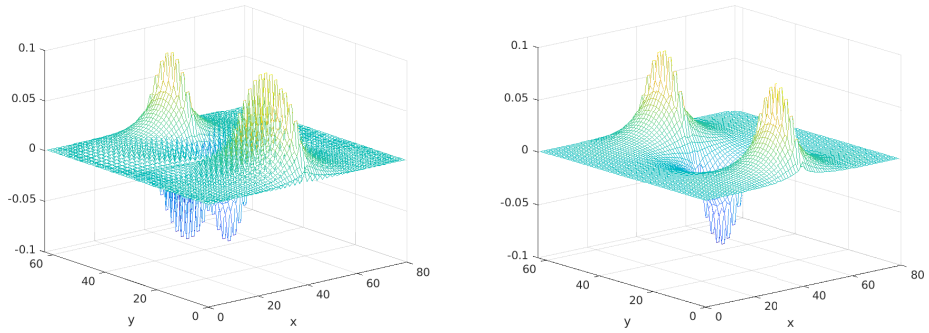


Figure 3.6: First eigenvector of S_{2LOSM} on the left and of S^A_{2LOSM} on the right.

3.4 Multilevel generalization

In this section we generalize the two-level Algorithm 1 to a multilevel setting. As emphasized in the multigrid literature [109, 152], the coarse problem $\widetilde{A}_{\ell-1}\widetilde{\mathbf{e}}_{\ell-1} = R_\ell \mathbf{r}^{n_1}_\ell$, may still be too large and therefore one could use recursively a two-level method to solve it. However in the nonoverlapping case, the smoothing property of the OSM depends strongly on the transmission conditions and therefore, moving from one grid to another, they need to be tuned according to the new mesh and this implies also that the residuals must be properly modified.

Suppose that at the continuous level we do some smoothing steps of the double-sided OSM, see [74], with free parameters p and q. According to equation (3.1.3), the residual will be zero inside the domain and on the interface Γ we have the two functions

$$r_1 := -\partial_x u_1^n - p u_1^n + \partial_x u_2^n + p u_2^n \quad \text{and} \quad r_2 := -\partial_x u_1^n + q u_1^n + \partial_x u_2^n - q u_2^n. \tag{3.4.1}$$

Suppose now that we want to change the parameters in the transmission conditions to a new couple (p_c, q_c). We are thus interested in the system

$$\begin{array}{llll} \mathscr{L} e_{1,c} = 0 & \text{in} & \Omega_1, & \partial_x e_1 + p_c e_1 - \partial_x e_2 - p_c e_2 = r_{1,c} \quad \text{on } \Gamma, \\ \mathscr{L} e_{1,2} = 0 & \text{in} & \Omega_2, & -\partial_x e_2 + q_c e_2 + \partial_x e_1 - q_c e_1 = r_{2,c} \quad \text{on } \Gamma, \end{array} \tag{3.4.2}$$

for some choice of $r_{1,c}$ and $r_{2,c}$. We would like to choose $r_{1,c}$ and $r_{2,c}$ such that the solutions of (3.4.2) and of (3.1.4) are identical. In this way the discrete solution of (3.4.2) with new parameters p_c, q_c on a coarse mesh would lead to a good coarse correction.

Therefore we look at the expression of $r_{1,c}$ and $r_{2,c}$ such that $e_{1,c} = e_1$ and $e_{2,c} = e_2$. We observe that the transmission conditions in (3.4.1) with the two parameters p and q imply

$$(e_1 - e_2) = \frac{r_1 - r_2}{p+q}, \quad (\partial_x e_1 - \partial_x e_2) = \frac{q r_1 + p r_2}{p+q}.$$

Thus $r_{1,c}$ and $r_{2,c}$ should be such that

$$\begin{aligned} e_{1,c} - e_{2,c} &= \frac{r_{1,c} - r_{2,c}}{p_c + q_c} = \frac{r_1 - r_2}{p+q} = e_1 - e_2, \\ \partial_x e_{1,c} - \partial_x e_{2,c} &= \frac{p_c r_{2,c} + q_c r_{1,c}}{p_c + q_c} = \frac{p r_2 + q r_1}{p+q} = \partial_x e_1 - \partial_x e_2, \end{aligned} \tag{3.4.3}$$

so that $e_{j,c} \equiv e_j$, $j = 1,2$ since they satisfy the same PDE in the interior of the subdomains and the same boundary conditions on the interface. Direct calculations from (3.4.3) lead to

$$r_{1,c} := r_{2,c} + \frac{p_c + q_c}{p+q}(r_1 - r_2), \quad r_{2,c} := r_1 \frac{q - q_c}{p+q} + r_2 \frac{p + q_c}{p+q}. \tag{3.4.4}$$

Moving to a discrete setting, we define $\mathbf{r}_\ell$ as the residual on the fine grid computed with parameters p_ℓ, q_ℓ and with $\mathbf{r}_{l,c}$ the modified residual where the role of p_c and q_c in eq (3.4.4) is now played by $p_{\ell-1}, q_{\ell-1}$, i.e. the smoothing parameters we want to use on the coarse grid. We call $\mathscr{G}$ the operator which takes $\mathbf{r}_\ell$ and returns the modified residual according to eq (3.4.4), i.e $\mathbf{r}_{l,c} = \mathscr{G}(\mathbf{r}_l, p_l, q_l, p_{l-1}, q_{l-1})$. Thanks to these observations, the multilevel optimized Schwarz method to solve the linear system $\widetilde{A}_{l_{\max}} \widetilde{\mathbf{u}}_{l_{\max}} = \mathbf{f}_{l_{\max}}$ consists in multiple calls of the MOSM function described by Algoritm 2 until convergence is reached. In the overlapping case, the smoothing property of the OSM is guaranteed by the overlap and so there is no need to tune the parameters p_ℓ and q_ℓ on each mesh. We can always use the parameters solution of (3.3.6) without losing in efficiency. Therefore, in the overlapping case, we just consider $\mathscr{G}$ as the identity operator. According to the value of γ in Algorithm 2 we obtain a V-cycle ($\gamma = 1$) or W-cycle ($\gamma = 2$). In the numerical section we consider only the V-cycle, since the W-cycle shows a similar behaviour.

3.5 Numerical results

Every experiment starts with a random initial guess with values between -1 and 1, and the right hand side is equal to $f = 1$. We use the acronyms OSMo(p) and OSM(p) to indicate a one-level OSM with a single sided optimized parameter p with and without overlap.

Algorithm 2: Function MOSM($\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^0, \widetilde{\mathbf{f}}_\ell$)

- If $\ell = \ell_{min}$, then return $\widetilde{\mathbf{u}}_{\ell\min} \leftarrow \widetilde{A}_{\ell_{\min}}^{-1} \mathbf{f}_{\ell_{\min}}$.
- For $n = 1 : n_1$, $\widetilde{\mathbf{u}}_\ell^n \leftarrow S_\ell(\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^{n-1}, \widetilde{\mathbf{f}}_\ell)$.
- $\mathbf{r}_{l,c} \leftarrow \mathcal{G}(\widetilde{\mathbf{f}}_\ell - \widetilde{A}_\ell \widetilde{\mathbf{u}}_\ell^{n_1}, p_\ell, q_\ell, p_{\ell-1}, q_{\ell-1})$.
- Set $\widetilde{\mathbf{e}}_{\ell-1} = 0$.
- Call γ times $\widetilde{\mathbf{e}}_{l-1} \leftarrow \text{MOSM}(\widetilde{A}_{\ell-1}, \widetilde{\mathbf{e}}_{l-1}, R_l \mathbf{r}_{l,c})$.
- $\widetilde{\mathbf{u}}_\ell^{n_1} \leftarrow \widetilde{\mathbf{u}}_\ell^{n_1} + I_\ell \widetilde{\mathbf{e}}_{l-1}$.
- For $n = n_1 + 1 : n_2$, $\widetilde{\mathbf{u}}_\ell^n \leftarrow S_\ell(\widetilde{A}_\ell, \widetilde{\mathbf{u}}_\ell^{n-1}, \widetilde{\mathbf{f}}_\ell)$.
- Return $\widetilde{\mathbf{u}}_\ell^{n_2}$.

OSMV(p,q) indicates a V-cycle OSM with two optimized parameters p, q. The optimized parameters are obtained by maximizing the smoothing property of the OSM scheme according to the Remark 3.2.8. MGV stands for a multigrid V-cycle with a Jacobi smoother with damping parameter $w = \frac{2}{3}$. The number of pre- and post-smoothing steps is set equal to $n_1 = n_2 = 2$ on each level, except on the coarsest one where the linear system is solved directly, for all the multilevel schemes, i.e. for MGV, OSMV(p), OSMV(p,q) and OSMoV(p). For each equation we compute the exact solution $\widetilde{\mathbf{u}}_{\text{exact}}$ solving directly the augmented system and we present a table containing the number of iterations required to reach a relative tolerance of Tol := 10^{-6}, i.e.

$$\frac{\|\widetilde{\mathbf{u}}_l^n - \widetilde{\mathbf{u}}_{\text{exact}}\|_\infty}{\|\widetilde{\mathbf{u}}_{\text{exact}}\|_\infty} \leq \text{Tol}.$$

3.5.1 Elliptic problems and scalability

We first consider the discrete setting described in Sections 3.1 and 3.3.1 with overlap fixed to $a = 0.0625$. We study the diffusion equation

$$-\nabla \cdot \nu(x,y)\nabla u = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \partial\Omega, \tag{3.5.1}$$

where $\nu(x,y) = \nu_1$ in Ω_1 and $\nu(x,y) = \nu_2$ in Ω_2. We define the ratio $\lambda = \frac{\nu_1}{\nu_2}$ as a measure of the heterogeneity. If heterogeneity is present, we do not consider overlapping methods. The coarsest grid corresponds to $\ell = 3$ and the finest to $\ell = 9$, leading respectively to 56 and 261632 degrees of freedom. Table 3.1 shows the number of iterations to reach the tolerance for the different methods. For the homogeneous case, i.e. $\lambda = 1$, the V-cycle OSM is faster than both the one-level OSM and the multigrid method in terms of iterations counts. In presence of heterogeneity, multigrid performance remains similar, while all the methods based on optimized Schwarz methods, both one-level and multilevel variants, become faster. This is in accordance with the discussion in Chapter 2. However to have faster convergence, OSMs do require the jump in the diffusion coefficient to be aligned along the interfaces between the subdomains and to proper rescale the transmission conditions according the diffusivity constants of the adjacent subdomains. If this is not the case, OSMs could even diverge. Therefore the method is not robust with respect

λ	OSM(p)	OSM(p,q)	OSMo(p)	OSMV(p)	OSMV(p,q)	OSMoV(p)	MGV
1	164	57	13	4	4	2	11
10^5	6	5	-	1	1	-	11

# Levels	OSMV(p)	OSMV(p,q)	OSMoV(p)
2	4	4	2
4	4	4	2
6	4	4	2

Table 3.1: Top table: number of iterations to reach the tolerance for a diffusion problem for the V-cycle OSMs and the multigrid scheme with a point Jacobi smoother. Bottom table: number of iterations to reach convergence as the number of levels increases in the multilevel methods for $\lambda = 1$.

ϵ	OSMV(p)	MGV	MGV-Line Jacobi
10^{-1}	4	59	6
10^{-3}	5	4769	6

Table 3.2: Number of iterations to reach the tolerance for the anisotropic Laplace equation for the V-cycle OSM, the multigrid scheme with a point Jacobi smoother and with a Line Jacobi smoother.

to arbitrary decompositions into subdomains in case of jumping diffusion coefficients. Some recent developments considering discontinuities across the interfaces are available in [108].

We then study the robustness of the methods with respect to the number of levels. We fix the finest grid to $\ell = 9$ and Table 3.1 shows that the number of iterations remains constant as the number of levels increases.

We then consider the anisotropic version of (3.5.1) where $\nu(x, y) = \text{diag}(\epsilon, 1)$. If ϵ is small, then we have a higher diffusivity in the y direction than in the x direction and multigrid performance deteriorates due to the inefficiency of classical smoothers, see Chapter 5 of [152]. Table 3.2 shows that the OSMV(p) does not suffer the anisotropy while multigrid becomes inefficient. A Jacobi line smoother fixes multigrid but it also makes each iteration much more expensive as it requires to solve N_x tridiagonal systems of dimension $N_y \times N_y$, where N_x and N_y are the number of unknowns in the x and y direction. For three dimensional problems one needs to perform plane relaxations which could be very complicated to implement in complex geometries [148]. The OSM smoother solves instead two (or several for a many subdomains decomposition) larger sparse subdomain problems. Depending on the implementation and architecture, these two costs can be comparable.

ν	a_1	a_2	OSM(p)	OSM(p,q)	OSMV(p)	OSMV(p,q)	MGV
1	1	1	166	57	5	4	15
1	20	1	99	44	8	7	16
1	20	20	101	48	7	6	18

Table 3.3: Number of iterations to reach the tolerance for the advection-diffusion equation in different physical regimes.

Next we solve the advection diffusion equation

$$-\nu\Delta u + \mathbf{a}\cdot\nabla u = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \partial\Omega,$$

where $\nu \in \mathbb{R}$ and $\mathbf{a} = (a_1, a_2)^\top \in \mathbb{R}^2$. We refer to [95] for the analysis of the one-level OSM for advection-diffusion PDEs in bounded domains. The coarsest mesh is $\ell = 5$, equivalent to 992 degrees of freedom, so that on the coarsest level we still have a rough description of the boundary layer due to the advection. Table 3.3 shows the number of iterations to reach convergence in different physical regimes. Note that being faster in terms of iteration numbers does not mean being faster in computational time: let us study the computational cost of a two-level optimized Schwarz method and of a multigrid scheme using point Jacobi. We denote with N, N_c and M the number of degrees of freedom on the first level, on the second level and in each subdomain on the fine mesh. N_{sub} indicates the total number of subdomains while N^{it} and N^{it}_{MG} are the number of iterations of the two-level optimized Schwarz method and of the multigrid scheme. Then, the computational cost (CC) of the two level optimized Schwarz method is $CC^{\text{MOSM}} = O(N^{\text{it}}((n_1+n_2)N_{sub}M^\gamma + N_c^\gamma))$ while for multigrid[3] $CC^{\text{MG}} = O(N^{\text{it}}_{\text{MG}}((n_1+n_2)N + N_c^\gamma))$, where γ is an exponent which depends on the structure of the matrix and on the linear solver used. It is clear that the subdomain solvers can represent a bottleneck due to the term M^γ. One solution is to increase N_{sub} so that M becomes smaller and one can then do the computations in parallel. We then study the scalability properties of OSMoV(p) with two levels, i.e. a two-level method which uses an overlapping optimized Schwarz method with one optimized parameter p as a smoother on the fine level. We consider a square domain Ω divided into several subdomains by the partitioning tool Metis, see Figure 3.7 for an example of a decomposition. As we increase the number of subdomains, we keep the size of each subdomain approximately constant (around 400 degrees of freedom), so that the global problem becomes larger. In this setting, we solve (3.5.1) with $\nu(x,y) = 1$ and $f = 1$. The numbers of pre- and post-smoothing steps are equal to $n_1 = n_2 = 2$ and the overlap is constant, equal to four times the mesh size. Table 3.7 shows that the two-level optimized Schwarz method is scalable and offers a comparison with the RAS method equipped with the well-known Nicolaides coarse space, see Section 4.2 of [61] for a detailed description. The OSMoV(p) requires much less iterations but, at least in its two-level variant, it requires to solve a larger and more expensive coarse problem.

[3] We suppose to use a point wise Jacobi smoother which has a linear cost in N. However in several situations, e.g. Table 3.2, one has to rely on more expensive smoothers.

N. subdomains	4	16	64	128
OSMoV(p)	3	3	4	4
RAS+Nicolaides	8	27	52	57

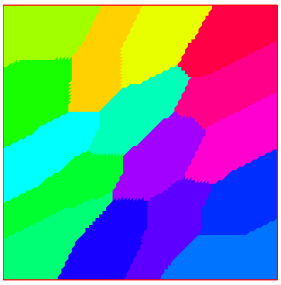


Figure 3.7: On the left, number of iterations to reach convergence as the number of subdomains increases. On the right, example of decomposition into 16 subdomains using Metis.

ω	OSM(p)	OSMV(p)	MGV
5π	15	2	4
25π	25	4	6
50π	34	9	10
100π	60	40	129

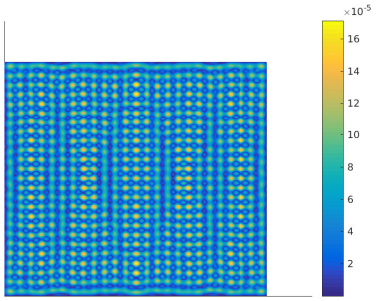


Table 3.4: Convergence behavior for the Helmholtz equation with different wavenumbers for a two-level method. Fine mesh labeled by $\ell = 10$ and coarse mesh by $\ell = 9$. On the right solution for $\omega = 25\pi$.

3.5.2 Helmholtz equation with a dispersion correction

We consider the Helmholtz equation in a square cavity open on the vertical edges with transparent Robin boundary conditions and with homogeneous Dirichlet conditions on the horizontal ones. Both OSMs and multigrid do not converge in general for Helmholtz problems when used iteratively, see Refs [71, 91] for a detailed discussion. The oscillatory nature of the Helmholtz equation makes it difficult to design efficient two-level solvers. Some recents developments are available in [105, 46, 13, 147]. In this paragraph we consider GMRES preconditioned by the one-level OSM [91], the V-cycle OSM and the multigrid scheme. We show two numerical experiments. In the first one, we test the two-level OSM with a fine mesh $\ell = 10$, approximately one million degrees of freedom, and a coarse mesh $\ell = 9$ and we compare the iterations required to converge for an increasing sequence of wave numbers ω. In Table 3.4, we see that the V-cycle OSM and multigrid schemes are extremely fast especially for low wave number. This is not surprising; in order for the coarse correction to be effective, we need a good representation of the error on the coarse mesh, therefore the lower the oscillations are the better representation we have and we are basically in an elliptic regime. As ω increases, the V-cycle OSM deteriorates

$\ell_{\min}$	$G_{\ell_{\min}}$	OSMV(p)	MGV
7	10.24	9	16
6	5.12	16	78
5	2.56	24	>200

$\ell_{\min}$	$G_{\ell_{\min}}$	OSMV(p)	MGV
7	10.24	9	8
6	5.12	16	20
5	2.56	26	>200

Table 3.5: Convergence behavior of the V-cycle OSM and of the multigrid scheme for $\omega = 25\pi$ as the number of points per wavelength on the coarsest grid $G_{\ell_{\min}}$ is reduced. The right table refers to the dispersion correction. The finest grid corresponds to $\ell_{\max} = 8$ with $G_{\ell_{\max}} = 20.48$

and multigrid becomes highly ineffective.

In the second experiment we investigate the robustness of the multilevel methods with respect to the coarseness of the meshes. In order to provided a good coarse correction, the coarse mesh should have at least a resolution of approximately ten points per wavelength $G_\ell := \frac{2\pi}{h_\ell \omega} \approx 10$. Moreover a coarse mesh amplifies the numerical dispersion; therefore, since this requirement sets a practical constraint on the use of multigrid for Helmholtz problems, some methods have been developed for dispersion correction such as optimized finite difference schemes, see [147, 45]. In the following, we do not use some specific new finite difference stencils to contain the numerical dispersion, but instead on each level, we modify the frequency ω of the Helmholtz equation. Indeed, in Section 2 of [45], it is shown that choosing the Helmholtz frequency on each level such that

$$\omega_\ell(\theta) = \left| \sqrt{h_\ell^{-2}(4 - 2\cos(\omega h_\ell \cos(\theta)) - 2\cos(\omega h_\ell \sin(\theta)))} \right|, \tag{3.5.2}$$

reduces the numerical dispersion and specifically it removes the dispersion in the direction defined by the angle θ. Thus, supposing that on the finest grid $\ell_{\max}$, the numerical dispersion is negligible, on each coarser mesh we discretize the Helmholtz equation with a modified frequency $\omega_\ell(\theta)$. We choose the angle $\theta = \frac{\pi}{8}$ since it is very close to the value found numerically which minimizes the maximum of the Euclidean distance between the points lying on the continuous dispersion relation $\{\xi \in \mathbb{R}^2 : \|\xi\| = \omega\}$ and the discrete one $\{\xi \in \mathbb{R}^2 : h^{-2}(4 - 2\cos(h_\ell \xi_1) - 2\cos(h_\ell \xi_2)) = \omega(\theta)\}$, for $\omega = 25\pi$.

Table 3.5 shows that multigrid is very sensitive to the coarseness of the meshes and that the dispersion correction improves its convergence behaviour up to $G_\ell \approx 5$. The V-cycle OSM is instead more robust than multigrid and it is unaffected by the correction of the frequency ω.

3.5.3 Helmholtz-Laplace heterogeneous coupling

We study the MOSM for the Helmholtz-Laplace coupling discussed in Chapter 2.4,

$$-\Delta u - \omega^2 u = f \quad \text{in} \quad \Omega_1, \quad -\Delta u = f \quad \text{in} \quad \Omega_2, \qquad u = 0 \quad \text{on} \quad \partial\Omega.$$

We consider the finest grid $\ell = 9$ and the coarsest $\ell = 7$ such that we have more than 10 points per wavelength. We see in Table 3.6 that the multigrid V-cycle diverges as an itera-

ω	OSM(p)	OSM-V(p)	MG-V
5π	165	6	13
25π	80	9	div

Table 3.6: Number of iterations to reach the tolerance for the different methods in the Helmholtz-Laplace coupling.

tive method for a large wave number. The MOSM still converges but the coarse correction clearly becomes less effective. The one-level OSM instead improves its performance for increasing ω as long as the mesh size does not increase, as discussed in the numerical results of Chapter 2.4. If we choose the coarsest grid equal to $l = 8$, then also the multigrid V-cycle converges but it requires 111 iterations, which illustrates the higher sensitivity of the multigrid scheme compared to the MOSM for wave problems, see also subsection 3.5.2.

CHAPTER 4

Substructured Two-Level and Multilevel Domain Decomposition methods

"Wisdom is brilliant, she never fades. By those who love her, she is readily seen, by those who seek her, she is readily found.
She anticipates those who desire her by making herself known first.
Whoever gets up early to seek her will have no trouble but will find her sitting at the door."

— Book of Wisdom, Chapter 6, 12-14.

Let us consider a vector space V[1], an invertible linear operator $A : V \to V$ and an element $\boldsymbol{b} \in V$. As we have seen in Chapter 3, a two-level domain decomposition method consists of a classical one-level domain decomposition method also called "smoother", e.g. one of the methods presented in Sections 1.2 and 1.3, and a coarse correction step performed on a coarse space $V_c \subset V$. In order for the coarse correction to be cheap, we assume that $\dim V_c \ll \dim V$. Once V_c is defined, the mappings between V and V_c are realized by a restriction operator $R : V \to V_c$ and a prolongation operator $P : V_c \to V$. We will denote the restriction of A on V_c with A_c, and set $A_c := RAP$. We emphasize that the efficiency of a two-level method relies on good team-play between the one-level smoother and the coarse space. Indeed, the coarse space should contain those functions which are slowing down the convergence of the one-level smoother. Heuristically, these "bad" functions can be easily identified by running the one-level smoother for a few iterations and looking directly at the form of the error and of the residual. For one-level domain decomposition methods, the error and residual have a particular form after each iteration. The error is harmonic inside the subdomains and it is predominant in the overlap region, where we simply merge the different subdomain contributions. This observation

[1] The space V will be assumed of infinite dimension in Sections 4.1 and 4.2 (except for Theorem 4.2.4) and of finite dimension in Section 4.3

has motivated the development of different techniques to define coarse functions inside the overlap and then to extend them in the interior of subdomains. In this direction we can refer to the manuscripts [57, 63, 68, 72, 73, 113, 118, 144, 145]. On the other hand, if one uses a nonoverlapping method (e.g. the OSM see Chapter 3) or an overlapping method with a partition of unity corresponding to an algebraic nonoverlapping partition of the unknowns, one gets that the residual is non-zero only along the interfaces between the nonoverlapping (either physical or algebraic) subdomains. Therefore another mainstream idea is to define functions on the interfaces and then to extend them in the interior of the subdomains, see e.g. [1, 11, 31, 33, 39, 84, 90, 89, 94, 113, 99].

In this chapter, we introduce new computational frameworks to solve the linear system $A\boldsymbol{u} = \boldsymbol{b}$, and we call them substructured two-level and multilevel domain decomposition methods. The term "substructured" is commonly used in the literature to refer to nonoverlapping domain decomposition methods, see for instance the monographs [139, 151], and this particular meaning essentially derives from [138]. We will instead use the term "substructured" to indicate that both the one-level domain decomposition method and the coarse space are defined on the interfaces, independently of the type of (overlapping or non-overlapping) decomposition of the domain. These interfaces will coincide with the physical interfaces between the subdomains in the case of a nonoverlapping decomposition, or they will consist of the parts of the boundary of a subdomain which lie in the interior of another subdomain, in the case of an overlapping decomposition. See Section 4.1 for a precise mathematical definition.

The content of this chapter is based on a long collaboration with Gabriele Ciaramella, which resulted in the manuscripts [41, 43, 42]. In [41] we presented the new computational framework for the two subdomain case, introducing a substructured two-level method based on a spectral coarse space, called Spectral 2-level Substructured (S2S) method, and another substructured two-level method based on a geometric coarse space, called Geometric 2-level Substructured (G2S) method.
The S2S method is based on a coarse space defined as the span of certain interface functions. One can choose freely these interface functions, even though not all choices are equivalent from the convergence point of view. An effective choice is to define the coarse space as the span of the eigenfunctions of the one-level substructured smoother which are associated to the largest eigenvalues in modulo. Clearly, computing the eigenfunctions of the one-level smoother is not always desirable. A partial remedy is to compute separately the eigenvectors of some local operators defined on each subdomain. Another approach is to approximate the leading eigenvectors of the one-level smoother, and we propose a numerical procedure based on the principal component analysis to obtain good approximations in an inexpensive way. Nevertheless, we emphasize that one could choose the interface functions freely, and for instance one could use some of the interface functions used by other two-level domain decomposition in volume, see for instance [90, 89, 94, 113]. In Section 4.6, we will show numerical experiments using the interface functions of the SHEM (Spectral Harmonically Enriched Multiscale) coarse space ([90]) as a basis for our substructured coarse space.

On the other hand, the G2S method is essentially a two-grid interface method, for which the coarse correction is performed on a coarse interface grid. The G2S framework does not require the explicit knowledge of the coarse space functions and it is suitable to be extended to a multilevel method, whenever the dimension of the coarse space is too large. We will show that the G2S method converges much faster than a two-level method in volume using RAS as a smoother, underlying that the G2S method is not just a different implementation of already known methods. For a discussion about the relation between the G2S method and a two-grid volume method, we refer to [43, Section 5]. The S2S method is presented in Section 4.2 while the G2S method is discussed in Section 4.3. Both methods have been generalized to the many-subdomain case and further analyzed respectively in [42] in [43]. In this chapter we are going to provide a complete overview, but several details will not be treated. The interested reader can refer to [43] and [42].

At this point, one could ask what are the advantages and disadvantages of working at the substructured level. Let us start considering a spectral two-level method in volume and the S2S method. On the one hand, most of the two-level methods in volume construct some functions in the overlap region or on the interfaces, which are then extended inside the subdomains through subdomain solves. All these methods inevitably need this extension step since they are defined in volume. Clearly, working at the substructured level, the S2S method does not need to extend the interface basis functions. On the other hand, as we will see in Section 4.2, if one wants to build explicitly the substructured coarse matrix A_c, then one would have to perform subdomain solves. Thus, if A_c is build explicitly, the two costs compensate. From the memory storage point of view, the substructured methods require less storage as the interface functions are represented by only the degrees of freedom on the substructures. Thus, for a three-dimensional problem with mesh size h, a discrete interface coarse function is an array of size $O(1/h^2)$ which is much smaller than $O(1/h^3)$, which is the size of an array corresponding to a coarse function in volume. Thus, for this reason the resulting interface restriction and prolongation operators are much smaller matrices and thus their action is cheaper to compute. Furthermore, we numerically observed that a SHEM coarse space and a spectral S2S coarse space of same dimension lead to a very similar convergence behaviour.

Concerning the G2S method, we first remark that the size of the coarse matrix A_c is much smaller than the size of a coarse matrix in volume. This implies that the coarse correction is cheaper to compute and that less levels are needed to ease the burden of the direct solve. Similarly to the S2S method, if one wants to assemble A_c of a G2S method explicitly, then one has to perform subdomain solves. This step can be done in parallel in a pre-computation phase. Furthermore, the G2S method exhibits a much faster convergence than a corresponding two-grid domain decomposition method, see for instance Figure 4.6. Regarding the geometric interpolation and restriction operators, we emphasize that if a two-level method in volume performs an interpolation in a n dimensional space, then the G2S method performs an interpolation in $n-1$ dimensional space. For a three-dimensional problem this property can be highly attractive.

4.1 Substructured Parallel Schwarz method

We consider the model problem

$$\mathscr{L}u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{4.1.1}$$

and we suppose it admits a unique solution $u \in H_0^1(\Omega)$. We decompose Ω into N overlapping Lipschitz subdomains Ω'_j, that is $\Omega = \cup_{j\in\mathscr{I}}\Omega'_j$ with $\mathscr{I} := \{1,2,\ldots,N\}$. We now introduce a notation inspired by [40]. For any $j \in \mathscr{I}$, we consider the set of neighbouring indices $\mathscr{N}_j := \left\{\ell \in \mathscr{I} : \Omega'_j \cap \partial\Omega'_\ell \neq \emptyset\right\}$. Given a $j \in \mathscr{I}$, we define the substructure of Ω'_j as $S_j := \cup_{\ell\in\mathscr{N}_j}\left(\partial\Omega'_\ell \cap \Omega'_j\right)$, that is the union of all the portions of $\partial\Omega'_\ell$ which lie in the interior of Ω'_j. The sets S_j are open and their closures are $\overline{S_j} = S_j \cup \partial S_j$, with $\partial S_j := \cup_{\ell\in\mathscr{N}_j}\left(\partial\Omega'_j \cap \partial\Omega'_\ell\right)$. The substructure of Ω is defined as $S := \cup_{j\in\mathscr{I}}\overline{S_j}$. We denote with $\mathscr{E}_j^0 : L^2(S_j) \to L^2(S)$ the extension by zero operator.
Given a bounded set Γ with boundary $\partial\Gamma$, we denote by $\rho_\Gamma(x)$ the function representing the distance of $x \in \Gamma$ from $\partial\Gamma$. We can then introduce the $H_{00}^{1/2}(\Gamma)$ space

$$H_{00}^{1/2}(\Gamma) := \{v \in H^{1/2}(\Gamma) : v/\rho_\Gamma^{1/2} \in L^2(\Gamma)\}, \tag{4.1.2}$$

which is also known as the Lions-Magenes space; see, e.g., [124, 139, 149]. Notice that $H_{00}^{1/2}(\Gamma)$ can be equivalently defined as the space of functions in $H^{1/2}(\Gamma)$ such that their extensions by zero to a superset $\widetilde{\Gamma}$ of Γ are in $H^{1/2}(\widetilde{\Gamma})$; see, e.g., [149]. Now, we consider a set of partition of unity functions $\chi_j : \overline{S_j} \to [0,1]$, $j = 1,\ldots,N$, such that $\sum_{j\in\mathscr{I}}\mathscr{E}_j^0\chi_j \equiv 1$ and

$$\chi_j(x) = \begin{cases} (0,1] & \text{for } x \in S_j, \\ \{1\} & \text{for } x \in \overline{S_j \setminus \cup_{\ell\in\mathscr{N}_j} S_\ell}, \\ \{0\} & \text{for } x \in \partial S_j \setminus \partial\Omega. \end{cases}$$

Further, we assume that the functions χ_j, $j \in \mathscr{I}$, satisfy the condition $\chi_j/\rho_{\mathscr{S}_j}^{1/2} \in L^\infty(\mathscr{S}_j)$. For any $j \in \mathscr{I}$, we define $\Gamma_j^{\text{int}} := \partial\Omega'_j \cap \left(\cup_{\ell\in\mathscr{N}_j}\Omega'_\ell\right)$, and introduce the following trace and restriction operators

$$\tau_j : H^1(\Omega'_j) \to H^{\frac{1}{2}}(S_j) \text{ and } \tau_j^{\text{int}} : H^{\frac{1}{2}}(S) \to H^{\frac{1}{2}}(\Gamma_j^{\text{int}}).$$

The operator τ_j takes a volume function defined over Ω'_j and returns the trace over the interior substructure S_j. On the other hand, τ_j^{int} takes a function defined over the whole substructure S and returns its restriction over Γ_j^{int}, that is the boundary of Ω'_j which lie in the interior of other subdomains.

It is well known that (4.1.1) is equivalent to the domain decomposition system (see, e.g., [139])

$$\mathscr{L}u_j = f_j \text{ in } \Omega'_j,\ u_j = \sum_{\ell\in\mathscr{N}_j}\mathscr{E}_\ell^0(\chi_\ell\tau_\ell u_\ell) \text{ on } \Gamma_j^{\text{int}},\ u_j = 0 \text{ on } \partial\Omega'_j \setminus \Gamma_j^{\text{int}}, \tag{4.1.3}$$

where $f_j \in L^2(\Omega'_j)$ is the restriction of f on Ω'_j. We emphasize that the properties of the partition of unity functions χ_ℓ guarantee that $\chi_\ell \tau_\ell u_\ell$ lies in $H^{\frac{1}{2}}_{00}(S_\ell)$ and $\mathscr{E}^0_\ell(\chi_\ell \tau_\ell u_\ell) \in H^{\frac{1}{2}}_{00}(S)$. Moreover, for $\ell \in \mathscr{N}_j$ it holds that $\tau^{\text{int}}_j \mathscr{E}^0_\ell(\chi_\ell \tau_\ell u_\ell) \in H^{1/2}_{00}(\Gamma^{\text{int}}_j)$ if $\Gamma^{\text{int}}_j \subsetneq \partial\Omega_j$, and $\tau^{\text{int}}_j \mathscr{E}^0_\ell(\chi_\ell \tau_\ell u_\ell) \in H^{1/2}(\Gamma^{\text{int}}_j)$ if $\Gamma^{\text{int}}_j = \partial\Omega_j$. Given a $j \in \mathscr{I}$ such that $\partial\Omega'_j \setminus \Gamma^{\text{int}}_j \neq \emptyset$, we define the extension operator $\mathscr{E}_j : H^{\frac{1}{2}}_{00}(\Gamma^{\text{int}}_j) \times L^2(\Omega'_j) \to H^1(\Omega'_j)$ as $\mathscr{E}_j(v) := w$, where w solves the problem

$$\mathscr{L} w = f_j \text{ in } \Omega'_j,\ w = v \text{ on } \Gamma^{\text{int}}_j,\ w = 0 \text{ on } \partial\Omega'_j \setminus \Gamma^{\text{int}}_j \tag{4.1.4}$$

for a $v \in H^{\frac{1}{2}}_{00}(\Gamma^{\text{int}}_j)$. Otherwise, if $\Gamma^{\text{int}}_j = \partial\Omega'_j$, we define $\mathscr{E}_j : H^{\frac{1}{2}}(\Gamma^{\text{int}}_j) \times L^2(\Omega'_j) \to H^1(\Omega'_j)$ as $\mathscr{E}_j(v) := w$, where w solves the problem

$$\mathscr{L} w = f_j \text{ in } \Omega'_j,\ w = v \text{ on } \Gamma^{\text{int}}_j, \tag{4.1.5}$$

for a $v \in H^{\frac{1}{2}}(\Gamma^{\text{int}}_j)$. Using linearity, the domain decomposition system (4.1.3) can be written as

$$u_j = \mathscr{E}_j(0, f_j) + \mathscr{E}_j\Big(\tau^{\text{int}}_j \sum_{\ell \in \mathscr{N}_j} \mathscr{E}^0_\ell(\chi_\ell \tau_\ell u_\ell), 0\Big),\ j \in \mathscr{I}. \tag{4.1.6}$$

We now apply the operator $\chi_j \tau_j$ on both sides of (4.1.6). Defining $v_j := \chi_j \tau_j u_j,\ j \in \mathscr{I}$, then system (4.1.6) becomes

$$v_j = g_j + \sum_{\ell \in \mathscr{N}_j} G_{j,\ell}(v_\ell),\ j \in \mathscr{I}, \tag{4.1.7}$$

where $g_j := \chi_j \tau_j \mathscr{E}(0, f_j)$ and the operators $G_{j,\ell} : H^{\frac{1}{2}}_{00}(S_\ell) \to H^{\frac{1}{2}}_{00}(S_j)$ are defined as

$$G_{j,\ell}(\cdot) := \chi_j \tau_j \mathscr{E}_j\big(\tau^{\text{int}}_j \mathscr{E}^0_\ell(\cdot), 0\big). \tag{4.1.8}$$

The system (4.1.7) is the substructured form of (4.1.3). The equivalence between (4.1.3) and (4.1.7) is explained by the following theorem.

Theorem 4.1.1 (Relation between (4.1.3) and (4.1.7))**.** *Let $u_j \in H^1(\Omega_j),\ j \in \mathscr{I}$, solve* (4.1.3)*, then $v_j := \chi_j \tau_j(u_j),\ j \in \mathscr{I}$, solve* (4.1.7)*. Let $v_j \in H^{\frac{1}{2}}(S_j),\ j \in \mathscr{I}$, solve* (4.1.7)*, then $u_j := \mathscr{E}_j(\tau^{int}_j \sum_{\ell \in \mathscr{N}_j} \mathscr{E}^0_\ell(v_\ell), f_j),\ j \in \mathscr{I}$, solve* (4.1.3)*.*

Proof. We have shown that $v_j := \chi_j \tau_j(u_j)$ satisfies the first statement while deriving the substructured system (4.1.7). To obtain the second statement, we proceed as follows. First we observe that

$$u_j = \mathscr{E}_j(0, f_j) + \mathscr{E}_j(\tau^{\text{int}}_j \sum_{\ell \in \mathscr{N}_j} \mathscr{E}^0_\ell(v_\ell), 0), \tag{4.1.9}$$

which multiplied by $\chi_j \tau_j$ and using (4.1.7) implies $v_j = \chi_j \tau_j u_j$. Thus replacing $v_\ell = \chi_\ell \tau_\ell u_\ell$ into (4.1.9) leads to (4.1.6). □

Given any function $w \in H_0^1(\Omega)$ and initializing $u_j^0 := w|_{\Omega_j'}$, $j \in \mathscr{I}$, the parallel Schwarz method (PSM) is given by

$$\mathscr{L} u_j^n = f_j \text{ in } \Omega_j,\ u_j^n = \sum_{\ell \in \mathscr{N}_j} \mathscr{E}_\ell^0(\chi_\ell \tau_\ell u_\ell^{n-1}) \text{ on } \Gamma_j^{\text{int}},\ u_j^n = 0 \text{ on } \partial\Omega_j \setminus \Gamma_j^{\text{int}}, \tag{4.1.10}$$

for $n \in \mathbb{N}^+$. The substructured form of the PSM is

$$v_j^n = g_j + \sum_{\ell \in \mathscr{N}_j} G_{j,\ell}(v_\ell^{n-1}),\ j \in \mathscr{I}, \tag{4.1.11}$$

initialized by $v_j^0 := \chi_j \tau_j(u_j^0) \in H_{00}^{\frac{1}{2}}(S_j)$. We emphasize that the iteration (4.1.11) is well posed, that is $v_j^n \in H_{00}^{\frac{1}{2}}(S_j)$ for $j \in \mathscr{I}$ and $n \in \mathbb{N}$. Equations (4.1.11) and (4.1.7) allow us to obtain the substructured PSM in error form, that is

$$e_j^n = \sum_{\ell \in \mathscr{N}_j} G_{j,\ell}(e_\ell^{n-1}),\ j \in \mathscr{I}, \tag{4.1.12}$$

for $n \in \mathbb{N}^+$, where $e_j^n := v_j - v_j^n$, for $j \in \mathscr{I}$ and $n \in \mathbb{N}$. Equation (4.1.7) can be written in the matrix form $A\mathbf{v} = \mathbf{b}$, where $\mathbf{v} = [v_1, \ldots, v_N]^\top$, $\mathbf{b} = [g_1, \ldots, g_N]^\top$ and the entries of A are

$$[A]_{j,j} = I_{d,j} \text{ and } [A]_{j,\ell} = -G_{j,\ell},\ j,k \in \mathscr{I},\ j \neq k, \tag{4.1.13}$$

where $I_{d,j}$ are the identities on $L^2(S_j)$, $j \in \mathscr{I}$. Similarly, we define the operator G as

$$[G]_{j,j} = 0 \text{ and } [G]_{j,\ell} = G_{j,\ell},\ j,k \in \mathscr{I},\ j \neq k,$$

which allows us to write equations (4.1.11) and (4.1.12) as $\mathbf{v}^n = G\mathbf{v}^{n-1} + \mathbf{b}$ and $\mathbf{e}^n = G\mathbf{e}^{n-1}$, respectively, where $\mathbf{v}^n := [v_1^n, \ldots, v_N^n]^\top$ and $\mathbf{e}^n := [e_1^n, \ldots, e_N^n]^\top$. Notice that $G = I - A$, where $I := \text{diag}_{j=1,\ldots,N}(I_{d,j})$.

4.2 S2S method

In the next two subsections we focus on the S2S method which relies on a coarse space defined as the span of certain interface basis functions. Ideally, one could define the coarse space as the span of some of the eigenfunctions of the smoothing operator G. This leads to a very efficient method. However, computing these eigenfunctions can be cumbersome from the computational point of view. In these cases, we will discuss how to approximate them through a Principal Component Analysis (PCA) which shares similarities with the randomized SVD. In alternative, one could also decide to use the eigenvectors of the subdomain smoothing operators G_j, which hopefully are less expensive to compute, and we will provide a convergence analysis in the case of two subdomains. It is even possible to build the coarse space heuristically, by inserting functions which are thought to be relevant for the convergence of the method. For instance, one could insert the lowest Fourier modes defined on each substructure. Finally, one can also use the extensive literature available for two-level domain decomposition methods in volume. In Section 4.6 we

Algorithm 3: Two-level substructured domain decomposition method

Require: $\boldsymbol{u}^0$ (initial guess)
1: $\boldsymbol{u}^n = G\boldsymbol{u}^{n-1} + \boldsymbol{b}$, $n = 1, \dots, n_1$ (dd pre-smoothing steps)
2: $\boldsymbol{r} = \boldsymbol{r} - A\boldsymbol{u}^{n_1}$ (compute the residual)
3: Solve $A_c \boldsymbol{u}_c = R\boldsymbol{r}$ (solve the coarse problem)
4: $\boldsymbol{u}^0 = \boldsymbol{u}^{n_1} + P\boldsymbol{u}_c$ (coarse correction)
5: $\boldsymbol{u}^n = G\boldsymbol{u}^{n-1} + \boldsymbol{b}$, $n = 1, \dots, n_2$ (dd post-smoothing steps)
6: Set $\boldsymbol{u}^0 = \boldsymbol{u}^{n_2}$ (update)
7: Repeat from 1 to 6 until convergence

present numerical experiments using the interface functions involved in the construction of the SHEM coarse space [90, 89]. When needed, we will specify the coarse space used by the S2S method. We will denote with S2S-G, S2S-G_j, S2S-PCA and S2S-HEM, S2S methods with coarse spaces made of respectively eigenfunctions of G, eigenfunctions of G_j, random functions obtained through a PCA, and a coarse space obtained through the SHEM coarse space, see Section 4.6 for implementation details.

In spite of the choice of the interface functions, we now provide a general definition of the S2S method. We introduce the space $V := \otimes_{j=1}^{N} H_{00}^{\frac{1}{2}}(S_j)$ and an operator $A : V \to V$. The space V can be equipped with an inner product structure. The inner product is defined as $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{j=1}^{N} \langle v_j, w_j \rangle_j$, where $v_j, w_j \in H_{00}^{\frac{1}{2}}(S_j)$ and $\langle , \rangle_j$ is the inner product on $H_{00}^{\frac{1}{2}}(S_j)$. We aim to solve the linear system $A\boldsymbol{u} = \boldsymbol{b}$. Let us suppose to have available a coarse space $V_c \subset V$ such that $V_c = \operatorname{span}\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_{N_c}\}$. The dimension of the coarse space is N_c. We define the restriction and prolongation operators $R : V \to \mathbb{R}^{N_c}$ and $P : \mathbb{R}^{N_c} \to V$ as

$$\begin{aligned} R\boldsymbol{v} &= (\langle \boldsymbol{v}, \boldsymbol{\psi}_1 \rangle, \langle \boldsymbol{v}, \boldsymbol{\psi}_2 \rangle, \dots, \langle \boldsymbol{v}, \boldsymbol{\psi}_{N_c} \rangle)^\top, && \forall v \in V, \\ P\mathbf{w} &= \sum_{j=1}^{N_c} w_j \boldsymbol{\psi}_j, && \forall \mathbf{w} \in \mathbb{R}^{N_c}. \end{aligned} \tag{4.2.1}$$

The restriction of A on V_c is defined in a Galerkin fashion, that is $A_c : V_c \to V_c$, with $A_c := RAP$.
The S2S method is then defined by Algorithm 3. The integers n_1 and n_2 are the numbers of pre- and post-smoothing steps. A direct calculation reveals that one iteration of the two-level method can be written in the form of a stationary method

$$\boldsymbol{u}^{\text{new}} = G^{n_2}(I - PA_c^{-1}RA)G^{n_1}\boldsymbol{u}^{\text{old}} + \widetilde{M}\boldsymbol{b}, \tag{4.2.2}$$

where I is the indentity operator over V. Here, $\widetilde{M}$ is an operator which acts on the right-hand side vector $\boldsymbol{b}$ and which can be regarded as the preconditioner corresponding to our two-level method. A direct calculation shows that

$$\widetilde{M} = I + PA_c^{-1}R + PA_c^{-1}RA,$$

which corresponds to a two-level preconditioner where the coarse correction is treated in a multiplicative way. Defining the errors $\boldsymbol{e}^{\text{new}} := \boldsymbol{v} - \boldsymbol{v}^{\text{new}}$ and $\boldsymbol{e}^{\text{old}} := \boldsymbol{v} - \boldsymbol{v}^{\text{old}}$, iteration (4.2.2) becomes

$$\boldsymbol{e}^{\text{new}} = T\boldsymbol{e}^{\text{old}}, \quad T := G^{n_2}(I - PA_c^{-1}RA)G^{n_1}. \tag{4.2.3}$$

In order for algorithm (3) to be well-defined, we need first to show that A_c is invertible. This is discussed in the next Lemma where we need the orthogonal projection operator onto V_c.

Definition 4.2.1. Given a space V and a subspace $V_c = \text{span}\{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{N_c}\}$, we define the orthogonal projector $\mathbb{P}_{V_c}$ as the linear operator $\mathbb{P}_{V_c} : V \to V_c$ such that

$$\forall\, \boldsymbol{v} \in V, \quad \langle \boldsymbol{v} - \mathbb{P}_{V_c}\boldsymbol{v}, \boldsymbol{\psi}_j \rangle = 0, \quad j = 1, \dots, N_c.$$

Lemma 4.2.2 (Invertibility of a coarse operator A_c)**.** *Let V be an inner product space and V_c be a finite-dimensional subspace of V given by the span of the basis functions $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{N_c}$. Let $\mathbb{P}_{V_c}$ be the orthogonal projection operator onto V_c. Consider an invertible operator $A : V \to V$ and the matrix $A_c = RAP \in \mathbb{R}^{N_c \times N_c}$, where P and R are defined as in* (4.2.1)*. Then A_c has full rank if and only if $\mathbb{P}_{V_c}(A\boldsymbol{v}) \neq 0\ \forall\, \boldsymbol{v} \in V_c \setminus \{0\}$.*

Proof. We first show that if $\mathbb{P}_{V_c}(A\mathbf{v}) \neq 0$ for any $\mathbf{v} \in V_c \setminus \{0\}$, then $A_c = RAP$ has full rank. This result follows from the rank-nullity theorem, if we show that the only element in the kernel of A_c is the zero vector. To do so, we recall the definitions of P and R given in (4.2.1). Clearly, $P\mathbf{z} = 0$ if and only if $\mathbf{z} = 0$. For any $\mathbf{z} \in \mathbb{R}^{N_c}$ the function $P\mathbf{z}$ is in V_c. Since A is invertible, then $AP\mathbf{z} = 0$ if and only if $\mathbf{z} = 0$. Moreover, by our assumption it holds that $\mathbb{P}_{V_c}(AP\mathbf{z}) \neq 0$. Now, we notice that $R\mathbf{w} \neq 0$ for all $\mathbf{w} \in V_c \setminus \{0\}$, and $R\mathbf{w} = 0$ for all $\mathbf{w} \in V_c^{\perp}$, where $V_c^{\perp}$ denotes the orthogonal complement of V_c in V with respect to $\langle \cdot, \cdot \rangle$. Since $(V, \langle \cdot, \cdot \rangle)$ is an inner-product space, we have $AP\mathbf{z} = \mathbb{P}_{V_c}(AP\mathbf{z}) + (I - \mathbb{P}_{V_c})(AP\mathbf{z})$ with $(I - \mathbb{P}_{V_c})(AP\mathbf{z}) \in V_c^{\perp}$. Hence, $RAPz = R\mathbb{P}_{V_c}(AP\mathbf{z}) \neq 0$ for any non-zero $\mathbf{z} \in \mathbb{R}^{N_c}$.

Now we show that, if $A_c = RAP$ has full rank, then $\mathbb{P}_{V_c}(Av) \neq 0$ for any $\mathbf{v} \in V_c \setminus \{0\}$. We proceed by contraposition and prove that if there exists a $\mathbf{v} \in V_c \setminus \{0\}$ such that $A\mathbf{v} \in V_c^{\perp}$, then $A_c = RAP$ is not full rank. Assume that there is a $\mathbf{v} \in V_c \setminus \{0\}$ such that $A\mathbf{v} \in V_c^{\perp}$. Since $\mathbf{v}$ is in V_c, there exists a nonzero vector $\mathbf{z} \in \mathbb{R}^{N_c}$ such that $\mathbf{v} = P\mathbf{z}$. Hence $AP\mathbf{z} \in V_c^{\perp}$. We can now write that $A_c\mathbf{z} = R(AP\mathbf{z}) = 0$, which implies that A_c is not full rank. □

4.2.1 Spectral coarse space based on the eigenvectors of G: convergence analysis and PCA

We start considering the case in which the functions $\{\boldsymbol{\psi}_j\}_{j=1}^{N_c}$ are eigenfunctions of G, that is $G\boldsymbol{\psi}_j = \lambda_j \boldsymbol{\psi}_j$. We suppose that 1 is not an eigenvalue of G, so that the operator $A = I - G$ is invertible. We do not assume any orthogonality between the eigenfunctions. With such a choice of V_c, it holds $A(V_c) \subseteq V_c$ which implies $\mathbb{P}_{V_c}(A\boldsymbol{v}) \neq 0\ \forall\, \boldsymbol{v} \in V_c \setminus \{0\}$. Thus, due to Lemma 4.2.2, the matrix A_c is invertible. In the proof of the convergence Theorem 4.2.4 we will need a matrix representation of the orthogonal projector.

Lemma 4.2.3 (Matrix representation of the orthogonal projector)**.** *Once fixed the basis* $\{\boldsymbol{\psi}_j\}_{j=1}^{N_c}$ *for* V_c*, the action of* $\mathbb{P}_{V_c}$ *has the matrix representation* $P_{V_c} = P(RP)^{-1}R$*,*

Proof. Given a general $\boldsymbol{v} \in V$, we express $P_{V_c}\boldsymbol{v} = \sum_i^{N_c} \alpha_i \boldsymbol{\psi}_i = P\boldsymbol{\alpha}$, with $(\boldsymbol{\alpha})_j = \alpha_j$. Inserting into the orthogonal conditions $\langle \mathbb{P}_{V_c}\boldsymbol{v}, \boldsymbol{\psi}_j\rangle = \langle \boldsymbol{v}, \boldsymbol{\psi}_j\rangle$ we get

$$\sum_i^{N_c} \alpha_i \langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j\rangle = \langle \boldsymbol{v}, \boldsymbol{\psi}_j\rangle,$$

which can be rewritten as $RP\boldsymbol{\alpha} = R\boldsymbol{\psi}$ and the result follows. □

So far we have worked assuming V is an infinite dimensional space. To prove the main result of this subsection, we assume V is finite dimensional.

Theorem 4.2.4 (Convergence of the S2S method- Eigenfunctions of G)**.** *Consider a finite dimensional inner product space* V*, an invertible operator* $A : V \to V$*,* $A = I - G$*, and a coarse space* $V_c := span\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_{N_c}\}$ *where* $\boldsymbol{\psi}_j$ *are eigenvectors of* G *associated to the eigenvalues* λ_j*,* $j = 1, \dots, N_c$*. Furthermore we consider the operator* R *and* P *defined as* (4.2.1) *and* $A_c := RAP$*. Then, defining* $T := G^{n_2}(\mathbb{I} - PA_c^{-1}RA)G^{n_1}$*, we have*

$$\rho(T) = \max|\lambda|^{n_1+n_2} \quad \textit{such that} \quad \lambda \in \sigma(G) \setminus \{\lambda_1, \dots, \lambda_{N_c}\}.$$

Proof. We first introduce the operator $\widetilde{T} = (I - PA_c^{-1}RA)G^{n_1+n_2}$. The operators $\widetilde{T}$ and T have the same spectrum and thus we focus on $\widetilde{T}$. The proof of this theorem is divided into two parts. First we show that $\{\boldsymbol{\psi}_j\}_{j=1}^{N_c}$ are eigenvectors of $\widetilde{T}$ associated to the zero eigenvalue. Second, we show that all the other eigenvalues of G are still eigenvalues of $\widetilde{T}$ by constructing directly the corresponding eigenvector. Let us start with the first part. If we consider a $\boldsymbol{\psi}_j \in V_c$ we have

$$\begin{aligned}(I - PA_c^{-1}RA)G^{n_1+n_2}\boldsymbol{\psi}_j &= \lambda_j^{n_1+n_2}\boldsymbol{\psi}_j - \lambda_j^{n_1+n_2}PA_c^{-1}RA\boldsymbol{\psi}_j \\ &= \lambda_j^{n_1+n_2}(\boldsymbol{\psi}_j - (1-\lambda_j)PA_c^{-1}R\boldsymbol{\psi}_j).\end{aligned} \tag{4.2.4}$$

We now compute the action of A_c over a canonical vector $\boldsymbol{e}_j$, $j = 1, \dots, N_c$,

$$RAP\boldsymbol{e}_j = RA\boldsymbol{\psi}_j = (1-\lambda_j)R\boldsymbol{\psi}_j,$$

which, RAP being invertible, implies $\boldsymbol{e}_j = (1-\lambda_j)A_c^{-1}R\boldsymbol{\psi}$. Inserting this expression into (4.2.4) we obtain

$$(I - PA_c^{-1}RA)G^{n_1+n_2}\boldsymbol{\psi}_j = \lambda_j^{n_1+n_2}(\boldsymbol{\psi}_j - P\boldsymbol{e}_j) = \lambda_j^{n_1+n_2}(\boldsymbol{\psi}_j - \boldsymbol{\psi}_j) = 0.$$

We now focus on the remaining eigenvalues. For every eigenpair $(\boldsymbol{\psi}_k, \lambda_k)$ of G such that $\boldsymbol{\psi}_k \notin V_c$, we show that $(\boldsymbol{\phi}_k, \lambda_k^{n_1+n_2})$, with

$$\boldsymbol{\phi}_k := A^{-1}(\boldsymbol{\psi}_k - P_{V_c}\boldsymbol{\psi}_k) = \frac{1}{(1-\lambda_k)}\boldsymbol{\psi}_k - \boldsymbol{w}, \tag{4.2.5}$$

for some $\boldsymbol{w} \in V_c$, is an eigenpair of $\widetilde{T}$. We claim that $\boldsymbol{w} \in V_c$ since V_c, which is spanned by eigenvectors of G, is invariant under the action of A^{-1}. Using that $V_c \subset \mathrm{Ker}(\widetilde{T})$, we have

$$\widetilde{T}\boldsymbol{\phi}_k = \widetilde{T}\frac{1}{1-\lambda_k}\boldsymbol{\psi}_k = \lambda_k^{n_1+n_2}\left(\frac{1}{1-\lambda_k}\boldsymbol{\psi}_k - PA_c^{-1}R\boldsymbol{\psi}_k\right). \tag{4.2.6}$$

If the eigenvectors were orthonormal, we would have finished the proof, since $R\boldsymbol{\psi}_k = 0$. In a more general case, we proceed as follows. We observe that

$$PA_c^{-1}R\boldsymbol{\psi}_k = PA_c^{-1}RP(RP)^{-1}R\boldsymbol{\psi}_k = PA_c^{-1}R\mathbb{P}_{V_c}\boldsymbol{\psi}_k = \sum_{\ell=1}^{N_c}\gamma_\ell PA_c^{-1}R\boldsymbol{\psi}_\ell,$$

such that $\sum_{\ell=1}^{N_c}\gamma_\ell\boldsymbol{\psi}_\ell$ is the orthogonal projection of $\boldsymbol{\psi}_k$ onto V_c. Now, we recall that $R\boldsymbol{\psi}_\ell = A_c((1-\lambda_\ell)^{-1}\mathbf{e}_\ell)$, for $\ell = 1,\ldots,N_c$, and write $PA_c^{-1}R\boldsymbol{\psi}_\ell = \sum_{\ell=1}^{m}\gamma_\ell(1-\lambda_\ell)^{-1}\boldsymbol{\psi}_\ell = \sum_{\ell=1}^{N_c}\gamma_\ell A^{-1}\boldsymbol{\psi}_\ell = A^{-1}\mathbb{P}_{V_c}\boldsymbol{\psi}_k$. Replacing this equality into (4.2.6), we obtain $\widetilde{T}\boldsymbol{\phi}_k = \lambda_k^{n_1+n_2}\boldsymbol{\phi}_k$. □

Theorem 4.2.4 provides very interesting insights on the convergence of the S2S method. The coarse space V_c is such that the operator T has the same eigenvalues of the one-level smoother G, except for those eigenvalues corresponding to eigenvectors which are in V_c. These latter eigenvalues are actually mapped to zero. It follows that the choice of the coarse space is extremely important. On the one hand, if the one-level smoother G has a large eigenvalue, approximately equal to 1, and V_c does not contain the corresponding eigenvector, then the S2S will be as slow as the one-level smoother. On the other hand, even if G is not converging, e.g. it has an eigenvalue larger than 1, then if V_c contains the corresponding eigenvector, then the S2S method will converge. In other words, the coarse correction can transform a divergent method into a converging one (see [33, 30] for a similar result concerning the Neumann-Neumann method and [94] for the AS method). We will see in Section 4.6 that for high contrast jumping diffusion coefficients, the PSM has just few eigenvalues approximately equal to 1. Including these very few eigenvectors into the coarse space V_c permits to have a very fast domain decomposition solver. We also remark that if the coarse space V_c is made of eigenvectors of G, then theoretically only one coarse correction step would be sufficient to remove the error components related to the slow eigenvectors of G.

Constructing a coarse space based on the eigenvectors of G is not always feasible, since computing these eigenvectors can be even more expensive than solving the original linear system $A\boldsymbol{u} = \boldsymbol{b}$. In this paragraph we briefly discuss a randomized approach to obtain good estimates for the eigenvectors of G and we suppose that $\rho(G) < 1$, that is the smoother G is converging. The idea we present is to approximate the image of the smoother G^r for some positive integer r. Indeed taking a sufficiently large value of r, the image of G^r contains information about the "slow" eigenvectors of G which are responsible for the slow convergence of the one-level algorithm. Motivated by this observation, we use a principal component analysis to extract information about the slowest eigenvectors from the image of G^r. We propose the following procedure

1. Consider a set of q linearly independent randomly generated vectors $\{\mathbf{x}_k\}_{k=1}^{q} \subset \mathbb{R}^{N^s}$, where N^s is the number of degrees of freedom on the product $\otimes_{j=1}^{N} S_j$, and define the matrix $X = [\mathbf{x}_1 \cdots \mathbf{x}_q]$. Here, $q \approx N_c$ and N_c is the desired dimension of the coarse space.
2. Use the vectors $\mathbf{x}_k$ as initial vectors and perform r smoothing steps to create the matrix $W = G^r X$. This computation can be performed in parallel and we assume that r is "small".
3. Compute the SVD of W: $W = U\Sigma V^\top$. This is cheap ($O(q(N^s)^2)$) because $W \in \mathbb{R}^{N^s \times q}$ is "small", since q is "small" and $\mathbf{v}_k$ are interface vectors.
4. Since the left-singular vectors (corresponding to the non-zero singular values) span the image of W, we define $V_c := \text{span}\{\mathbf{u}_j\}_{j=1}^{N_c}$ and $P := [\mathbf{u}_1, \cdots, \mathbf{u}_{2m}]$.

We emphasize that this procedure is numerically feasible since

- q is small since it is around the size of the coarse space, which also correspond to the size of the linear coarse problem involving the coarse matrix A_c.
- The number of smoothing steps r is small. Numerically we have observed that $r \approx 1-2$ for classical equations, and $r \approx 6$ for problems with jumping diffusion coefficients. Moreover, the smoothing steps can be done in parallel for the q columns of X.
- The size of the vectors $\mathbf{x}_k$ is relatively small and it is equal to the number of degrees of freedom on the substructures.
- The PCA technique is thought to be done in an off-line phase, to generate a spectral coarse space which can then be used repeatedly to solve the original linear system in a many-query context.

Numerically, we have also explored the use of the randomized SVD algorithm which shares similarities with our approach. However, we have not observed any significant advantages in terms of iteration numbers. In section 4.6, we will show that the PCA procedure permits to construct spectral coarse spaces which are extremely efficient.

4.2.2 Spectral coarse space based on the eigenvectors of G_j: convergence analysis

In this subsection, we provide a convergence analysis for a spectral coarse space which consists of eigenfunctions of the operators G_j. Our proof is restricted to the two subdomain case and we consider the decomposition introduced in Section 1.2. Since we do not have cross points, we can simplify the notation described in 4.1. First, we have $S_1 = \Gamma_2$,

$S_2 = \Gamma_1$ and $\mathbf{v} = [v_1, v_2]^\top = [\tau_1(u_1), \tau_2(u_2)]^\top = [(u_1)_{|\Gamma_2}, (u_2)_{|\Gamma_1}]^\top \in H_{00}^{\frac{1}{2}}(\Gamma_2) \times H_{00}^{\frac{1}{2}}(\Gamma_1)$. The linear system $A\mathbf{u} = \mathbf{b}$ reads

$$\begin{pmatrix} I_2 & -G_1 \\ -G_2 & I_1 \end{pmatrix} \begin{pmatrix} u_1^n \\ u_2^n \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix},$$

where I_j are the identity operators on $L^2(\Gamma_j)$. Consider the two spaces $\mathscr{H}_1 := H_{00}^{\frac{1}{2}}(\Gamma_1)$ and $\mathscr{H}_2 := H_{00}^{\frac{1}{2}}(\Gamma_2)$ and define $\mathscr{H} := \mathscr{H}_2 \times \mathscr{H}_1$. Let $\{\psi_k^1\}_{k\in\mathbb{N}}$ be a basis of $\mathscr{H}_1$ and $\{\psi_k^2\}_{k\in\mathbb{N}}$ a basis of $\mathscr{H}_2$. Let us introduce an inner product $\langle\cdot,\cdot\rangle_1$ for $\mathscr{H}_1$, an inner product $\langle\cdot,\cdot\rangle_2$ for $\mathscr{H}_2$, and define $\langle(a,b),(c,d)\rangle := \langle a,c\rangle_2 + \langle b,d\rangle_1$ for all $(a,b),(c,d) \in \mathscr{H}$. Assume that the coarse space $V_c \subset \mathscr{H}$ is the span of the basis functions $(\psi_1^2, 0), \ldots, (\psi_m^2, 0)$ and $(0, \psi_1^1), \ldots, (0, \psi_m^1)$, for a finite $m > 0$, which are orthonormal with respect to $\langle\cdot,\cdot\rangle$. The operators $P : \mathbb{R}^{2m} \to \mathscr{H}$ and $R : \mathscr{H} \to \mathbb{R}^{2m}$ are then defined as

$$\begin{aligned} P\begin{bmatrix} \mathbf{v} \\ \mathbf{w} \end{bmatrix} &:= \left[\sum_{k=1}^m (\mathbf{v})_k \psi_k^2, \quad \sum_{k=1}^m (\mathbf{w})_k \psi_k^1 \right]^\top, \\ R\begin{bmatrix} f \\ g \end{bmatrix} &:= \left[\langle \psi_1^2, f\rangle_2, \quad \cdots, \quad \langle \psi_m^2, f\rangle_2, \quad \langle \psi_1^1, g\rangle_1, \quad \cdots, \quad \langle \psi_m^1, g\rangle_1 \right]^\top, \end{aligned} \tag{4.2.7}$$

for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^m$ and any $(f,g) \in \mathscr{H}$. The restriction of A on V_c is the operator $A_c : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ given by $A_c = RAP$. It is possible to show that the matrix A_c is invertible using Theorem 4.2.2 and we refer the reader to Lemma 3.1 in [41] for a detailed proof. The S2S iteration in the error form reads as usual

$$\boldsymbol{e}^{\text{new}} = T\boldsymbol{e}^{\text{old}}, \quad T := G^{n_2}(I - PA_c^{-1}RA)G^{n_1}.$$

To provide a convergence analysis, we study the operator T and we introduce the operator norm

$$\|S\|_{\text{op}} := \sup_{\|\mathbf{v}\|_{2,\infty}=1} \|S\mathbf{v}\|_{2,\infty} \text{ for any } S \in \mathscr{L}(\mathscr{H}),$$

where $\mathscr{L}(\mathscr{H})$ is the space of linear operators on $\mathscr{H}$ and $\|\mathbf{v}\|_{2,\infty} := \max\{\|v_2\|_{\mathscr{H}_2}, \|v_1\|_{\mathscr{H}_1}\}$ with $\|v_j\|_{\mathscr{H}_j} := \langle v_j, v_j\rangle_j^{1/2}$, for $j = 1,2$ and $\mathbf{v} = (v_2, v_1) \in \mathscr{H}$. Moreover, we also define the contraction factor $\rho(T) := \lim_{n\to\infty} (\|T^n\|_{\text{op}})^{1/n}$.

We now make the further hypothesis that the interfaces Γ_1 and Γ_2 can be mapped one to the other by simple rotation, translation and scaling. This hypothesis implies that $\mathscr{H}_1 = \mathscr{H}_2 =: \mathscr{H}_0$ and we define the scalar product $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_1 = \langle\cdot,\cdot\rangle_2$. Further, we assume that there exists a set of basis functions $\{\psi_1, \psi_2, \psi_3, \ldots\} \subset \mathscr{H}_0$, orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle$, that diagonalizes the operators G_j:

$$G\begin{bmatrix} \psi_k \\ \psi_k \end{bmatrix} = \begin{bmatrix} 0 & G_1 \\ G_2 & 0 \end{bmatrix} \begin{bmatrix} \psi_k \\ \psi_k \end{bmatrix} = \begin{bmatrix} \rho_1(k)\psi_k \\ \rho_2(k)\psi_k \end{bmatrix}, \tag{4.2.8}$$

where $\rho_j(k)$ are the eigenvalues of G_j, for $j = 1,2$. Thus, the operators G_j share a common orthogonal eigenfunctions basis but they can have different eigenvalues. The coarse

space is defined as $V_c = (\text{span}\{\psi_1, \psi_2, \cdots, \psi_m\})^2$. Prolongation and restriction operators are defined as in (4.2.7). To analyze the convergence behavior, we expand the error as $\mathbf{e}^0 = \left[\sum_{j=1}^{\infty} (\mathbf{v})_j^0 \psi_j, \quad \sum_{j=1}^{\infty} (\mathbf{w})_j^0 \psi_j\right]^\top$ and study the operator norm of T.

Theorem 4.2.5 (Convergence of the S2S method- Eigenfunctions of G_j)**.** *Consider the coarse space*
$V_c = (span\{\psi_1, \psi_2, \cdots, \psi_m\})^2$ *and the operators P and R defined in* (4.2.7)*. The S2S method applied to the model problem* (4.1.1) *is a direct method for all the error components* (ψ_k, ψ_ℓ) *with* $k, \ell \leq m$*, that is* $T\left[\psi_k, \psi_\ell\right]^\top = 0$ *for all* $k, \ell \leq m$*. Moreover, if the eigenvalues* $\rho_j(k)$*,* $j = 1, 2$*, are in absolute value non-increasing functions of* k*, the contraction factor of the S2S, defined as* $\rho_{S2S}(T) := \lim_{n\to\infty} (\|T^n\|_{op})^{\frac{1}{n}}$*, is given by*

$$\rho_{S2S}(T) = \begin{cases} |\rho_1(m+1)\rho_2(m+1)|^{\frac{n_1+n_2}{2}}, \text{ if } n_1, n_2 \text{ are both even or odd,} \\ |\rho_1(m+1)\rho_2(m+1)|^{\frac{n_1+n_2-1}{2}} \max\{|\rho_1(m+1)|, |\rho_2(m+1)|\}, \text{ otherwise.} \end{cases}$$

Proof. Let us suppose that both n_1 and n_2 are even. The other cases can be treated similarly to this one. For n_1 even we define $\pi^{n_1}(k) := \rho_1^{\frac{n_1}{2}}(k)\rho_2^{\frac{n_1}{2}}(k)$ and study the action of the operator T on a vector $\left[\psi_k, \psi_\ell\right]^\top$:

$$T\begin{bmatrix}\psi_k \\ \psi_\ell\end{bmatrix} = G^{n_2}(\mathbb{I} - PA_c^{-1}RA)G^{n_1}\begin{bmatrix}\psi_k \\ \psi_\ell\end{bmatrix}.$$

We begin with the case $k \leq m$ and $\ell \leq m$. First, let us compute the action of the operator RAG^{n_1} on $\left[\psi_k, \psi_\ell\right]^\top$. Since the operators G_j are diagonalized by the basis $\{\psi_k\}_k$ using (4.2.8) one obtains $G^{n_1}\begin{bmatrix}\psi_k \\ \psi_\ell\end{bmatrix} = \begin{bmatrix}\pi^{n_1}(k)\psi_k \\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}$. The action of A on $\left[\pi^{n_1}(k)\psi_k, \pi^{n_1}(\ell)\psi_\ell\right]^\top$ is

$$A\begin{bmatrix}\pi^{n_1}(k)\psi_k \\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix} = \begin{bmatrix}I_d & -G_1 \\ -G_2 & I_d\end{bmatrix}\begin{bmatrix}\pi^{n_1}(k)\psi_k \\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix} = \begin{bmatrix}\pi^{n_1}(k)\psi_k \\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix} - \begin{bmatrix}\pi^{n_1}(\ell)\rho_1(\ell)\psi_\ell \\ \pi^{n_1}(k)\rho_2(k)\psi_k\end{bmatrix}.$$

Since A is invertible and has the form $A = \mathbb{I} - G$, the eigenvalues $\rho_j(k)$ must be different from one. Hence, the product $A\left[\pi^{n_1}(k)\psi_k, \pi^{n_1}(\ell)\psi_\ell\right]^\top \neq 0$. Now, the application of the restriction operator R on $A\left[\pi^{n_1}(k)\psi_k, \pi^{n_1}(\ell)\psi_\ell\right]^\top$ gives us

$$RA\begin{bmatrix}\pi^{n_1}(k)\psi_k \\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix} = \begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k \\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix} - \begin{bmatrix}\pi^{n_1}(\ell)\rho_1(\ell)\mathbf{e}_\ell \\ \pi_1^{n_1}(k)\rho_2(k)\mathbf{e}_k\end{bmatrix} = \Lambda\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k \\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix},$$

where $\mathbf{e}_k$ and $\mathbf{e}_\ell$ are canonical vectors in $\mathbb{R}^m$ and $\Lambda := \begin{bmatrix}I & -\rho_1(\ell)I \\ -\rho_2(k)I & I\end{bmatrix}$, with I the $m \times m$ identity matrix. We have then obtained

$$RAG^{n_1}\begin{bmatrix}\psi_k \\ \psi_\ell\end{bmatrix} = \Lambda\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k \\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}. \tag{4.2.9}$$

Now, by computing

$$A_c\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}=R\begin{bmatrix} I_d & -G_1\\ -G_2 & I_d\end{bmatrix}\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}=R\begin{bmatrix}\pi^{n_1}(k)\psi_k-\pi^{n_1}(\ell)\rho_1(\ell)\psi_\ell\\ \pi^{n_1}(\ell)\psi_\ell-\pi^{n_1}(k)\rho_2(k)\psi_k\end{bmatrix}=\Lambda\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix},$$

one obtains the action of A_c^{-1} on $\Lambda\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}$, that is

$$\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}=A_c^{-1}\Lambda\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}. \tag{4.2.10}$$

Using (4.2.9) and (4.2.10) we have

$$\begin{aligned}(\mathbb{I}-PA_c^{-1}RA)G^{n_1}\begin{bmatrix}\psi_k\\ \psi_\ell\end{bmatrix}&=\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}-PA_c^{-1}\Lambda\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}\\ &=\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}-P\begin{bmatrix}\pi^{n_1}(k)\mathbf{e}_k\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}=\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}-\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}=0.\end{aligned} \tag{4.2.11}$$

This means that the S2S method is a direct method for all the pairs (ψ_k,ψ_ℓ) with $k\le m$ and $\ell\le m$. The result for n_1 odd follows by similar calculations.

Next, let us consider the case $k>m$ and $\ell\le m$. Recalling that the basis $\{\psi_k\}_k$ is orthonormal, one has

$$RAG^{n_1}\begin{bmatrix}\psi_k\\ \psi_\ell\end{bmatrix}=R\left(\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}-\begin{bmatrix}\pi^{n_1}(\ell)\rho_1(\ell)\psi_\ell\\ \pi^{n_1}(k)\rho_2(k)\psi_k\end{bmatrix}\right)=\begin{bmatrix}0 & -\rho_1(\ell)I\\ 0 & I\end{bmatrix}\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}.$$

Similarly as before, we compute

$$A_c\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}=RA\begin{bmatrix}0\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}=R\begin{bmatrix}-\pi^{n_1}(\ell)\rho_1(\ell)\psi_\ell\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}=\begin{bmatrix}0 & -\rho_1(\ell)I\\ 0 & I\end{bmatrix}\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix},$$

which implies that

$$\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}=A_c^{-1}\begin{bmatrix}0 & -\rho_1(\ell)I\\ 0 & I\end{bmatrix}\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}.$$

Thus, we have

$$\begin{aligned}T\begin{bmatrix}\psi_k\\ \psi_\ell\end{bmatrix}&=G^{n_2}\left(\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}-PA_c^{-1}\begin{bmatrix}0 & -\rho_1(\ell)I\\ 0 & I\end{bmatrix}\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}\right)\\ &=G^{n_2}\left(\begin{bmatrix}\pi^{n_1}(k)\psi_k\\ \pi^{n_1}(\ell)\psi_\ell\end{bmatrix}-P\begin{bmatrix}0\\ \pi^{n_1}(\ell)\mathbf{e}_\ell\end{bmatrix}\right)=\begin{bmatrix}\pi^{n_1+n_2}(k)\psi_k\\ 0\end{bmatrix}.\end{aligned} \tag{4.2.12}$$

Hence for any pair (ψ_k,ψ_ℓ) with $k>m$ and $\ell\le m$, the S2S is a direct method only for the ℓ^{th} error component, which belongs to the coarse space. The component k is not affected by the coarse correction and only affected by the smoothing steps. For the remaining case $k>m$ and $\ell>m$, the same arguments as before imply that

$$T\begin{bmatrix}\psi_k\\ \psi_\ell\end{bmatrix}=G^{n_2}(\mathbb{I}-PA_c^{-1}RA)G^{n_1}\begin{bmatrix}\psi_k\\ \psi_\ell\end{bmatrix}=G^{n_2}G^{n_1}\begin{bmatrix}\psi_k\\ \psi_\ell\end{bmatrix}=\begin{bmatrix}\pi^{n_1+n_2}(k)\psi_k\\ \pi^{n_1+n_2}(\ell)\psi_\ell\end{bmatrix}. \tag{4.2.13}$$

We can now study the norm of T. To do so, we first use (4.2.11), (4.2.12) and (4.2.13), and that $\{\psi_k, \psi_\ell\}_{k,\ell}$ is a basis of $\mathscr{H}$, to write

$$T\mathbf{v} = T\begin{bmatrix}\sum_{k=1}^{\infty}\mathbf{c}_k\psi_k\\ \sum_{\ell=1}^{\infty}\mathbf{d}_\ell\psi_\ell\end{bmatrix} = T\begin{bmatrix}\sum_{k=m+1}^{\infty}\pi(k)\mathbf{c}_k\psi_k\\ \sum_{\ell=m+1}^{\infty}\pi(\ell)\mathbf{d}_\ell\psi_\ell\end{bmatrix},$$

for any $\mathbf{v}\in\mathscr{H}$. Since $|\rho_1(k)|$ and $|\rho_2(k)|$ are non-increasing functions of k, $|\pi(k)|$ is also a non-increasing function of k. Therefore, using that the basis $\{\psi_k, \psi_\ell\}_{k,\ell}$ is orthonormal, we get

$$\|T\|_{\text{op}} = \sup_{\|\mathbf{v}\|_{2,\infty}=1}\|T\mathbf{v}\|_{2,\infty} \leq \max\left(|\pi^{n_1+n_2}(k)|, |\pi^{n_1+n_2}(\ell)|\right) = |\pi^{n_1+n_2}(m+1)|.$$

This upper bound is achieved at $\mathbf{v} = [\psi_{m+1}, 0]^\top$. Hence, $\|T\|_{\text{op}} = |\pi^{n_1+n_2}(m+1)|$. Now, a similar direct calculation leads to $\|T^n\|_{\text{op}} = |\pi^{n(n_1+n_2)}(m+1)|$, which implies that $\rho_{S2S}(T) = \lim_{n\to\infty}(\|T^n\|_{\text{op}})^{1/n} = |\pi^{n_1+n_2}(m+1)|$. □

Theorem 4.2.5 shows that, similarly to the case of a coarse space based on the eigenfunctions of G, the choice of the basis functions ψ_k^j can affect drastically the convergence of the method. We conclude this section dedicated to the S2S method with two remarks and we refer the interested reader to [42] for further details. It is natural to pose the following question: given an integer N_c, which is the coarse space of dimension N_c such that the spectral radius of the S2S method is minimized? In other words, which is the coarse space of dimension N_c leading to the fastest convergence? One would be tempted to say that the spectral coarse space based on the eigenfunctions of G is optimal, as its convergence is determined by the largest eigenvalue associated to an eigenvector not included in the coarse space. However, we remark that the PCA and HEM coarse spaces, since they are not based on eigenfunctions of G and A, lead to a two-level method with substantially different eigenvalues and eigenvectors. It can happen that these coarse spaces do not map exactly the "slowest" eigenvectors of G into the kernel of the S2S method, but they can take care of them very efficiently, while still be faster on the remaining part of the spectrum. We finally remark that the substructured matrix A is not symmetric but it has eigenvalues strictly positive assuming $\sigma(G)\subset[0,1)$. In the case of highly jumping diffusion coefficients, the largest eigenvalues of G tends to one, which means that the smallest eigenvalue of A tends to zero. If one uses a coarse space which is not made of eigenfunctions of A, it can be that the coarse matrix A_c has some negative eigenvalue, which then lead to a divergent method. We studied the effects of including a perturbed eigenvector into the coarse space and we refer the interested reader to [42]. A simple numerical solution to this problem is to apply few times the smoother G to the basis of the coarse space, in such a way to make each element of the basis closer to the eigenfunctions of G.

In the following section we aim to answer the following questions: is it possible to use a two-level substructured method without providing directly a definition of the coarse space V_c? Is it possible to define a multilevel substructured domain decomposition method? The G2S method is a positive answer to these needs.

4.3 G2S method

In this section, we consider a discretization of the substructures such that each S_j is approximated by a mesh of N_j points, $j \in \mathscr{I}$. We denote the discrete substructures by $\mathscr{S}_j^{N_j}$, $j \in \mathscr{I}$ and we set $N^s := \sum_{j\in\mathscr{I}} N_j$. We then introduce finite-dimensional discretizations of the operators $G_{j,\ell}$ denoted by $G_{h,j,\ell} : \mathbb{R}^{N_j \times N_\ell}$. Similarly as in (4.1.13), we define the block operators $A_h \in \mathbb{R}^{N^s \times N^s}$ and $G_h \in \mathbb{R}^{N^s \times N^s}$ as

$$\begin{aligned} &[A_h]_{j,j} = I_{h,j},\ [A_h]_{j,\ell} = -G_{h,j,\ell},\ j,k \in \mathscr{I},\ j \neq k, \\ &[G_h]_{j,j} = 0,\ [G_h]_{j,\ell} = G_{h,j,\ell},\ j,k \in \mathscr{I},\ j \neq k, \end{aligned} \tag{4.3.1}$$

where $I_{h,j} \in \mathbb{R}^{N_j \times N_j}$ are identity matrices. Notice that $A_h = I_h - G_h$, where $I_h = \mathrm{diag}(I_{h,1}, \ldots, I_{h,N})$. Therefore, the substructured problem $A\mathbf{v} = \mathbf{b}$ becomes

$$A_h \mathbf{v} = \mathbf{b}_h,$$

where $\mathbf{b}_h = [\mathbf{b}_{h,1}, \ldots, \mathbf{b}_{h,N}]$, and the PSM is then

$$\mathbf{v}^n = G_h \mathbf{v}^{n-1} + \mathbf{b}_h. \tag{4.3.2}$$

We emphasize that the matrices G_h and A_h are never assembled explicitly and their action on given vectors is computed directly. Note that the action of $G_{h,j,\ell}$ on a given vector requires a subdomain solve which is performed exactly. Concerning the invertibility of A_h it is sufficient to assume $\rho(G_h) < 1$, that is the discrete PSM converges.

To define a two-grid method, we introduce coarser discretizations of the substructures $\mathscr{S}_j^{M_j}$, $j \in \mathscr{I}$, where $M_j < N_j$ points. The total number of discrete coarse points is $M^s := \sum_{j\in\mathscr{I}} M_j$. For each $j \in \mathscr{I}$ we introduce restriction and prolongation matrices $R_j \in \mathbb{R}^{M_j \times N_j}$ and $P_j \in \mathbb{R}^{N_j \times M_j}$. These could be classical interpolation operators used, e.g., in multigrid methods. If for example $\mathscr{S}_j$ is a one-dimensional interval, then the full weighting restriction matrix and the prolongation matrix are

$$R_j := \frac{1}{2}\begin{bmatrix} \frac{1}{2} & 1 & \frac{1}{2} & & & \cdots & & & \\ & & \frac{1}{2} & 1 & \frac{1}{2} & \cdots & & & \\ & & & & \frac{1}{2} & \cdots & & & \\ & & & & & \cdots & & & \\ & & & & & \cdots & \frac{1}{2} & 1 & \frac{1}{2} \end{bmatrix}, \quad P_j := 2R_j^\top. \tag{4.3.3}$$

We remark that using P_j we are assuming that the function we interpolate is zero at the boundary of the substructure, and this holds true since each function v_j belongs to $H_{00}^{\frac{1}{2}}(S_j)$. The global restriction and prolongation matrices are defined block-wise as $R := \mathrm{diag}(R_1, \ldots, R_N) \in \mathbb{R}^{M^s \times N^s}$ and $P := \mathrm{diag}(P_1, \ldots, P_N) \in \mathbb{R}^{N^s \times M^s}$. The restriction of A_h on the coarse level is then defined as $A_{2h} := RA_hP$. Notice that this matrix can be either precomputed exactly or assembled in an approximate way. The G2S procedure is defined by Algorithm 3, replacing the continuous operators with their discrete counterparts, and

with the specific choice of the geometrical restriction and prolongation operators. One iteration of the G2S method in the error form can therefore be written as

$$\mathbf{e}^{\text{new}} = T_h \mathbf{e}^{\text{old}} \text{ with } T_h := G_h^{n_2}(\mathbb{I}_h - P A_{2h}^{-1} R A_h) G_h^{n_1}. \tag{4.3.4}$$

We conclude this paragraph with some remarks.

- The G2S method is a classical two-grid type iteration, but instead of having the classical grids in volume, we consider two discrete levels on the substructures. This has the advantage of performing all restriction and interpolation operations on interfaces which are one dimension smaller than the original domain Ω. Thus, dealing with a problem in 3D, the G2S method requires to perform restriction and interpolation operations in 2D.
- We refer the interested reader to [43, Section 3.2.3] for a discussion of analogies and differences between the G2S method and a two-grid volume method. In particular, it is shown that the G2S method is spectrally equivalent to a two-grid method in volume which is very different from a standard two-grid method in volume using PSM as a smoother.
- We insist that the G2S method does not require the explicit construction of a coarse space V_c, but it exploits directly a discretization of the interfaces. However, it is possible to show that the G2S method coincides with a S2S method with a precise choice for the spectral coarse space, see [43, Section 3.2.1].
- The dimension of the coarser problem A_{2h} is equal to the number of unknowns on the substructures on the coarse mesh, and thus it is generally very small if compared to the size of a coarse matrix of a two-grid method in volume. Thus, generally less levels are required to obtain a sufficiently small coarse matrix to invert.
- If needed, it is clear that a simple recursion allows us to embed the G2S method into a multi-grid framework. We discuss further implementation details in Section 4.4.
- The G2S method permits to have more freedom in the choice of the grids with respect to volume methods. Since the coarse correction is added only on the substructures, one could use a coarse grid which is coarser and coarser as the distance from the substructures increases. A stretched mesh is not natural in the volume case, as it would lead to a poor volume coarse correction away from the substructures and would require a more sophisticated volume interpolation step.

4.3.1 Convergence analysis

In this section, we analyze the convergence behavior of the G2S method. To do so, we consider our model problem (3.1.1) and assume the two-subdomain decomposition depicted in Figure 4.1. We recall that such a decomposition implies $S_1 = \Gamma_2$ and $S_2 = \Gamma_1$ which are segments of same length $\widetilde{L}$. For a given $\ell \in \mathbb{N}^+$, $\ell \geq 2$, we discretize (3.1.1) using

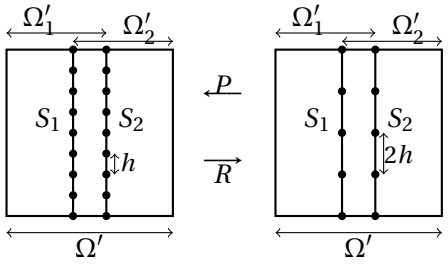


Figure 4.1: Two-subdomain decomposition, substructures and their discretizations.

a uniform grid of $N_h = 2^\ell - 1$ points on each substructure so that the grid size is $h = \frac{\widetilde{L}}{N_h+1}$. We also introduce a coarser mesh of $N_c = 2^{\ell-1} - 1$ points on each substructure and mesh size $h_c = \frac{1}{N_c+1}$. We define the geometric prolongation operator $P_{2h}^h \in \mathbb{R}^{2N_h \times 2N_c}$ as $P_{2h}^h := \text{diag}(\widetilde{P}, \widetilde{P})$ and the geometric restriction operator $R_{2h}^h \in \mathbb{R}^{2N_c \times 2N_h}$ as $R_{2h}^h := \text{diag}(\widetilde{R}, \widetilde{R})$. $\widetilde{R}$ and $\widetilde{P}$ correspond to the matrices P_j, R_j in (4.3.3).

We suppose that the operators $G_{h,1}$ and $G_{h,2}$ have as eigenvectors the discrete Fourier modes given by $(\boldsymbol{\psi}_k)_j = \sin(k\pi hj)$, for $j,k = 1,\dots,N_h$. The eigenvalues are $\rho_j(k)$, $k = 1,\dots,N_h$, $j = 1,2$. It is well-known that the actions of $\widetilde{R}$ and $\widetilde{P}$ on the combination of a low-frequency mode $\boldsymbol{\psi}_k$ with its high-frequency companion $\boldsymbol{\psi}_{\widetilde{k}}$, with $\widetilde{k} = N_h - k + 1$, are

$$\widetilde{R}\begin{bmatrix}\boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}}\end{bmatrix} = \boldsymbol{\phi}_k\begin{bmatrix}c_k^2 & -s_k^2\end{bmatrix}, \quad \widetilde{P}\boldsymbol{\phi}_k = (c_k^2\boldsymbol{\psi}_k - s_k^2\boldsymbol{\psi}_{\widetilde{k}}) = \begin{bmatrix}\boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}}\end{bmatrix}\begin{bmatrix}c_k^2 \\ -s_k^2\end{bmatrix}, \tag{4.3.5}$$

where $c_k = \cos(k\pi\frac{h}{2})$, $s_k = \sin(k\pi\frac{h}{2})$ for $k = 1,\dots,N_c$ and $(\boldsymbol{\phi}_k)_j = \sin(k\pi 2hj)$, for $k = 1,\dots,\frac{N_h+1}{2} - 1$ and $j = 0,\dots,\frac{N_h+1}{2}$; see, e.g., [109]. The vectors $\boldsymbol{\phi}_k$ are Fourier modes on the coarse grid. Before studying the convergence of the method, we discuss the well-possessedness of the G2S iteration (4.3.4).

Lemma 4.3.1 (Invertibility of A_{2h})**.** *Assume that $\rho_1(k), \rho_2(k) \in [0,1)$ for all k and that $\rho_1(k) \geq \rho_1(\widetilde{k})$ and $\rho_2(k) \geq \rho_2(\widetilde{k})$ for any $k = 1,\dots,N_c$ and $\widetilde{k} = N_h - k + 1$. The matrix $A_{2h} := R_{2h}^h A_h P_{2h}^h \in \mathbb{R}^{2N_c \times 2N_c}$ has full rank.*

Proof. We refer the interested reader to the proof of Lemma 3.2 [43] □

We now derive sharp estimates for the spectral radius of T_h. The first step is this technical Lemma.

Lemma 4.3.2. *Consider the G2S matrix $T_h := G_h^{n_2}(I - P_{2h}^h A_{2h}^{-1} R_{2h}^h A_h) G_h^{n_1}$. The action of T_h on $\begin{bmatrix}\boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}}\end{bmatrix}$ is given by*

$$T_h\begin{bmatrix}\boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}}\end{bmatrix}\widetilde{G}_k, \tag{4.3.6}$$

where $\widetilde{G}_k := D_{n_2}(k)(D_{n_1}(k) - V(k)\Lambda_2^{-1}(k)\Lambda_1(k))$ *with*

$$\Lambda_1(k) := V(k)^\top H(k) D_{n_1}(k), \qquad \Lambda_2(k) := V(k)^\top H(k) V(k),$$

$$V(k) := \begin{bmatrix} c_k^2 & 0 \\ -s_k^2 & 0 \\ 0 & c_k^2 \\ 0 & -s_k^2 \end{bmatrix}, \qquad H(k) := \begin{bmatrix} 1 & 0 & -\rho_1(k) & 0 \\ 0 & 1 & 0 & -\rho_1(\widetilde{k}) \\ -\rho_2(k) & 0 & 1 & 0 \\ 0 & -\rho_2(\widetilde{k}) & 0 & 1 \end{bmatrix},$$

and $D_n(k)$ *is given by*

$$D_n(k) := \begin{bmatrix} \pi(k)^n & 0 & 0 & 0 \\ 0 & \pi(\widetilde{k})^n & 0 & 0 \\ 0 & 0 & \pi(k)^n & 0 \\ 0 & 0 & 0 & \pi(\widetilde{k})^n \end{bmatrix}, \quad D_n(k) := \begin{bmatrix} 0 & 0 & \pi_{21}(k,n) & 0 \\ 0 & 0 & 0 & \pi_{21}(\widetilde{k},n) \\ \pi_{12}(k,n) & 0 & 0 & 0 \\ 0 & \pi_{12}(\widetilde{k},n) & 0 & 0 \end{bmatrix}$$

for n even and for n odd, respectively, whose entries are $\pi(k) := (\rho_1(k)\rho_2(k))^{1/2}$, $\pi_{12}(k,n) := \rho_1(k)^{\frac{n-1}{2}}\rho_2(k)^{\frac{n+1}{2}}$, *and* $\pi_{21}(k,n) := \rho_1(k)^{\frac{n+1}{2}}\rho_2(k)^{\frac{n-1}{2}}$.

Proof. We consider the case in which both n_1 and n_2 are even. The other cases can be obtained by similar arguments. Since n_1 is even, we have that

$$G_h^{n_1} = \begin{bmatrix} (G_{h,1}G_{h,2})^{n_1/2} & 0 \\ 0 & (G_{h,2}G_{h,1})^{n_1/2} \end{bmatrix}.$$

Because of the relation $(G_{h,1}G_{h,2})^{n_1/2}\boldsymbol{\psi}_k = (G_{h,2}G_{h,1})^{n_1/2}\boldsymbol{\psi}_k = \pi^{n_1}(k)\boldsymbol{\psi}_k$, where $\pi(k) := (\rho_1(k)\rho_2(k))^{1/2}$, we get

$$G_h^{n_1}\begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} \begin{bmatrix} \pi(k) & 0 & 0 & 0 \\ 0 & \pi(\widetilde{k}) & 0 & 0 \\ 0 & 0 & \pi(k) & 0 \\ 0 & 0 & 0 & \pi(\widetilde{k}) \end{bmatrix}^{n_1} = \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} D_{n_1}(k).$$

Similarly, we obtain that $G_h^{n_2}\begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} D_{n_2}(k)$. Moreover, direct calculations reveal that

$$A_h\begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} \begin{bmatrix} 1 & 0 & -\rho_1(k) & 0 \\ 0 & 1 & 0 & -\rho_1(\widetilde{k}) \\ -\rho_2(k) & 0 & 1 & 0 \\ 0 & -\rho_2(\widetilde{k}) & 0 & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} H(k) \tag{4.3.7}$$

and

$$R_{2h}^h\begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \begin{bmatrix} c_k^2 & -s_k^2 & 0 & 0 \\ 0 & 0 & c_k^2 & -s_k^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} V(k)^\top, \tag{4.3.8}$$

where we used (4.3.5). It follows that $R_{2h}^h A_h G_h \begin{bmatrix} \boldsymbol{\psi}_k \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \Lambda_1(k)$. Let us study the action of the coarse matrix A_{2h} on $\begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix}$. We use (4.3.5), (4.3.7) and (4.3.8) to write

$$\begin{aligned} A_{2h} \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} &= R_{2h}^h A_h P_{2h}^h \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} = R_{2h}^h A_h \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} V(k) \\ &= R_{2h}^h \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} H(k) V(k) = \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} V(k)^\top H(k) V(k). \end{aligned}$$

Thus, we have $A_{2h} \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \Lambda_2(k)$. Hence, recalling Lemma 4.3.1 we get

$$\begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} = A_{h,c}^{-1} \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \Lambda_2(k). \tag{4.3.9}$$

A direct calculation reveals that the eigenvalues of $\Lambda_2(k)$ are

$$\lambda_{1,2} = c_k^4 + s_k^4 \pm \sqrt{(c_k^4 \rho_1(k) + s_k^4 \rho_1(\widetilde{k}))(c_k^4 \rho_2(k) + s_k^4 \rho_2(\widetilde{k}))},$$

and they are nonzero for $k = 1, \ldots, N_c$. Hence, $\Lambda_2(k)$ is invertible and, using (4.3.9), we get

$$A_{h,c}^{-1} \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \Lambda_1(k) = A_{h,c}^{-1} \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \Lambda_2(k) \Lambda_2^{-1}(k) \Lambda_1(k) = \begin{bmatrix} \boldsymbol{\phi}_k & 0 \\ 0 & \boldsymbol{\phi}_k \end{bmatrix} \Lambda_2^{-1}(k) \Lambda_1(k),$$

Summarizing our results and using the definition of T_h, we conclude that

$$T_h \begin{bmatrix} \boldsymbol{\psi}_k \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\psi}_k \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} D_{n_2}(k) \left[D_{n_1}(k) - \begin{bmatrix} c_k^2 & 0 \\ -s_k^2 & 0 \\ 0 & c_k^2 \\ 0 & -s_k^2 \end{bmatrix} \Lambda_2^{-1}(k) \Lambda_1(k) \right]$$

and our claim follows. □

Using Lemma 4.3.2, it is possible to factorize the iteration matrix T_h. This factorization is obtained in the following theorem.

Theorem 4.3.3 (Factorization of the iteration matrix T_h)**.** *There exists an invertible matrix Q such that $T_h = Q \widetilde{G} Q^{-1}$, where the G2S iteration matrix T_h is defined in Lemma 4.3.2 and*

$$\widetilde{G} = \begin{bmatrix} \widetilde{G}_1 & & & & & \\ & \ddots & & & & \\ & & \widetilde{G}_{N_c} & & & \\ & & & \gamma_1(\frac{N_h+1}{2}) & & \\ & & & & \gamma_2(\frac{N_h+1}{2}) & \end{bmatrix},$$

where the matrices $\widetilde{G}_k \in \mathbb{R}^{4\times 4}$ are defined in Lemma 4.3.2 and $\gamma_j(\frac{N_h+1}{2})$ depend on n_1, n_2 and the eigenvalues $\rho_j(\frac{N_h+1}{2})$ of $G_{h,j}$, for $h = 1, 2$.

Proof. We define the invertible matrix

$$Q = \begin{bmatrix} \boldsymbol{\psi}_1 & \boldsymbol{\psi}_{N_h} & 0 & 0 & \cdots & \boldsymbol{\psi}_{N_c} & \boldsymbol{\psi}_{N_c+2} & 0 & 0 & \boldsymbol{\psi}_{\frac{N_h+1}{2}} & 0 \\ 0 & 0 & \boldsymbol{\psi}_1 & \boldsymbol{\psi}_{N_h} & \cdots & 0 & 0 & \boldsymbol{\psi}_{N_c} & \boldsymbol{\psi}_{N_c+2} & 0 & \boldsymbol{\psi}_{\frac{N_h+1}{2}} \end{bmatrix}.$$

Equation (4.3.6) says that $T_h \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_k & \boldsymbol{\psi}_{\widetilde{k}} \end{bmatrix} \widetilde{G}_k$, for every $k = 1,\dots,N_c$ and $\widetilde{k} = N_h - k - 1$. Moreover, notice that the frequency $\boldsymbol{\psi}_{\frac{N_h+1}{2}}$ is mapped to zero by the restriction operator, $R_{2h}^h \begin{bmatrix} \boldsymbol{\psi}_{\frac{N_h+1}{2}} & 0 \\ 0 & \boldsymbol{\psi}_{\frac{N_h+1}{2}} \end{bmatrix} = 0$, and we get

$$T_h \begin{bmatrix} \boldsymbol{\psi}_{\frac{N_h+1}{2}} & 0 \\ 0 & \boldsymbol{\psi}_{\frac{N_h+1}{2}} \end{bmatrix} = G_h^{n_2} G_h^{n_1} \begin{bmatrix} \boldsymbol{\psi}_{\frac{N_h+1}{2}} & 0 \\ 0 & \boldsymbol{\psi}_{\frac{N_h+1}{2}} \end{bmatrix} = \begin{bmatrix} \gamma_1(\frac{N_h+1}{2})\boldsymbol{\psi}_{\frac{N_h+1}{2}} & 0 \\ 0 & \gamma_2(\frac{N_h+1}{2})\boldsymbol{\psi}_{\frac{N_h+1}{2}} \end{bmatrix},$$

where the expressions of $\gamma_1(\frac{N_h+1}{2})$ and $\gamma_2(\frac{N_h+1}{2})$ depend on n_1 and n_2. For instance if n_1+n_2 is an even number, then $\gamma_1(\frac{N_h+1}{2}) = \gamma_2(\frac{N_h+1}{2}) := (\rho_1(\frac{N_h+1}{2})\rho_2(\frac{N_h+1}{2}))^{\frac{n_1+n_2}{2}}$. Hence, we conclude that $T_h Q = Q\widetilde{G}$ and our claim follows. □

The factorization of T_h proved in Theorem 4.3.3 allows one to obtain precise convergence results of a G2S method. Clearly, an optimal result would be a direct calculation of the spectral radii of the matrices $\widetilde{G}_k$. However, this is in general a difficult task that requires cumbersome calculations. Nevertheless, in Theorem 4.3.4 we are capable to obtain an explicit expression for the spectral radii of $\widetilde{G}_k$ under some reasonable assumptions that are in general satisfied in case of Schwarz methods. Notice also that Theorem 4.3.4 guarantees that only one (pre- or post-) smoothing step is necessary for the G2S method to converge.

Theorem 4.3.4. *Assume that* $1 > \rho_1(k) = \rho_2(k) = \rho(k) \geq 0$ *for any* k *and that* $\rho(k)$ *is a decreasing function of* k*. The convergence factor of the G2S method is*

$$\rho_{G2S}(T_h) = \max_{k\in\{1,\dots,N_c,\frac{N_h+1}{2}\}} \left(\frac{c_k^4(1-\rho(k))\rho(\widetilde{k})^{n_1+n_2} + s_k^4(1-\rho(\widetilde{k}))\rho(k)^{n_1+n_2}}{c_k^4(1-\rho(k)) + s_k^4(1-\rho(\widetilde{k}))} \right) < 1.$$

Proof. The convergence factor of the G2S is given by the spectral radius of the iteration matrix T_h. Theorem 4.3.3 implies that

$$\rho_{G2S}(T_h) = \max\Big\{ \max_{k\in\{1,\dots,N_c\}} \rho(\widetilde{G}_k), \gamma_1\left(\frac{N_h+1}{2}\right), \gamma_2\left(\frac{N_h+1}{2}\right) \Big\}.$$

G2S	G2S C.C.	Volume two-level	Volume C.C.
$\mathbf{v}^{n+\frac{1}{2}} = G_h\mathbf{v}^n + \mathbf{b}_h$	$O(N_{\text{sub}}^3)$	$\mathbf{u}_v^{n+\frac{1}{2}} = N\mathbf{u}_v^n + M^{-1}\mathbf{b}_v$	$O(N_{\text{sub}}^3)$
$\mathbf{r}^{n+\frac{1}{2}} = \mathbf{b}_h - A_h\mathbf{v}^{n+\frac{1}{2}}$	$O(N_{\text{sub}}^3)$	$\mathbf{r}_v^{n+\frac{1}{2}} = \mathbf{b}_v - A_v\mathbf{u}_v^{n+\frac{1}{2}}$	$O((N^v)^2)$
$\mathbf{v}_c^{n+1} = A_{2h}^{-1}(R\mathbf{r}^{n+\frac{1}{2}})$	$O((M^s)^{\gamma_s})$	$\mathbf{u}_{vc}^{n+1} = A_{vc}^{-1}(R_v\mathbf{r}_v^{n+\frac{1}{2}})$	$O((M^v)^{\gamma_v})$
$\mathbf{v}^{n+1} = \mathbf{v}^{n+\frac{1}{2}} + P\mathbf{v}_c^{n+1}$	$O(M^sN^s)$	$\mathbf{u}_v^{n+1} = \mathbf{u}_v^{n+\frac{1}{2}} + P_v\mathbf{u}_{vc}^{n+1}$	$O(M^vN^v)$

Table 4.1: Computational cost (C.C.) per iteration.

Regardless of the values of n_1 and n_2, direct calculations show that the matrices $\widetilde{G}_k$ have four eigenvalues:

$$\begin{aligned}
\lambda_1(k) = \lambda_2(k) &= 0,\\
|\lambda_3(k)| &= \frac{c_k^4(1-\rho(k))\rho(\widetilde{k})^{n_1+n_2} + s_k^4(1-\rho(\widetilde{k}))\rho(k)^{n_1+n_2}}{c_k^4(1-\rho(k)) + s_k^4(1-\rho(\widetilde{k}))},\\
|\lambda_4(k)| &= \frac{c_k^4(1+\rho(k))\rho(\widetilde{k})^{n_1+n_2} + s_k^4(1+\rho(\widetilde{k}))\rho(k)^{n_1+n_2}}{c_k^4(1+\rho(k)) + s_k^4(1+\rho(\widetilde{k}))}.
\end{aligned}$$

Moreover, we observe that

$$|\lambda_3(k)| - |\lambda_4(k)| = \frac{2c_k^4s_k^4(\rho(k)-\rho(\widetilde{k}))(\rho(k)^{n_1+n_2} - \rho(\widetilde{k})^{n_1+n_2})}{((\rho(k)+1)c_k^4 + s_k^4(\rho(\widetilde{k})+1))((1-\rho(k))c_k^4 + s_k^4(1-\rho(\widetilde{k}))} \geq 0,$$

where we used the monotonicity of $\rho(k)$. On the other hand, since $\rho_1(k) = \rho_2(k) = \rho(k)$, we have $\gamma_1(\frac{N_h+1}{2}) = \gamma_2(\frac{N_h+1}{2}) = \rho(\frac{N_h+1}{2})^{n_1+n_2}$. Therefore we have that

$$\max\left\{\max_{k\in\{1,\dots,N_c\}}\rho(\widetilde{G}_k), \rho\left(\frac{N_h+1}{2}\right)^{n_1+n_2}\right\} = \max\left\{\max_{k\in\{1,\dots,N_c\}}|\lambda_3(k)|, \rho\left(\frac{N_h+1}{2}\right)^{n_1+n_2}\right\},$$

and the result follows by observing that $\lambda_3\left(\frac{N_h+1}{2}\right) = \rho\left(\frac{N_h+1}{2}\right)^{n_1+n_2}$, since $\rho(\widetilde{k}) = \rho(k)$ for $k = \frac{N_h+1}{2}$. □

4.4 Implementation details of two-level substructured methods

In this section, we study the computational costs (C.C.) of one iteration of the G2S and of a two-grid method in volume which uses the same smoother. Let N^v be the size of the volume problem and N^s the size of the substructured problem ($N^s \ll N^v$). The size of each subdomain is N_{sub}. The coarse spaces are of dimension M^s for the G2S method and M^v for the volume method. The restriction and prolongation operators in volume are denoted by R_v and P_v. For simplicity we assume $n_1 = 1$, $n_2 = 0$. The computational costs of one iteration are reported in Table 4.1. For simplicity, we assume that restriction and prolongation operations are classical matrix-vector products. Since the dimension

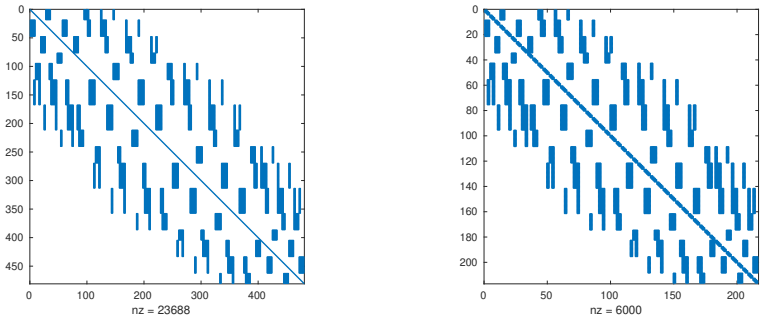


Figure 4.2: Sparsity pattern for matrices A_h (left) and A_{2h} (right).

of the substructured coarse space is smaller, the G2S could require much less computational effort in the solution of the coarse problem. However, we remark that the coarse matrix A_{2h} is typically block-dense, where the block structure is related to the connectivity among the subdomains. We report in Figure 4.2 the sparsity patter of A_h and A_{2h} for a regular decomposition of a square into 4x4 subdomains. On the other hand, A_{vc} is typically a sparse matrix and the sparsity pattern depends on the discretization method. In both cases, there are sophisticated algorithms for the solution of the corresponding linear systems, and thus we use two parameters γ_s and γ_v to indicate the computational cost of the coarse solvers. Assuming that the two coarse problems have the same, negligible, computational cost, Table 4.1 shows that in both cases the dominating costs correspond to the application of the smoothing operators. This application costs $O(N_{\text{sub}}^3)$ for both approaches. However, the G2S method requires twice a cost of order $O(N_{\text{sub}}^3)$, since the computation of the residual involves the substructured matrix A_h and thus it requires subdomain solves. If we could avoid this extra cost, then the G2S methods would be faster than a volumetric method, since all other operations are performed on arrays of much smaller sizes. Furthermore, we remark that the G2S method requires in general less iterations than the corresponding method in volume as we will show in Section 4.6. To avoid the two applications of the smoother in the G2S method, we exploit the special form of the matrix $A_h = I_h - G_h$ and propose two new versions of Algorithm 3. These are called G2S-B1 and G2S-B2 and given by Algorithms 4 and 5. These substructured algorithms require only one smoothing step per iteration and thus they are potentially cheaper than a two-grid method in volume using the same smoother. We remark that the G2S-B1 and G2S-B2 require to store the matrix $\widetilde{P} := G_h P$. This matrix is anyway computed in a precomputation phase to assemble the coarse matrix $A_{2h} = RA_hP = RP - RG_hP = RP - R\widetilde{P}$. Hence no extra cost is required. Moreover, we now prove that G2S and G2S-B1 are equivalent and they have the same spectral properties of G2S-B2.

Theorem 4.4.1 (Equivalence between G2S, G2S-B2 and G2S-B1)**.** *Algorithm 5 and Algorithm 4 have the same convergence behavior. Moreover,*

(a) Algorithm 4 generates the same iterates of Algorithm 3.

Algorithm 4: G2S-B1

Require: $\mathbf{v}^0$ and $\widehat{P} = G_h P$.

1: $\mathbf{v}^1 = G_h\mathbf{v}^0 + \mathbf{b}_h$,
2: $\mathbf{w} = G_h\mathbf{v}^1$,
3: $\mathbf{r} = \mathbf{b}_h - \mathbf{v}^1 + \mathbf{w}$,
4: $\mathbf{v}_c = A_{2h}^{-1} R\mathbf{r}$,
5: $\mathbf{v}^0 = \mathbf{v}^1 + P\mathbf{v}_c$,
Iterations:
6: $\mathbf{v}^1 = \mathbf{w} + \widehat{P}\mathbf{v}_c + \mathbf{b}_h$,
7: $\mathbf{w} = G_h\mathbf{v}^1$,
8: $\mathbf{r} = \mathbf{b}_h - \mathbf{v}^1 + \mathbf{w}$,
9: $\mathbf{v}_c = A_{2h}^{-1} R\mathbf{r}$,
10: $\mathbf{v}^0 = \mathbf{v}^1 + P\mathbf{v}_c$,
11: Repeat 6 to 10 until convergence.

Algorithm 5: G2S-B2

Require: $\mathbf{v}^0$ and $\widehat{P} = G_h P$.

1: $\mathbf{v} = G_h\mathbf{v}^0$,
2: $\mathbf{r} = \mathbf{b}_h - \mathbf{v}^0 + \mathbf{v}$,
3: $\mathbf{v}_c = A_{2h}^{-1} R\mathbf{r}$,
4: $\mathbf{v}^0 = \mathbf{v} + \widehat{P}\mathbf{v}_c + \mathbf{b}_h$,
5: Repeat 1 to 5 until convergence.

(b) Algorithm 5 corresponds to the stationary iterative method

$$\boldsymbol{\nu}^n = G_h(\mathbb{I}_h - PA_{2h}^{-1}RA_h)\boldsymbol{\nu}^{n-1} + \widetilde{M}\boldsymbol{b}_h,$$

where $G_h(\mathbb{I}_h - PA_{2h}^{-1}RA_h)$ is the iteration matrix and $\widetilde{M}$ the relative preconditioner.

Proof. For simplicity, we suppose to work with the error equation and thus $\mathbf{b}_h = 0$. We call $\widetilde{\mathbf{v}}^0$ the output of the first five steps of Algorithm 4 and $\widehat{\mathbf{v}}^0$ the output of Algorithm 3. Then given an initial guess $\mathbf{v}^0$, we have

$$\begin{aligned}\widetilde{\mathbf{v}}^0 &= \mathbf{v}^1 + P\mathbf{v}_c = \mathbf{v}^1 + PA_{2h}^{-1}R(-\mathbf{v}^1 + \mathbf{w}) \\ &= G_h\mathbf{v}^0 + PA_{2h}^{-1}R(-A_hG_h\mathbf{v}^0) = (\mathbb{I}_h - PA_{2h}^{-1}RA_h)G_h\mathbf{v}^0 = \widehat{\mathbf{v}}^0.\end{aligned}$$

Similar calculations show that also steps 6-10 of G2S-B1 are equivalent to an iteration of 3. For the second part of the Theorem, we write one iteration of Algorithm 5 as

$$\mathbf{v}^1 = \mathbf{v} + \widehat{P}\mathbf{v}_c = G\mathbf{v}^0 + G_hPA_{2h}^{-1}R(-A_h\mathbf{v}^0) = G_h(\mathbb{I}_h - PA_{2h}^{-1}RA_h)\mathbf{v}^0.$$

Hence, Algorithm 5 performs a post-smoothing step instead of a pre-smoothing step as Algorithm 4 does. The method still has the same convergence behavior since the matrices $G_h(\mathbb{I}_h - PA_{2h}^{-1}RA_h)$ and $(\mathbb{I}_h - PA_{2h}^{-1}RA_h)G_h$ have the same eigenvalues. □

Notice that Algorithm 4 requires for the first iteration two applications of the smoothing operator G_h, namely two subdomains solves. The next iterations, given by Steps 6-10, need only one application of the smoothing operator G_h. Theorem 4.4.1 (a) shows that Algorithm 4 is equivalent to Algorithm 3. This means that each iteration after the first one of Algorithm 4 is computationally less expensive than one iteration of a volume two-level

DD method. Since two-level DD methods perform generally few iterations, it could be important to get rid of the expensive first iteration. For this reason, we introduce Algorithm 5, which overcome the problem of the first iteration. Theorem 4.4.1 (b) guarantees that Algorithm 5 is exactly an S2S method with no pre-smoothing and one post-smoothing step. Moreover, it has the same convergence behavior of Algorithm 4. These implementation tricks can be readily generalized to a general number of pre- and post-smoothing steps and they can also be applied to the S2S method.

Concerning the specific implementation details for the G2S, we remark that one can lighten the off-line assembly of the matrix $A_{2h} = R^h_{2h} A_h P^h_{2h}$, using instead the matrix

$$\widetilde{A}_{2h} := \begin{bmatrix} I_{2h,2} & -G_{2h,1} \\ -G_{2h,2} & I_{2h,1} \end{bmatrix}, \tag{4.4.1}$$

which corresponds to a direct discretization of A on the coarse level.

4.5 Extension to a multilevel framework

Two-grid methods in volume are not very efficient for the solution of large problems due to the dimension of the coarse problem which is roughly about one fourth of the dimension of the full problem in two dimensions. We have already remarked that the size of the substructured coarse matrix is usually quite small, since it corresponds to the number of unknowns on a coarse discretization of the substructures. Nevertheless, there can be problems for which the direct solution of the coarse problem is inconvenient also in a substructured framework. For instance, if we considered several subdomains, then we would have several substructures and therefore the size of the substructured coarse matrix increases.

The G2S method is suitable to be generalized to multilevel framework following a classical multigrid strategy [109]. Given a sequence of grids on the substructures labeled from the coarsest to the finest by $\{\ell_{\min}, \ell_{\min}+1, \dots, \ell_{\max}\}$, we denote by $P^\ell_{\ell-1}$ and $R^\ell_{\ell-1}$ the interpolation and restriction operators between grids ℓ and $\ell-1$. To build the substructured matrices on the different grids we have two possible choices. The first one corresponds to the standard Galerkin projection. Letting $A_{\ell_{\max}}$ be the substructured matrix on the finest grid, we can define the coarse matrices $A_\ell := R^{\ell+1}_\ell A_{\ell+1} P^{\ell+1}_\ell$, for $\ell \in \{\ell_{\min}, \ell_{\min}+1, \dots, \ell_{\max}-1\}$. The second choice consists in defining A_ℓ directly as the discretization of (4.1.13) on the grid labeled by ℓ, and corresponds exactly to (4.4.1) for the two-grid case. The two choices are not equivalent. On the one hand, the Galerkin approach leads to a faster method in terms of iterations. However, the Galerkin matrices A_ℓ do not have the block structure as in (4.1.13). For instance, $A_{\ell_{\max}-1} = R^{\ell_{\max}}_{\ell_{\max}-1} A_{\ell_{\max}} P^{\ell_{\max}}_{\ell_{\max}-1} = R^{\ell_{\max}}_{\ell_{\max}-1} P^{\ell_{\max}}_{\ell_{\max}-1} - R^{\ell_{\max}}_{\ell_{\max}-1} G_{\ell_{\max}} P^{\ell_{\max}}_{\ell_{\max}-1}$. Thus, the identity matrix is replaced by the sparse matrix $R^{\ell_{\max}}_{\ell_{\max}-1} P^{\ell_{\max}}_{\ell_{\max}-1}$. On the other hand, defining A_ℓ directly on the current grid ℓ as in (4.4.1) leads to a minor increase of the iteration number, but it permits to preserve the original block-diagonal structure (which is important if one wants to use G2S-B1 and G2S-B2). The difference be-

Algorithm 6: Geometric multilevel substructured DD method - GMS($\mathbf{u}^0$,$\mathbf{b}$,ℓ)

if $\ell = \ell_{\min}$, **then**
  set $\mathbf{v}^0 = A_{\ell_{\min}}^{-1}\mathbf{b}$. (direct solver)
else
  $\mathbf{v}^n = G_\ell(\mathbf{v}^{n-1},\mathbf{b})$, $n = 1,\ldots,n_1$ (DD pre-smoothing steps)
  $\mathbf{r} = \mathbf{b} - A_\ell \mathbf{v}^{n_1}$ (compute the residual)
  $\mathbf{v}_c$ =GMS($\mathbf{0}, R_{\ell-1}^{\ell}\mathbf{r}, \ell-1$). (recursive call)
  $\mathbf{v}^0 = \mathbf{v}^{n_1} + P_{\ell-1}^{\ell}\mathbf{v}_c$ (coarse correction)
  $\mathbf{v}^n = G_\ell(\mathbf{v}^{n-1},\mathbf{b})$, $n = 1,\ldots,n_2$ (DD post-smoothing steps)
  Set $\mathbf{v}^0 = \mathbf{v}^{n_2}$ (update)
end if
return $\mathbf{u}^0$.

tween the two approaches is also studied numerically in Section 4.6. In spite of the choice of A_ℓ, the geometric multilevel substructured domain decomposition method (GMS) is described in Algorithm 6, which is a substructured multi-grid V-cycle.

4.6 Numerical Experiments

In this section we present numerical experiments to validate the computational framework presented in this chapter. Subsection 4.6.3 considers a Laplace equation in a two subdomains decomposition. We consider a two dimensional problem and the aim is to provide an overview of the different methods in this simple example. We then consider a three dimensional problem, showing the effects of the implementation tricks discussed in section 4.4 and we present a comparison of computational times. In subsection 4.6.2, we consider the Laplace equation in a many-subdomain decomposition and, while discussing the numerical results, we provide further implementation details. Finally subsection 4.6.3 deals with a more challenging diffusion problem with jumping coefficients.

4.6.1 Laplace equation on 2D and 3D boxes

We consider the Poisson equation $-\Delta u = f$ in a rectangle $\Omega = (-1,1)\times(0,1)$ with homogeneous Dirichlet boundary condition. The domain Ω is decomposed into two overlapping rectangles $\Omega_1 = (-1,\delta)\times(0,1)$ and $\Omega_2 = (-\delta,1)\times(0,1)$, where 2δ is the length of the overlap. We discretize the problem using a standard second-order finite difference scheme based on a uniform grid of $N_y = 2^\ell - 1$ interior points in direction y and $N_x = 2N_y + 1$ interior points in direction x. The overlap is $2\delta = h(N_{ov}+1)$ where h is the mesh size and where N_{ov} represents the number of interior points in the overlap in direction x. The results of our numerical experiments are shown in Figures 4.3.

For the G2S method we use the one-dimensional interpolation operator P_{2h}^h defined in (4.3.3) and $R_{2h}^h = \frac{1}{2}(P_{2h}^h)^\top$. For the S2S method and the SHEM method, we used coarse

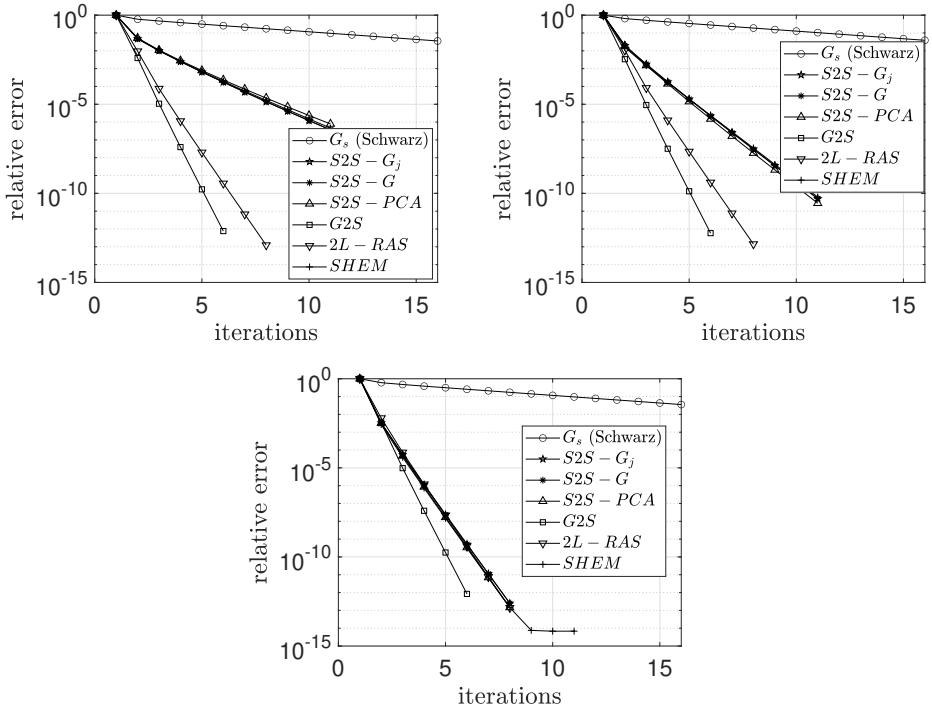


Figure 4.3: Convergence curves for $\ell = 6$, $N_{ov} = 4$, and $N_c = 10$ (top-left), $N_c = 20$ (top-right), $N_c = 40$ (bottom).

spaces of dimension N_c. This means that for the SHEM method and S2S-G_j, we include $N_c/2$ sine Fourier functions on each interface. We also used the S2S method together with N_c coarse functions generated randomly by the PCA procedure and this is denoted by "S2S-PCA". To generate the PCA coarse space, we set $q = 2N_c$ and $r = 2$.

The figures show the decay of the relative errors with respect to the number of iterations. All the methods are stopped if the relative error is smaller that 10^{-12}. The G2S and the two-grid RAS methods outperform the other methods in terms of iterations numbers and the G2S method outperforms the two-grid RAS method. Notice that, while the G2S coarse space has dimension about N_y, the one corresponding to the two-grid RAS method has dimension about $N_x N_y/4 \approx N_y^2/2 \gg N_y$. The spectral methods perform very well since already for $N_c = 10$ they achieve an error of about 10^{-6} with less than 10 iterations and we emphasize that the PCA coarse space has the same performance. Increasing the dimension of the coarse space, the convergence of the S2S methods and of the SHEM method drastically improves. Notice that if $N_c = 40$, the dimension of the coarse spaces for S2S and SHEM is 40, while the dimension of the coarse spaces of G2S and 2L-RAS are about 60 and 1900, respectively. The slower performance of 2L-RAS with respect to G2S can be traced back to the interpolation step. This operation breaks the harmonicity of the obtained correction, which therefore does not lie anymore in the space where the errors

# (volume)	G2S	G2S-B1	G2S-B2	2L-RAS
539	4	4	4	6
6075	5	5	4	6
56699	4	4	4	6
488187	4	4	4	6

Table 4.2: Number of iterations performed by the different methods and for different number of degrees of freedom.

# (volume)	G2S	G2S-B1	G2S-B2	2L-RAS
539	0.023 (0.005)	0.010 (0.003)	0.010 (0.003)	0.039 (0.06)
6075	0.143 (0.028)	0.102 (0.024)	0.070 (0.017)	0.190 (0.03)
56699	2.700 (0.675)	1.598 (0.399)	1.280 (0.320)	4.128 (0.688)
488187	126.0980 (31.524)	78.363 (19.591)	63.131 (15.783)	189.162 (31.527)

Table 4.3: Computational times performed by the different methods. In parentheses we indicate the computational time per iteration.

lie; see, e.g., [94]. One could use interpolators which extend harmonically the correction inside the overlapping subdomains although this would increase significantly the computational cost of each iteration, see the discussion in Chapter 3.

Next, we repeat the same experiments on a three-dimensional box $\Omega = (-1,1) \times (0,1) \times (0,1)$ decomposed into two overlapping subdomains $\Omega_1 = (-1,\delta) \times (0,1) \times (0,1)$ and $\Omega_2 = (-\delta,1) \times (0,1) \times (0,1)$. Since we are interested in computational times, we solve the problem (up to a tolerance of 10^{-10} on the relative error) using the G2S method, its equivalent forms G2S-B1 and G2S-B2, introduced in Section 4.4, and 2L-RAS. The length of the overlap is $\delta = hN_{ov}$, where h is the grid size and N_{ov} is fixed to 4. Hence the overlap is proportional to the grid size. The results are shown in Tables 4.2 and 4.3. It is clear that the G2S methods outperforms 2L-RAS, in terms of iteration numbers and computational times. In particular, G2S-B1 and G2S-B2 require per iteration about half of the computational time required by 2L-RAS. The experiments have been performed on a workstation with 8 processors Intel Core i7-6700 CPU 3.40GHz and with 32 GB of RAM.

4.6.2 Decompositions into many subdomains

In this paragraph, we consider a square domain Ω decomposed into $M \times M$ nonoverlapping square subdomains Ω_j, $j = 1,...,M^2 = N$. Each subdomain Ω_j contains N_{sub}^2 interior degrees of freedom, with $N_{sub} := 2^\ell - 1$. Extending the subdomains Ω_j by N_{ov} points, we obtain the overlapping subdomains Ω_j' with overlap $\delta = 2N_{ov}h$. On each subdomain Ω_j, we locate the discrete substructure $\mathscr{S}_j^{N_j}$, marked with blue lines in Figure 4.4, which is made by four (one-dimensional) segments. On this domain, we consider a classical Laplace equation.

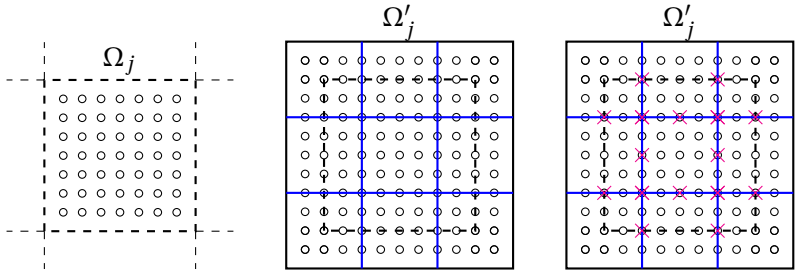


Figure 4.4: A nonoverlapping subdomain Ω_j is enlarged by $N_{ov} = 2$ points in each direction. The discrete substructure $\mathscr{S}_j^{N_j}$ is denoted by a blue line. On the right panel, the coarse discrete substructure $\mathscr{S}_j^{M_j}$ is marked by red crosses.

Figure 4.5 compares several versions of the S2S method with respect to the SHEM coarse space. We follow [90] for the implementation details of SHEM. We specifically compare a S2S method with a coarse space made by eigenfunctions of G (S2S-G), a S2S method with a coarse space obtained with the PCA procedure (S2S-PCA), and a S2S method with a coarse space which is derived by the SHEM coarse space (S2S-HEM, that is S2S Harmonically Enriched Multiscale). In more detail, we first create the SHEM coarse space solving interface eigenvalue problems, see equation (8) in [90], and we extend these interface functions into the interior of the subdomains. We then restricted these volume functions on the substructures to obtain a basis for the S2S-HEM coarse space. For the PCA approach, we generated $q = 2\times\dim V_c$ random vectors $\boldsymbol{x}_k$ and we set $r = 2$. The result we plot is averaged over 30 different random coarse spaces. The size of the coarse space is set by the SHEM coarse space. In the left panel, we consider only the multiscale functions without solving any eigenvalue problem along the interfaces. In the center panel, we include the first eigenfunctions on each interface, and on the right we include the first and the second eigenfunctions. In all cases we observe that the methods have a similar convergence, which is slightly faster for the substructured methods. As we remarked at the end of Section 4.2, S2S-G is not necessarily the fastest.

We now focus on the G2S method. For each discrete substructure $\mathscr{S}_j^{N_j}$, the geometric interpolation operator P_j acts block-wise on each one dimensional interval, i.e. $P_j = \mathrm{diag}\{\widetilde{P}_1, \widetilde{P}_2, \widetilde{P}_3, \widetilde{P}_4\}$, where each $\widetilde{P}_k$, $k = 1,..,4$, corresponds to the prolongation matrix (4.3.3). The results of our numerical experiments for the G2S method are reported in Figure 4.6. The left panel shows the dependence of the spectral radius on the size of the overlap for the one-level substructured method G, the RAS method, the G2S method and the 2L-RAS method for $N = 16$, $\ell = 5$. We then study the robustness of the method with respect to an increasing number of subdomains. We first keep the size of each subdomain fixed, $N_{sub} = 2^5 - 1$, and thus we consider larger global problems as N grows. Then, we fix a global domain Ω with approximately $17 \cdot 10^3$ interior degrees of freedom, and we get smaller subdomains as N grows. In both cases, we observe that the spectral radius of both 2L-RAS and G2S does not deteriorate as the number of subdomains increases. Moreover,

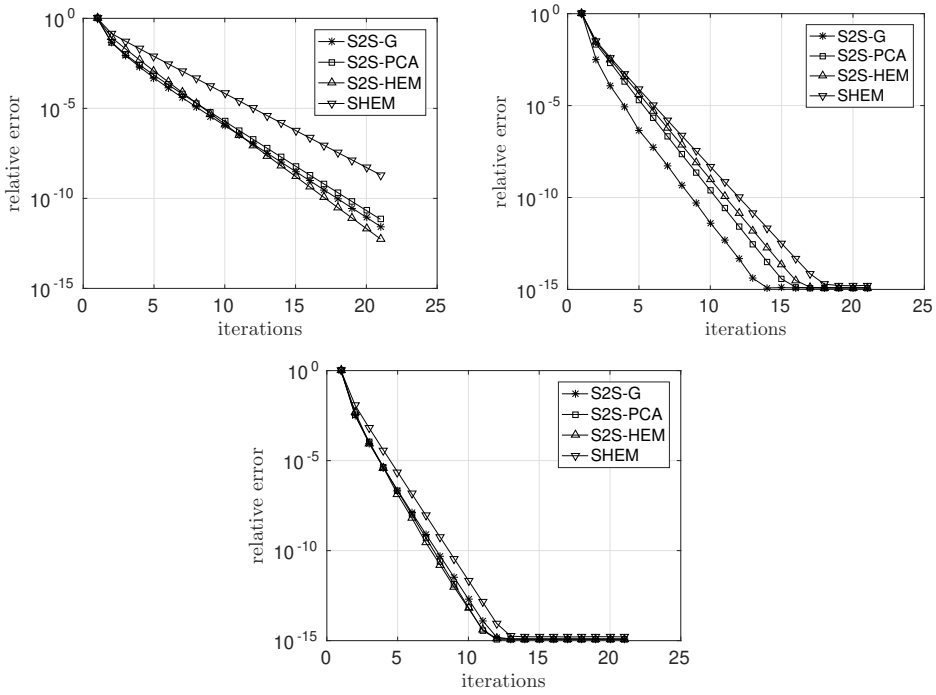


Figure 4.5: Convergence behavior of the different methods for a Laplace equation wit $N = 16$, $\ell = 4$ and $N_{ov} = 2$. The dimension of the coarse space is 36 (top-left), 84 (top-right), 132 (bottom).

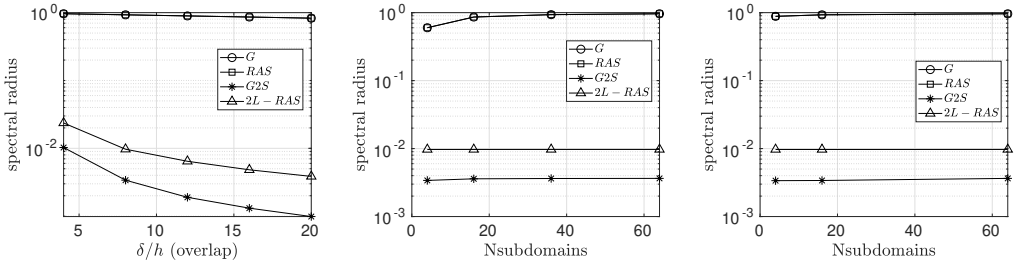


Figure 4.6: Dependence of spectral radius on the overlap (left) and robustness of the two-level methods when increasing the number of subdomains for subdomains with same size (center) and global problem fixed (right).

these numerical experiments confirm that the G2S method is faster in terms of iteration count compared to the 2L-RAS method.

4.6.3 Diffusion problem with jumping diffusion coefficients

In this section, we test the S2S and G2S methods for the solution of a diffusion equation $-\text{div}(\alpha\nabla u) = f$ in a square domain $\Omega := (0,1)^2$ with $f := \sin(4\pi x)\sin(2\pi y)\sin(2\pi xy)$. The

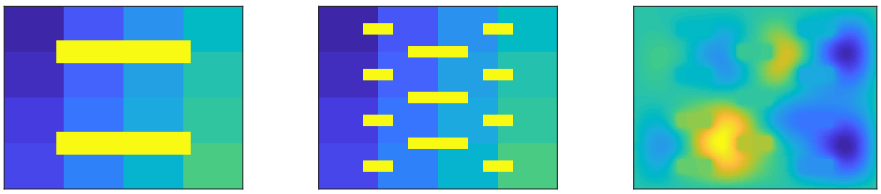


Figure 4.7: Decomposition of Ω into 16 subdomains with two different patterns of channels (left and center) and solution of the equation with the central pattern (right).

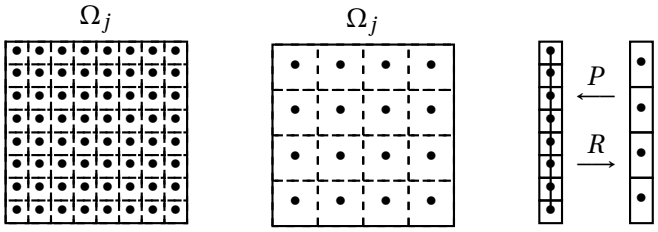


Figure 4.8: Illustration of the action of the restriction operator in volume (left) and of the restriction and interpolation operators on a one-dimensional substructure (right).

domain Ω is decomposed into 16 non-overlapping subdomains and we suppose $\alpha = 1$ everywhere except in some channels where α takes the values 10^2, 10^4 and 10^6. We consider two configurations represented in Figure 4.7. We use a finite-volume discretization, where each non-overlapping subdomain is discretized with $N_{\text{sub}} = 2^\ell$ cells and it is enlarged by N_{ov} cells to create an overlapping decomposition with overlap $\delta = 2N_{ov}h$. We further assume that the discontinuities of the diffusion coefficient are aligned with the edges of the cells and they do not cross any cell.

Concerning the geometric methods, the mapping between the fine and coarse mesh is illustrated in Figure 4.8. At the volume level, the restriction operator maps four fine cells to a single coarse cell by averaging the four cell values and the interpolation operator is its transpose. At the substructured level, the restriction operator maps two fine cells to a single coarser cell by averaging. The interpolation operator splits one coarse cell to two fine cells assigning the same coarse value to each new cell. It still holds that the interpolation operators is the transpose of the restriction operator. In this setting, we study the robustness of the G2S method with respect to the mesh size and the amplitudes of the jumps of α and we compare it to the 2L-RAS method. In Table 4.4 we report the number of iterations to reach a relative error of Tol = 10^{-12}. The iterations performed by the G2S method are the numbers on the left in each cell of the table, while the iterations of the 2L-RAS are the numbers in brackets on the right. These results show that the G2S method is robust both with respect the jumps of the diffusion coefficient and the mesh size, and that it outperforms the 2L-RAS method.

$\dim V_c$	456	840	1608
α \ N^v	4096	16384	65536
10^2	7 (39)	7 (41)	7 (41)
10^4	7 (42)	7 (41)	7 (41)
10^6	7 (39)	7 (40)	7 (40)

$\dim V_c$	456	840	1608
α \ N^v	4096	16384	65536
10^2	8 (45)	7 (41)	7 (41)
10^4	8 (42)	7 (41)	7 (41)
10^6	8 (39)	7 (39)	7 (40)

Table 4.4: Number of iterations performed by the G2S and 2L-RAS (in brackets) methods with $N_{ov} = 2$ and for different values of jumps of α and different numbers of degrees of freedom N^v. The dimension of the substructured coarse space is $\dim V_c$. The left table refers to the two channels configuration and the right table to the multiple channels one.

α	S2S-G	S2S-PCA	S2S-EHM	SHEM
10^2	11-9-7	11-9-7	10-9-7	12-10-7
10^4	11-9-7	11-9-7	11-9-7	12-11-7
10^6	11-9-7	10-8-7	10-8-7	10-9-7

Table 4.5: For each spectral method and value of α, we report the number of iterations to reach a relative error smaller than 10^{-8} with a coarse space of dimension 84 (left), 132 (center) and 180 (right). The discretization parameters are $N^v = 16384$ and $N_{ov} = 2$.

We then investigate the performances of the S2S methods and we compared them with the SHEM coarse space in the multiple channel configuration. We set $\ell = 4$, which corresponds to $N^v = 4096$ degrees of freedom, and $N_{ov} = 2$. Table 4.5 shows the number of iterations to reach a relative error smaller than 10^{-8} for the S2S-G, S2S-PCA, S2S-HEM and SHEM methods. We consider coarse spaces of dimension 84, 132 and 180, which, for the SHEM and S2S-HEM methods, correspond to multiscale coarse spaces enriched by respectively the first, second and third eigenvectors of the interface problem (8) in [90]. For PCA coarse space, we set $q = 2N_c$ and $r = 6$. We remark that for smaller values of r, the S2S-PCA method diverges. This increase in the value of r can be explained noticing that for the multichannel configuration, the smoother G has several eigenvalues approximately 1 for large values of α. Thus the PCA procedure, which essentially relies on a power method idea to approximate the image of G, suffers due to the presence of several clustered eigenvalues, and hence does not provide accurate approximations of the eigenfunctions of G. We also observed that a straightforward use of the restriction of the SHEM functions could lead to a divergent S2S-EHM method. In order to improve this coarse space, we build a matrix whose columns are the restriction of the SHEM functions. We then use this matrix, instead of a random one, in the PCA procedure, obtaining a new coarse space which is then used in the S2S-EHM method. Table 4.5 shows that all spectral methods have very similar performance. We remark that all of them are robust with respect to the strength of the jumps.

CHAPTER 5

Application of optimized Schwarz methods to the Stokes-Darcy coupling

"Il existe sous la surface du sol, dans le terrains stratifiés, tantôt de véritables cours d'eau souterrains circulant, avec des vitesses sensibles, dans des fissures, fentes ou cavités naturelles"

— H. Darcy, Les fontaines publiques de la ville de Dijon, pag. 137.

Over the last decades, the filtration of fluids through porous media has increasingly drawn the attention of researchers due to the large number of applications in physical processes. Instances are blood simulations [47], groundwater and oil simulations [4], food processes [49] and soil-water evaporation with applications to nuclear waste disposal [107]. In this chapter, we study domain decomposition strategies to deal effectively with the Stokes-Darcy system. Seminal works in this direction have been done by Discacciati in his Ph.D. thesis [52], which culminated in the review article [55]. In the Chapters 2-3-4 of [52], the author reduces the global Stokes-Darcy system to a single interface equation involving the Steklov-Poincaré operator. Then a Dirichlet-Neumann preconditioner, much in the spirit of Section 1.3.2, is proposed, and numerical results show that the preconditioner is robust with respect to mesh size but not with respect to the physical parameters. Specifically, low values of the diffusion constants lead to a poor performance. Thus, Robin-Robin domain decomposition methods have been introduced in [56], where the Robin parameters are heuristically tuned to make the method robust with respect to the physical parameters. Robin-Robin domain decomposition methods for this system have also been proposed by other groups, we name in particular the work by Xiaoming He, Yanzhao Cao and collaborators [111, 23, 24]

In this chapter, we present our contribution to the definition of efficient domain decomposition strategies for the Stokes-Darcy coupling. We first propose a one-level OSM and we discuss the limitations of standard techniques to find optimized transmission con-

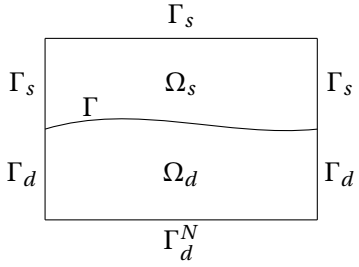


Figure 5.1: Stokes-Darcy domain

ditions for this complicated system. Then, we discuss the conditions for which we can rely on the standard Fourier approach, and we further apply the probing technique to find good estimates when those conditions are not satisfied. We then focus on two-level and multilevel solvers. First, we apply the multilevel optimized Schwarz framework introduced in Chapter 3 to design a two-level solver for the Stokes-Darcy coupling. Finally, we apply the substructured framework discussed in Chapter 4 to define two-level spectral and geometric substructured methods.

5.1 Definition of the model

Let $\Omega \subset \mathbb{R}^d$, $d = 2,3$, be a bounded domain decomposed into two nonoverlapping subdomains Ω_s and Ω_d, separated by a sufficiently regular interface Γ, see Figure 5.1. The unit normal vector pointing towards Ω_d is denoted with $\mathbf{n}$. We suppose that Ω, Ω_s and Ω_d are Lipschitz domains, and we define $\Gamma_s := \partial\Omega_s \setminus \Gamma$, and $\partial\Omega_d \setminus \Gamma = \Gamma_d \cup \Gamma_d^N$. We assume that Ω_s contains an incompressible fluid described by the Stokes equations

$$-\nabla \cdot T(\mathbf{u}_s, p_s) = \mathbf{f}, \quad \nabla \cdot \mathbf{u}_s = 0, \quad \text{in } \Omega_s, \tag{5.1.1}$$

where $T(\mathbf{u}_s, p_s) := 2\mu\nabla^s \mathbf{u}_s - p_f I$ is the stress tensor, $\nabla^s \mathbf{u}_s := \frac{1}{2}(\nabla \mathbf{u}_s + (\nabla \mathbf{u}_s)^\top)$ is the symmetrized gradient, $\mathbf{f}_s$ is a body external force and $\mu \in \mathbb{R}^+$ is the dynamic viscosity of the fluid. The unknowns are the fluid velocity field $\mathbf{u}_s$ and the pressure field p_s.
The lower domain Ω_d consists of a porous medium filled by a fluid which flows according to Darcy's law, discovered experimentally in 1886 [48],

$$\mathbf{u}_d = -K\nabla p_d + \mathbf{g}_d, \quad \nabla \cdot \mathbf{u}_d = 0 \quad \text{in } \Omega_d, \tag{5.1.2}$$

where $\mathbf{u}_d$ is the fluid velocity, p_d is the Darcy pressure and $\mathbf{g}_d$ is a body force. $K \in \mathbb{R}^{d\times d}$ is the permeability tensor of the porous medium. Generally, K is a symmetric positive definite tensor that can be diagonalized by introducing the so-called principal directions of anisotropy, i.e. $K = \text{diag}(k_1, ..., k_d)$, $k_i \in L^\infty(\Omega_d)$, $k_i > 0$ a.e. in Ω_d. Taking the divergence of the first equation in (5.1.2), we obtain a second order PDE only in terms of the pressure

$$-\nabla \cdot K\nabla p_d = -\nabla \cdot \mathbf{g}_d \quad \text{in } \Omega_d. \tag{5.1.3}$$

For the sake of simplicity, we impose homogeneous boundary conditions for the Stokes domain, $\mathbf{u}_s = 0$ on Γ_s. On the Darcy domain, we set $p_d = 0$ on Γ_d, and a no slip condition $K\nabla p_d \cdot \mathbf{n}_{\text{ext}} = 0$ on Γ_d^N, where $\mathbf{n}_{\text{ext}}$ is the unit outward normal on Γ_d^N.
The two physical models need to be coupled along the common interface Γ. There is not a unique choice for the coupling conditions and we refer the reader to Section 3 of [55] for a more detailed discussion of several cases. Generally, the continuity of the normal velocity, i.e. $\mathbf{u}_s \cdot \mathbf{n} = \mathbf{u}_d \cdot \mathbf{n}$, and the continuity of the normal stress, i.e. $-\mathbf{n} \cdot T \cdot \mathbf{n} = p_d$ are prescribed. We remark that this last condition actually allows the pressure to be discontinuous along Γ. In order to have a well-posed problem in Ω_s, we still need to impose a condition on the tangential velocity on Γ. In 1967, Beavers and Joseph found experimentally that the difference between the slip velocity is proportional to the shear rate of the free fluid [6]. In mathematical terms this is equivalent to impose

$$-\tau_j \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} = \frac{\epsilon}{\mu}(\mathbf{u}_s - \mathbf{u}_d) \cdot \tau_j, \quad j = 1, ..., d-1 \quad \text{on } \Gamma, \tag{5.1.4}$$

where τ_j are linear independent unit tangential vectors lying on the interface Γ, and ϵ is a constant depending on the physical structure of the porous medium. The well-posedness of the Stokes-Darcy system with (5.1.4) has been only proved in 2010 [25]. However, Saffman noticed in [141] that, in most applications, $\mathbf{u}_d$ is much smaller that $\mathbf{u}_s$ and thus can be neglected. Supposing $\mathbf{u}_d = 0$ in equation (5.1.4), we get the so called Beaver-Joseph-Saffman (BJS) condition. This condition has also been derived mathematically through homogenization theory in [134]. Another possible choice, which is much used in blood simulations, is to impose a zero tangential velocity, i.e. $\mathbf{u}_s \cdot \tau_j = 0$, $j = 1, ..., d-1$. To summarize, in this thesis we will suppose the following coupling conditions:

$$\begin{aligned} \mathbf{u}_s \cdot \mathbf{n} &= \mathbf{u}_d \cdot \mathbf{n}, \\ -\mathbf{n} \cdot (T(\mathbf{u}_s, p_s) \cdot \mathbf{n}) &= p_d, \\ -\epsilon\tau_j \cdot (T(\mathbf{u}_s, p_s) \cdot \mathbf{n}) &= \mu \mathbf{u}_s \cdot \tau_j, \quad j = 1, ..., d-1. \end{aligned} \tag{5.1.5}$$

The strong form of the coupled Stokes-Darcy system is

$$\begin{aligned} -\nabla \cdot T(\mathbf{u}_s, p_s) = \mathbf{f}_s, \quad \nabla \cdot \mathbf{u}_s = 0, &\qquad \text{in } \Omega_s, \\ -\nabla \cdot K\nabla p_d = -\nabla \cdot \mathbf{g}_d, &\qquad \text{in } \Omega_d, \\ \mathbf{u}_s = 0, &\qquad \text{on } \Gamma_s, \\ \mathbf{u}_d = 0 \text{ on } \Gamma_d, \quad K\nabla p_d \cdot \mathbf{n}_{\text{ext}} = 0, &\qquad \text{on } \Gamma_d^N, \\ \mathbf{u}_s \cdot \mathbf{n} = \mathbf{u}_d \cdot \mathbf{n}, &\qquad \text{on } \Gamma, \\ -\mathbf{n} \cdot (T(\mathbf{u}_s, p_s) \cdot \mathbf{n}) = p_d, &\qquad \text{on } \Gamma, \\ -\epsilon\tau_j \cdot (T(\mathbf{u}_s, p_s) \cdot \mathbf{n}) = \mu \mathbf{u}_s \cdot \tau_j, &\qquad \text{on } \Gamma. \end{aligned} \tag{5.1.6}$$

Introducing two positive real parameters s_1 and s_2 and two initial guesses λ_s^0 and λ_d^0, the optimized Schwarz method for system (5.1.6) computes for $n = 1, 2...$

$$
\begin{aligned}
(\mathbf{u}_s^n, p_s^n) = \text{Stokes Problem}(\mathbf{f}, \lambda_s^{n-1}) : & \\
-\nabla \cdot (T(\mathbf{u}_s^n, p_s^n)) = \mathbf{f}_s, \quad \nabla \cdot \mathbf{u}_s^n = 0, & \qquad \text{in } \Omega_s, \\
\mathbf{u}_s^n = 0, & \qquad \text{on } \Gamma_s, \\
-\epsilon \tau_j \cdot (T(\mathbf{u}_s^n, p_s^n) \cdot \mathbf{n}) = \mu \mathbf{u}_s^n \cdot \tau_j, & \qquad \text{on } \Gamma, \\
-\mathbf{n} \cdot (T(\mathbf{u}_s^n, p_s^n)) \cdot \mathbf{n} - s_2 \mathbf{u}_s^n \cdot \mathbf{n} = \lambda_s^{n-1}, & \qquad \text{on } \Gamma, \\
p_d^n = \text{Darcy Problem}(\mathbf{g}_d, \lambda_d^{n-1}) : & \\
-\nabla \cdot K \nabla p_d^n = -\nabla \cdot \mathbf{g}_d, & \qquad \text{in } \Omega_d, \\
\mathbf{u}_d = 0 \text{ on } \Gamma_d, \quad K \nabla p_d \cdot \mathbf{n}_{\text{ext}} = 0, & \qquad \text{on } \Gamma_d^N \\
p_d^n - s_1 \left(K \nabla p_d^n \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n} \right) = \lambda_d^{n-1}, & \qquad \text{on } \Gamma,
\end{aligned}
\tag{5.1.7}
$$

with the updating rules

$$
\begin{aligned}
\lambda_s^n &= p_d^n + s_2 \left(K \nabla p_d^n \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n} \right) = \left(1 + \frac{s_2}{s_1} \right) p_d^n - \frac{s_2}{s_1} \lambda_d^{n-1}, \\
\lambda_d^n &= -\mathbf{n} \cdot T(\mathbf{u}_s^n, p_s^n) \cdot \mathbf{n} + s_1 \mathbf{u}_s^n \cdot \mathbf{n} = \lambda_s^{n-1} + (s_1 + s_2) \mathbf{u}_s^n \cdot \mathbf{n}.
\end{aligned}
\tag{5.1.8}
$$

This domain decomposition algorithm has been studied in the literature by several authors. In [56], the authors defined the method and provided a convergence analysis based on energy estimates. A similar analysis has been carried out in [35] where the author included a rough estimation of the optimized parameters through a Von Neumman analysis. The first theoretical analysis devoted to establish optimized transmission conditions has been carried out in [53]. Unfortunately, the standard techniques used to derive optimized transmission conditions are not effective for this particular coupling. Part of the contribution of this thesis is to investigate the reasons for this failure and to propose alternative approaches. This is the main topic of Section 5.3, which is largely based on the proceeding paper [98]. Before concluding this Section, we refer the interested reader to [111] and [23] for a study of domain decomposition methods based on Robin boundary conditions for the Stokes-Darcy system equipped with the Beaver-Joseph coupling condition. To the best of our knowledge, optimized transmission conditions have not been derived for Beavers-Joseph coupling conditions. We also cite [24] where the authors proposed an interesting non-iterative marching in time scheme based on domain decomposition algorithms.

5.2 Weak formulation and well-posedness

In this Section we derive a weak formulation for the coupled Stokes-Darcy system (5.1.6) and for the domain decomposition algorithm (5.1.7), and we show their well-posedness.

First, we introduce the functional spaces and norms

$$\begin{aligned}
&H_s := \left\{\mathbf{w} \in (H^1(\Omega_s))^d : \mathbf{w} = 0 \text{ on } \Gamma_s\right\}, && \|\mathbf{w}\|^2_{H_s} := \textstyle\int_{\Omega_s} |\nabla \mathbf{w}|^2,\\
&Q_s := L^2(\Omega_s), && \|q\|^2_{Q_s} := \textstyle\int_{\Omega_s} |q|^2,\\
&H_d := \left\{\psi \in H^1(\Omega_d) : \psi = 0 \text{ on } \Gamma_d\right\}, && \|\psi\|^2_{H_d} := \textstyle\int_{\Omega_d} |\nabla \psi|^2,\\
&W := H_s \times H_d, && \|\underline{w}\|^2_W := (\|\mathbf{w}\|^2_{H_s} + \|\psi\|^2_{H_d})^{\frac{1}{2}},\\
&\Lambda := H^{\frac{1}{2}}_{00}(\Gamma), && \|\lambda\|_\Lambda := \|\lambda\|_{H^{\frac{1}{2}}_{00}},
\end{aligned} \tag{5.2.1}$$

as well as the bilinear forms

$$\begin{aligned}
\widetilde{a}_s(\mathbf{u},\mathbf{v}) &:= 2\mu \textstyle\int_{\Omega_s} \nabla^s \mathbf{u} : \nabla^s \mathbf{v} + \sum_{j=1}^{d-1} \int_\Gamma \frac{\mu}{\epsilon} (\mathbf{u})_{\tau_j} (\mathbf{v})_{\tau_j}, && \forall \mathbf{u},\mathbf{v} \in H_s,\\
a_s(\mathbf{u},\mathbf{v}) &:= \widetilde{a}_s(\mathbf{u},\mathbf{v}) + \textstyle\int_\Gamma s_2 (\mathbf{u}\cdot\mathbf{n})(\mathbf{v}\cdot\mathbf{n}), && \forall \mathbf{u},\mathbf{v} \in H_s,\\
b_s(\mathbf{v},q) &:= -\textstyle\int_{\Omega_s} q \nabla\cdot\mathbf{v}, && \forall \mathbf{v} \in H_s, \forall q \in Q_s,\\
\widetilde{a}_d(p_d,q_d) &:= \textstyle\int_{\Omega_d} K\nabla p_d \cdot \nabla q_d, && \forall p_d, q_d \in H_d,\\
a_d(p_d,q_d) &:= \widetilde{a}_d(p_d,q_d) + \textstyle\int_\Gamma \frac{1}{s_1} p_d q_d, && \forall p_d, q_d \in H_d,\\
C(p_d,\mathbf{v}) &:= \textstyle\int_\Gamma p_d (\mathbf{v}\cdot\mathbf{n}), && \forall p_d \in H_d, \forall \mathbf{v} \in H_s,\\
\mathscr{A}(\underline{v},\underline{w}) &:= 2\mu \textstyle\int_{\Omega_s} \nabla^s \mathbf{v} : \nabla^s \mathbf{w} + \sum_{j=1}^{d-1} \int_\Gamma \frac{\mu}{\epsilon} (\mathbf{v})_{\tau_j} (\mathbf{w})_{\tau_j} + \int_{\Omega_d} \nabla\phi \cdot K \nabla\psi\\
&\quad + \textstyle\int_\Gamma \phi(\mathbf{w}\cdot\mathbf{n}) - \int_\Gamma (\mathbf{v}\cdot\mathbf{n})\psi, && \forall \underline{v} = (\mathbf{v},\phi), \forall \underline{w} = (\mathbf{w},\psi) \in W,\\
\mathscr{B}(\underline{w},q) &:= b_s(\mathbf{w},q), && \forall \underline{w} = (\mathbf{w},\psi) \in W, \forall q \in Q_s,\\
\mathscr{F}(\underline{w}) &:= \textstyle\int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{w} + \int_{\Omega_d} \mathbf{g}_d \cdot \nabla\psi, && \forall \underline{w} = (\mathbf{w},\psi) \in W.
\end{aligned} \tag{5.2.2}$$

To have a well defined functional $\mathscr{F}$ acting on W, we assume that $\mathbf{f}_s$ and $\mathbf{g}_d$ are in $(L^2(\Omega_d))^d$.

Integrating by part (5.1.6) and using the coupling conditions (5.1.5), one gets the weak formulation,

$$\begin{aligned}
&\text{Find } \mathbf{u}_s \in H_s,\ p_s \in Q_s,\ p_d \in H_d \text{ such that for all } \mathbf{v} \in H_s,\ q_s \in Q_s,\ q_d \in H_d\\
&2\mu \int_{\Omega_s} \nabla^s \mathbf{u}_s : \nabla^s \mathbf{v} + \sum_{j=1}^{d-1} \int_\Gamma \mu (\mathbf{u}_s)_{\tau_j} (\mathbf{v})_{\tau_j} - \int_{\Omega_s} p \nabla\cdot\mathbf{v} + \int_{\Omega_d} \nabla p_d \cdot K \nabla q_d,\\
&+ \int_\Gamma p_d (\mathbf{v}\cdot\mathbf{n}) - \int_\Gamma (\mathbf{u}_s\cdot\mathbf{n}) q_d = \int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{v} + \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d,\\
&- \int_{\Omega_s} q_s \nabla\cdot\mathbf{u}_s = 0.
\end{aligned} \tag{5.2.3}$$

Using the bilinear forms defined in (5.2.2), equation (5.2.3) can be written as

$$\begin{aligned}
&\text{Find } \mathbf{u}_s \in H_s,\ p_s \in Q_s,\ p_d \in H_d \text{ such that for all } \mathbf{v} \in H_s,\ q_s \in Q_s,\ q_d \in H_d\\
&\widetilde{a}_s(\mathbf{u}_s,\mathbf{v}) + b_s(\mathbf{v},p_s) + \widetilde{a}_d(p_d,q_d) + C(p_d,\mathbf{v}) - C(q_d,\mathbf{u}_s) = \int_{\Omega_s} \mathbf{f}_s\cdot\mathbf{v} + \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d,\\
&b_s(\mathbf{u}_s,q_s) = 0,
\end{aligned} \tag{5.2.4}$$

which, in a more compact form, corresponds to

$$\begin{aligned} &\text{Find } \underline{u} = (\mathbf{u}_s, p_d) \in W \text{ and } p_s \in Q_s \text{ such that} \\ &\mathscr{A}(\underline{u}, \underline{v}) + \mathscr{B}(\underline{v}, p_s) = F(\underline{v}), \quad \forall \underline{v} \in W, \\ &\mathscr{B}(\underline{u}, q) = 0, \quad \forall q \in Q_s. \end{aligned} \tag{5.2.5}$$

System (5.2.5) is well-posed and it admits a unique solution $(\underline{u}, p_s) \in W \times Q_s$. The proof, based on Brezzi's saddle point theory, can be found in Chapter 2 of [52]. To solve a general Stokes-Darcy problem, one could directly introduce a finite dimensional approximation of (5.2.5). For instance, given two regular triangulations $\mathscr{T}_h^s$ and $\mathscr{T}_h^d$, we define the finite element spaces

$$\begin{aligned} H_s^h &:= \left\{ \boldsymbol{\psi}_s^h \in (C^0(\overline{\Omega}_s))^d : \boldsymbol{\psi}_s^h|_T \in (\mathbb{P}_2(T))^d, \forall T \in \mathscr{T}_h^s, \quad \boldsymbol{\psi}_s^h|_{\Gamma_s} = 0 \right\}, \\ Q_s^h &:= \left\{ \phi_s^h \in C^0(\overline{\Omega}_s) : \phi_s^h|_T \in \mathbb{P}_1(T), \forall T \in \mathscr{T}_h^s \right\}, \\ H_d^h &:= \left\{ \psi_d^h \in C^0(\overline{\Omega}_d) : \psi_d^h|_T \in \mathbb{P}_2(T), \forall T \in \mathscr{T}_h^d, \quad \psi_d^h|_{\Gamma_d} = 0 \right\}, \end{aligned} \tag{5.2.6}$$

and the basis $H_s^h = \text{span}\left\{(\boldsymbol{\psi}_{s,i}^h)_{i=1}^{i=N_v^s}\right\}$, $Q_s^h = \text{span}\left\{(\phi_{s,i}^h)_{i=1}^{i=N_p^s}\right\}$, $H_d^h = \text{span}\left\{(\psi_{d,i}^h)_{i=1}^{i=N_d}\right\}$. Then the discrete version of (5.2.5) is

$$\begin{pmatrix} \widetilde{A}_s & B_s^\top & C \\ B_s & 0 & 0 \\ -C & 0 & \widetilde{A}_d \end{pmatrix} \begin{pmatrix} \mathbf{u}_s^h \\ \mathbf{p}_s^h \\ \mathbf{p}_d^h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_s^h \\ 0 \\ \mathbf{g}_d^h \end{pmatrix}, \tag{5.2.7}$$

where

$$\begin{aligned} (\widetilde{A}_s)_{i,j} &:= \widetilde{a}_s(\boldsymbol{\psi}_{s,j}^h, \boldsymbol{\psi}_{s,i}^h), & (\widetilde{A}_d)_{i,j} &:= \widetilde{a}_d(\psi_{d,j}^h, \psi_{d,i}^h), \\ (B_s)_{i,j} &:= b_s(\boldsymbol{\psi}_{s,j}^h, \phi_{s,i}^h), & (C)_{i,j} &:= C(\psi_{d,j}^h, \boldsymbol{\psi}_{s,i}^h), \\ (\mathbf{f}_s^h)_i &:= \textstyle\int_{\Omega_s} \mathbf{f}_s \cdot \boldsymbol{\psi}_{s,i}^h, & (\mathbf{g}_d^h)_i &:= \textstyle\int_{\Omega_d} \mathbf{g}_d \cdot \nabla \psi_{d,i}^h, \end{aligned} \tag{5.2.8}$$

and $\mathbf{u}_s^h, \mathbf{p}_s^h$ and $\mathbf{p}_d^h$ are the coefficients of the solution in the corresponding basis. We have verified this approach, by numerically solving a Stokes-Darcy problem with $\Omega_s = (0,1) \times (0,1)$, $\Omega_d = (0,1) \times (-1,0)$, with solution

$$\begin{aligned} &\mathbf{u}_s^{\text{ex}} := (x^2y^2 + e^{-y}, -\frac{2}{3}xy^3 + 2 - \pi\sin(\pi x)), \quad p_s^{\text{ex}} = -(2 - \pi\sin(\pi x))\cos(2\pi y), \\ &p_d^{\text{ex}} := -(2 - \pi\sin(\pi x))(y+1). \end{aligned} \tag{5.2.9}$$

One can verify that these functions satisfy the coupling conditions (5.1.5) along the interface $\Gamma = (0,1) \times \{0\}$. The functions f_1, f_2 and $\mathbf{g}_d$ are chosen such that the functions (5.2.9) satisfy the interior equations. We solved this problem using the GDGMatlab library, a finite element library we developed during these years which implements Lagrangian continuous Finite element methods as well as nodal discontinuous Galerkin methods, including some of their hybrid versions. The library is freely available on GitHub[1]. Figure 5.2 shows on the left the sparsity pattern of the matrix in (5.2.7). We clearly observe that it

[1] Codes available at: https://github.com/vanzantom/GDGMatlab

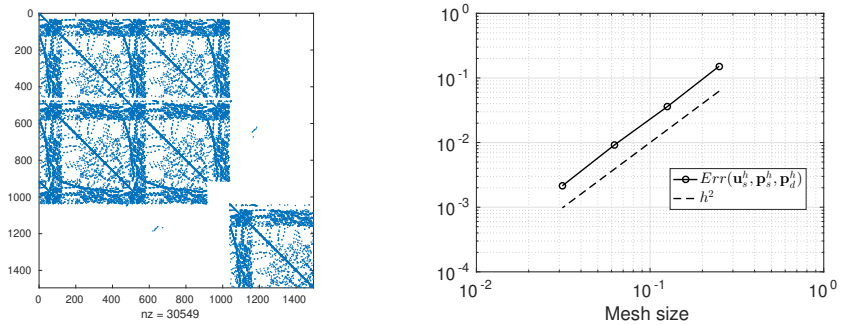


Figure 5.2: Sparsity pattern of the global Stokes-Darcy matrix on the left and error decaying as the mesh is refined on the right.

has a strong diagonal structure with small off-diagonal blocks, representing the coupling term C. On the right panel, we plot the error of the finite element approximation defined as

$$\text{err}(\mathbf{u}_s^h, \mathbf{p}_s^h, \mathbf{p}_d^h) := \|\mathbf{u}_s^h - \mathbf{u}_s^{\text{ex}}\|_{H_s} + \|\mathbf{p}_s^h - p_s^{\text{ex}}\|_{Q_s} + \|\mathbf{p}_d^h - p_d^{\text{ex}}\|_{H_d}.$$

The error decays as h^2, where h is a measure of the mesh size, in agreement with the classical results for the approximation error of the Taylor-Hood finite element space for the Stokes unknowns, and of quadratic finite element spaces for the Darcy pressure. Instead of solving directly the large system (5.2.7), it is naturally to use the sparsity pattern shown in Figure 5.2, to define a domain decomposition method, which solves alternately a saddle point problem, related to the Stokes domain, and a Laplace problem concerning the Darcy pressure. To investigate further the decoupling algorithm (5.1.7), we first show the well-possessedness of the local subproblems.

Let us remark that the weak formulation (5.2.5) has been obtained using the coupling conditions

$$-\mathbf{n}\cdot T(\mathbf{u}_s, p_s)\cdot\mathbf{n} = (p_d)_\Gamma \quad \text{and} \quad (K\nabla p_d - \mathbf{g}_d)\cdot\mathbf{n} = -(\mathbf{u}_s\cdot n)_\Gamma.$$

Thus, since (5.2.5) admits a unique solution in $W\times Q_s$, a posteriori we can state that $-\mathbf{n}\cdot T(\mathbf{u}_s, p_s)\cdot\mathbf{n}\in\Lambda$, as it is equal to the trace of a function $p_d\in H_d$ and $\partial\Gamma\subset\overline{\Gamma}_d$, where p_d is equal to zero. Similarly we have $(K\nabla p_d - \mathbf{g}_d)\cdot\mathbf{n}\in\Lambda$. We now introduce the extension operator $\mathscr{E}_s : \Lambda\times(L^2(\Omega_s))^d \to H_s\times Q_s$ as $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s) = \mathscr{E}_s(\lambda_s, \mathbf{f}_s)$ where $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s)$ is the unique solution of

$$\begin{aligned}
&\text{Find } (\widetilde{\mathbf{u}}_s, \widetilde{p}_s)\in H_s\times Q_s \text{ such that}\\
&a_s(\widetilde{\mathbf{u}}_s, \mathbf{v}) + b_s(\mathbf{v}, \widetilde{p}_s) = \int_{\Omega_s}\mathbf{f}_s\cdot\mathbf{v} - \int_\Gamma \lambda_s(\mathbf{v}\cdot\mathbf{n}), \qquad \forall\mathbf{v}\in H_s,\\
&b_s(\widetilde{\mathbf{u}}_s, q_s) = 0, \qquad\qquad\qquad\qquad\qquad\qquad \forall q_s\in Q_s.
\end{aligned} \tag{5.2.10}$$

Similarly, we define the extension operator $\mathscr{E}_d : \Lambda\times(L^2(\Omega_d))^d \to H_d$ by $\widetilde{p}_d = \mathscr{E}_d(\lambda_d, \mathbf{g}_d)$

where $\widetilde{p}_d$ is the unique solution of

$$\begin{aligned} &\text{Find } \widetilde{p}_d \in H_d \text{ such that} \\ &a_d(\widetilde{p}_d, q_d) = \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d + \frac{1}{s_1}\int_\Gamma \lambda_d q_d, \quad \forall q_d \in H_d. \end{aligned} \tag{5.2.11}$$

To show that the operators $\mathscr{E}_s$ and $\mathscr{E}_d$ are well defined, we need the following Lemma.

Lemma 5.2.1. *The bilinear forms $a_s(\cdot,\cdot) : H_s \times H_s \to \mathbb{R}$ and $a_d(\cdot,\cdot) : H_d \times H_d \to \mathbb{R}$ are coercive and continuous.*

Proof. The bilinear form $a_d(\cdot,\cdot)$ is continuous since, using Cauchy-Schwarz and the trace inequality

$$\begin{aligned} |a_d(p_d, q_d)| &\le k_\infty \|p_d\|_{H_d}\|q_d\|_{H_d} + \frac{1}{s_1}\|p_d\|_{L^2(\Gamma)}\|q_d\|_{L^2(\Gamma)} \\ &\le 2\max\left(k_\infty, \frac{(C_{tr}^d)^2}{s_1}\right)\|p_d\|_{H_d}\|q_d\|_{H_d}, \end{aligned}$$

where $k_\infty = \max_{i=1,\dots,d}\sup_{\mathbf{x}\in\Omega_d} K_i(\mathbf{x})$ and C_{tr}^d is the continuity constant of the trace operator in Ω_d. Defining instead $k_0 := \min_{i=1,\dots,d}\inf_{\mathbf{x}\in\Omega_d} K_i(\mathbf{x}) > 0$, we have immediately coercivity since

$$k_0\|p_d\|_{H_d}^2 \le \int_{\Omega_d} K\nabla p_d \cdot \nabla p_d + \frac{1}{s_1}\int_\Gamma p_d^2 = a_d(p_d, p_d).$$

Now we consider $a_s(\cdot,\cdot)$. Calling C_{tr}^s the continuity constant of the trace operator in Ω_s we have

$$\begin{aligned} |a_s(\mathbf{u},\mathbf{v})| &\le 2\mu\int_{\Omega_s} |\nabla^s\mathbf{u} : \nabla^s\mathbf{v}| + \sum_{j=1}^{d-1}\int_\Gamma \frac{\mu}{\epsilon}|(\mathbf{u})_{\tau_j}(\mathbf{v})_{\tau_j}| + \int_\Gamma s_2|(\mathbf{u}\cdot\mathbf{n})(\mathbf{v}\cdot\mathbf{n})| \\ &\le 2\mu\|\mathbf{u}\|_{H_s}\|\mathbf{v}\|_{H_s} + (d-1)\frac{\mu}{\epsilon}\|\mathbf{u}\|_{L^2(\Gamma)}\|\mathbf{v}\|_{L^2(\Gamma)} + s_2\|\mathbf{u}\|_{L^2(\Gamma)}\|\mathbf{v}\|_{L^2(\Gamma)} \\ &\le \gamma\|\mathbf{u}\|_{H_s}\|\mathbf{v}\|_{H_s}, \end{aligned}$$

where $\gamma := 2\mu + (d-1)\frac{\mu}{\epsilon}(C_{tr}^s)^2 + s_2(C_{tr}^s)^2$. □

The coercivity follows from Korn's inequality with constant δ as

$$\begin{aligned} a_s(\mathbf{u},\mathbf{u}) &\ge 2\mu\delta\|\mathbf{u}\|_{H_s}^2 + \sum_{j=1}^{d-1}\|\mathbf{u}\cdot\tau_j\|_{L^2(\Gamma)}^2 + s_2\|\mathbf{u}\cdot\mathbf{n}\|_{L^2(\Gamma)}^2 \\ &\ge 2\mu\delta\|\mathbf{u}\|_{H_s}^2. \end{aligned}$$

Theorem 5.2.2. *Problems* (5.2.10) *and* (5.2.11) *are well-posed.*

Proof. We denote the functional on the right side of equation (5.2.11) by $\mathcal{G}_d(\cdot)$. Then using Cauchy- Schwarz, we have

$$\begin{aligned}|\mathcal{G}_d(q_d)| &\leq \|\mathbf{g}_d\|_{(L^2(\Omega_d))^d}\|\nabla q_d\|_{L^2(\Omega_d)} + \frac{1}{s_1}\|\lambda_d\|_{L^2(\Gamma)}\|q_d\|_{L^2(\Gamma)},\\ &\leq 2\max\left(\|\mathbf{g}_d\|_{(L^2(\Omega_d))^d}, \frac{1}{s_1}C_{tr}^d\|\lambda_d\|_{L^2(\Gamma)}\right)\|q_d\|_{H_d}.\end{aligned}$$

It follows that $\mathcal{G}_d(\cdot)$ is a continuous functional on H_d and thus, since $a_d(\cdot,\cdot)$ is continuous and coercive, the Lax-Milgram theorem guarantees that the problem (5.2.11) is well-posed, i.e. there exists a unique $p_d \in H_d$ solution of problem (5.2.11). Regarding (5.2.10), we rely on Brezzi's theory for saddle-point problems. We call $\mathcal{G}_s(\cdot)$ the functional on the right hand side and we have

$$\begin{aligned}|\mathcal{G}_s(\mathbf{v})| &\leq \|\mathbf{f}_s\|_{(L^2(\Omega_s))^d}\|\mathbf{v}\|_{L^2(\Omega_s)} + \|\lambda_s\|_{L^2(\Gamma)}\|\mathbf{v}\cdot\mathbf{n}\|_{L^2(\Gamma)},\\ &\leq 2\max\left(C^p\|\mathbf{f}_s\|_{(L^2(\Omega_s))^d}, C_{tr}^s\|\lambda_s\|_{L^2(\Gamma)}\right)\|\mathbf{v}\|_{H_s},\end{aligned}$$

where C^p is the Poincaré constant. On the other hand we have the continuity of $b_s(\cdot,\cdot)$,

$$|b_s(\mathbf{u}_s, q_s)| \leq \int_{\Omega_s} |q_s\nabla\cdot\mathbf{u}_s| \leq \|q\|_{Q_s}\|\mathbf{u}_s\|_{H_s},$$

and, from Proposition 5.3.2 in [139], the inf-sup condition

$$\forall\, q \in Q_s, \quad \exists \mathbf{v} \in H_s, \quad \mathbf{v} \neq 0 : b_s(\mathbf{v}, q) \geq \beta\|\mathbf{v}\|_{H_s}\|q\|_{Q_s}$$

holds. We remark that $a_s(\cdot,\cdot)$ is continuous and coercive over all H_s due to Lemma 5.2.1, thus Brezzi's theory guarantees that (5.2.10) has an unique solution. □

5.3 One-level optimized Schwarz methods

In [53], the authors perform a Fourier analysis of the OSM (5.1.7) for a two dimensional problem ($d = 2$). In their analysis, they considered unbounded domains where one can use the Fourier transform. A separation of variables technique, as in [95], is not feasible unfortunately, since no analytical expression is available for the eigenvectors of the Stokes operator in bounded domains with Dirichlet boundary conditions. Furthermore, to simplify the calculations the authors assumed that $K = \mathrm{diag}(\eta_1, \eta_2)$ with $\eta_j > 0$, $j = 1, 2$ and since the solutions are required to go to zero as $y \to \infty$, they set $\Gamma_d^N = \emptyset$. They finally obtained that the convergence factor of algorithm (5.1.7) is

$$\rho(k, s_1, s_2) = \left|\frac{2\mu|k| - s_1}{2\mu|k| + s_2} \cdot \frac{1 - s_2\sqrt{\eta_1\eta_2}|k|}{1 + s_1\sqrt{\eta_1\eta_2}|k|}\right|, \tag{5.3.1}$$

for all the Fourier frequencies $k \in \mathbb{R}$. The optimal choice $s_1 = 2\mu|k|$ and $s_2 = \frac{1}{\sqrt{\eta_1\eta_2|k|}}$ would lead to a direct method which converges in just two iterations; however this choice corresponds to nonlocal operators once backtransformed[2]. Therefore a more practical choice

[2] See the extensive discussion in Section 2.1

is to set $s_1 = 2\mu p$ and $s_2 = \frac{1}{\sqrt{\eta_1\eta_2}p}$ for some $p \in \mathbb{R}$. An equivalent choice of optimized parameters has been treated in [53] where the authors obtained the following result.

Theorem 5.3.1 (Proposition 3.3 in [53])**.** *The unique solution of the min-max problem*

$$\min_{p} \max_{k\in[k_{\min},k_{\max}]} \rho(k,p), \tag{5.3.2}$$

is given by the unique root of the non linear equation $\rho(k_{\min},p) = \rho(k_{\max},p)$.

One could also consider double sided optimized transmission conditions, choosing $s_1 = 2\mu p$ and $s_2 = \frac{1}{\sqrt{\eta_1\eta_2}q}$ with $p,q \in \mathbb{R}$. In [54], the authors propose to choose the couple p,q such that $\rho(k_{\min},p,q) = \rho(\widehat{k},p,q) = \rho(k_{\max},p,q)$, i.e. they impose equioscillation. Even though often the solution of such min-max problems is indeed given by equioscillation, a priori there is no reason why this should be the case also for the Stokes-Darcy coupling. In fact for heterogenous problems, it has been observed that there can exist a couple of parameters which satisfies the equioscillation property, but leads to a non optimized convergence or even to a divergent method, see for instance Theorem 2.1.9, Theorem 2.4.6 and reference [78]. In Theorem 5.3.2 we refine Proposition 1 of [54].

Theorem 5.3.2. *The solutions of the min-max problem*

$$\min_{p,q\in\mathbb{R}} \max_{k\in[k_{\min},k_{\max}]} \rho(k,p,q) = \min_{p,q\in\mathbb{R}} \max_{k\in[k_{\min},k_{\max}]} 2\mu\sqrt{\eta_1\eta_2}\left|\frac{k-p}{1+2\mu\sqrt{\eta_1\eta_2}kp}\cdot\frac{k-q}{1+2\mu\sqrt{\eta_1\eta_2}kq}\right|, \tag{5.3.3}$$

are given by two pairs (p_i^*,q_i^*), $i=1,2$ *which satisfy the non linear equations* $|\rho(k_{\min},p_i^*,q_i^*)| = |\rho(\widehat{k},p_i^*,q_i^*)| = |\rho(k_{\max},p_i^*,q_i^*)|$, $\widehat{k}$ *being an interior maximum. Moreover* $p_2^* = q_1^*$ *and* $q_2^* = p_1^*$.

Proof. The proof is based on arguments presented in the proofs of Theorem 2.1.9 and Theorem 3 in [95]. We outline the main steps. We first observe that $\rho(k,p,q)$ is invariant under $p \leftrightarrow q$, hence we consider only $p < q$ and moreover $\rho(k,p,q) = 0$ for $k = q$ and $k = p$. The partial derivatives with respect to the parameters satisfy $\operatorname{sign}(\partial_p\rho) = \operatorname{sign}(p-k)$ and $\operatorname{sign}(\partial_q\rho) = \operatorname{sign}(q-k)$, therefore at optimality we conclude that p,q lie in $[k_{\min},k_{\max}]$, see the proof of Theorem 1 in [95]. Solving $\partial_k\rho = 0$, we get that there exists a unique interior maximum $\widehat{k}$, with $p < \widehat{k} < q$, so that we can restrict $\max_{k\in[k_{\min},k_{\max}]}\rho(k,p,q) = \max\{\rho(k_{\min},p,q),\rho(\widehat{k},p,q),\rho(k_{\max},p,q)\}$. Repeating the same arguments of Lemma 2.1.8, we obtain that at the optimum we must have $\rho(k_{\min},p,q) = \rho(k_{\max},p,q)$, so that we can express q as function of p and we can restrict the study to $\min_p \max\{\rho(k_{\min},p,q(p)),\rho(\widehat{k},p,q(p))\}$. Defining $\delta := 2\mu\sqrt{\eta_1\eta_2}$, the equioscillation constraint is equivalent to

$$l(p) := \frac{k_{\min}-p}{1+\delta k_{\min}p}\frac{1+\delta k_{\max}p}{k_{\max}-p} = \frac{k_{\max}-q(p)}{1+\delta q(p)k_{\max}}\frac{1+\delta q(p)k_{\min}}{k_{\min}-q(p)} =: g(p). \tag{5.3.4}$$

Since $\partial_p l(p) < 0$ and $\partial_p g(p) > 0$, $q(p)$ must be a decreasing function of p so that equation (5.3.4) is satisfied. Then using the sign of the derivatives of ρ with respect to p and q and

the explicit expression of $q(p)$, we have $\frac{d\rho(k_{\min},p)}{dp} > 0$ and $\frac{d\rho(\widehat{k},p)}{dp} < 0$ for $k_{\min} < p < q(p)$. These observations are sufficient to conclude that the solution of $\min_p \max\{\rho(k_{\min},p,q(p)),\rho(\widehat{k},p,q(p))\}$ is given by the unique p_1^*, such that $\rho(k_{\min},p_1^*,q(p_1^*)) = \rho(\widehat{k},p_1^*,q(p_1^*))$ and q_1^* given by $q_1^* = q(p_1^*)$. Due to the invariance $p \leftrightarrow q$, we get the same results in the case $q < p$ and we conclude that the other couple satisfies $p_2^* = q_1^*$ and $q_2^* = p_1^*$. □

In [53, 54], the authors studied extensively the methods obtained from Theorems 5.3.1-5.3.2 as preconditioners for GMRES. They observed that these optimized parameters do not lead to an optimized convergence and they proposed to minimize the L^1 norm of the convergence factor, that is

$$\min_p \frac{1}{k_{\max}-k_{\min}} \int_{k_{\min}}^{k_{\max}} \rho(k,p)\,dk. \tag{5.3.5}$$

The motivation lies in the assumption that the Krylov method can take care of isolated slow frequencies, and thus it would be better to have a convergence factor that is very small for a large set of frequencies with possibly high peaks. A similar approach was first discussed in [91] for the Helmholtz problem as the optimized Schwarz method does not converge for the Helmholtz frequency ω, and thus the authors proposed to solve the optimization problem $\min_p \max_{k\in[k_{\min},\omega^-]\cup[\omega^+,k_{\max}]} \rho(k,p)$. However, such a bad performance of the optimized parameters obtained from a min-max problem does not have comparison in the literature, and thus we investigated it in detail in [98]. We consider the domains $\Omega_s = (0,1)\times(0,1)$, $\Omega_d = (0,1)\times(-1,0)$ and a uniform structured mesh with mesh size $h = 0.02$, so that $k_{\min} = \pi$ and $k_{\max} = \pi/h$. We discretize the corresponding error equations of (5.1.7) with Taylor-Hood finite elements $\mathbb{P}_2^2 - \mathbb{P}_1$ for the Stokes unknowns and $\mathbb{P}_2$ elements for the Darcy pressure. The physical parameters are set equal to $\mu = 0.1$, $K = \mathrm{diag}(\eta_1,\eta_2)$, with $\eta_1 = \eta_2 = 1$. The stopping criterion for the iterative method is $\|u^n\|_{H^1} + \|v^n\|_{H^1} + \|p_s^n\|_{L^2} + \|p_d^n\|_{H^1} < 10^{-9}$ and similarly for GMRES the tolerance is 10^{-9}. Figure 5.3 shows the number of iterations to reach convergence. On the left panel we show with a circle the optimized parameter p obtained from Theorem 5.3.1 and with a square the optimized p obtained solving (5.3.5). We observe that indeed the solution of (5.3.5) leads to a faster convergence than the classical approach of Theorem 5.3.1 for preconditioned GMRES. This is in agreement with the results reported in [54, 53], where it has been shown numerically that the solution of (5.3.5) leads to an equivalent or faster convergence than Theorem 5.3.1 for a wide range of parameters. However, we remark that (5.3.5) leads to a faster method than (5.3.2) also for the iterative method and not only under Krylov acceleration! On the right panel of Figure 5.3 we observe that also Theorem 5.3.2 does not lead to an optimized convergence and the symmetry of the parameters has disappeared. To better understand the behaviour of the method, we set as initial condition one by one the sine functions, which correspond to the restriction of the Fourier basis $\{e^{-ikx}\}_k$ on bounded domains with Dirichlet boundary conditions. We then compute numerically an approximation of the convergence factor defining $\rho_v(k,p) = \left(\frac{\|v^3\|_{H^1}}{\|v^1\|_{H^1}}\right)$, $\rho_{p_d}(k,p) = \left(\frac{\|p_d^3\|_{H^1}}{\|p_d^1\|_{H^1}}\right)$, where v is the second component of the Stokes velocity and p_d is the Darcy pressure.

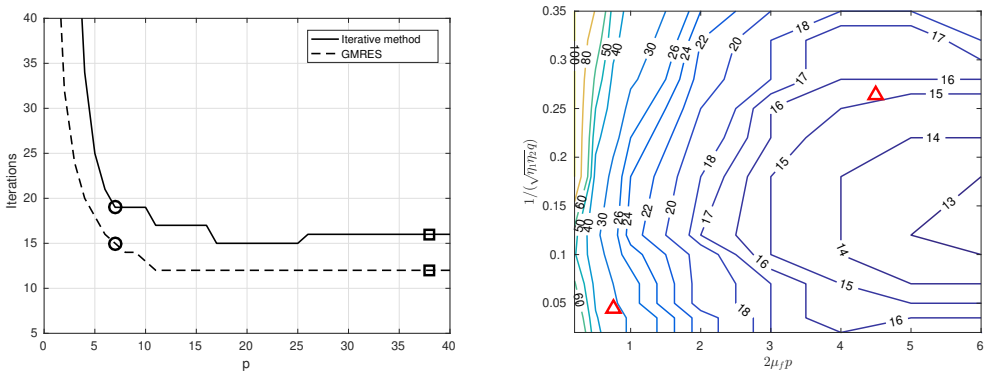


Figure 5.3: Number of iterations to reach the tolerance 10^{-9} for different optimized parameters. On the left, the circle represents the solution of Theorem 5.3.1, the square corresponds to the solution of (5.3.5). On the right the triangles correspond to the double solutions of Theorem 5.3.2 and the contour plot refers to the iterative method.

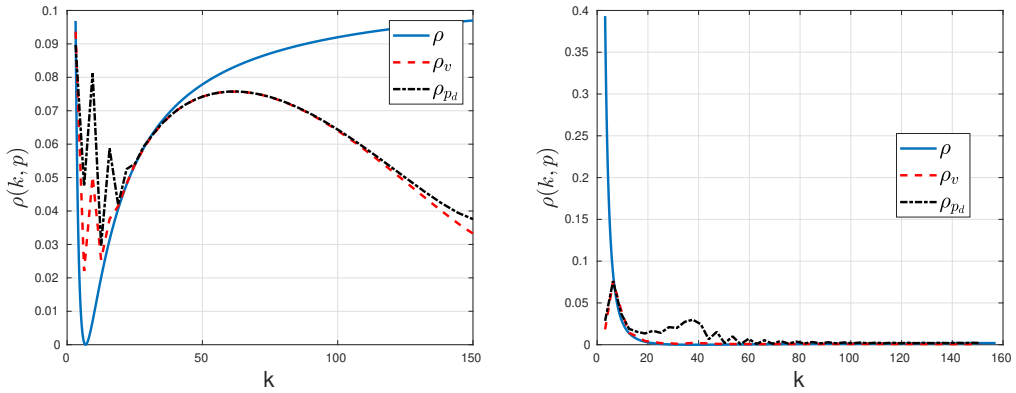


Figure 5.4: Comparison of the theoretical and numerical convergence factors. On the left, optimized parameter from Theorem 5.3.1 and on the right, optimized parameter from (5.3.5).

From the results presented in Figure 5.4, we observe two major issues: the first one is a poor approximation of high frequencies. This is due to the fact that the discrete approximation based on the finite element spaces $\mathbb{P}_2^2 - \mathbb{P}_1 - \mathbb{P}_2$ is not capable of representing properly the exponential boundary layer of the high frequencies near the interface. This can also be observed for the classical Laplace equation. We propose two remedies which can also be combined. We could first raise the order of the approximation of the finite element spaces to $\mathbb{P}_3^2 - \mathbb{P}_2 - \mathbb{P}_3$ and/or refine the mesh in the normal direction to the interface. Both remedies improve the representation of the high frequencies. The second issue lies in a unusual oscillatory behaviour of the low, odd frequencies. For instance, in the right panel of Figure 5.4, the first frequency $\sin(\pi x)$ is transformed after one iteration into a

complicated combination of higher frequencies so that actually the parameter p makes the method much faster than the theory predicts. In [98], we claimed that the reason for this phenomenon was that the sines do not form a separated variable solution for the Stokes operator with Dirichlet boundary condition. This argument was based on our experience developed with the tangential advection case, discussed in Section 2.2.3, where we have remarked that the unbounded analysis leads to inefficient optimized parameters since the two equations lack a common eigenbasis. While our statements remain true, we now believe a further reason lies in a compatibility condition which must be satisfied whenever dealing with a Stokes problem with Dirichlet boundary condition all along the boundary. Indeed, the imposed velocity field must satisfy

$$0 = \int_{\Omega_s} \text{div}\mathbf{u}_s = \int_{\partial\Omega_s} \mathbf{u}_s \cdot \mathbf{n}. \tag{5.3.6}$$

Working with the error equation, and imposing a velocity field only on Γ, condition (5.3.6) is satisfied by the even Fourier frequencies, but not by the odd frequencies. Thus, not only the odd Fourier frequencies are not eigenvectors of the Stokes operator, but they are even incompatible boundary condition, and this explains why the lowest Fourier mode is immediately transformed into a combination of higher even frequencies after just one iteration! We conclude that it is not possible to diagonalize the iteration as the formula of the convergence factor (5.3.1) assumes. In Section 5.3.1, we discuss how to recover optimized parameters for the Dirichlet case, and we show that actually imposing a normal stress condition on the upper boundary, allows one to recover good estimates also with the Fourier approach, since the velocity field does not need to satisfy equation (5.3.6) anymore.

We conclude this subsection considering the Stokes-Darcy system (5.1.7) with periodic boundary conditions on the vertical edges in order to make the bounded problem as similar as possible to the unbounded case. In this setting there exists a separated variable solution for the Stokes problem involving the Fourier basis $\{e^{-ikx}\}_k$, see [140]. In Figure 5.5 we show both the numerical and theoretical convergence factors computed for even frequencies $\{\sin(2k\pi x)\}_k$. The same results are obtained using the other periodic frequencies $\{\cos(2k\pi x)\}_k$. Comparing with Figure 5.4, we observe that now we have an excellent agreement between the numerical and theoretical convergence factors and thus we would expect that the optimized parameters from the min-max theorems provide optimized convergence. We thus start the optimized Schwarz method (5.1.7) with initial guesses given by a linear combination of periodic sine and cosine functions multiplied by random coefficients. Figure 5.6 shows that both Theorem 5.3.1 and 5.3.2 now lead to optimized convergence for the iterative method (5.1.7) and we also observe the symmetry of the optimized parameters in the right panel as Theorem 5.3.2 predicts. However concerning GMRES, we note that the optimized parameter from Theorem 5.3.1 is still a bit too small. This can be understood studying the eigenvalues of the preconditioned matrix system which are shown in Fig 5.7. Analyzing the large real eigenvalue, we have observed that the corresponding eigenvector is given by a zero velocity field $\mathbf{u}_s$, a constant pressure p_s and a linear Darcy pressure p_d. This constant mode is actually not

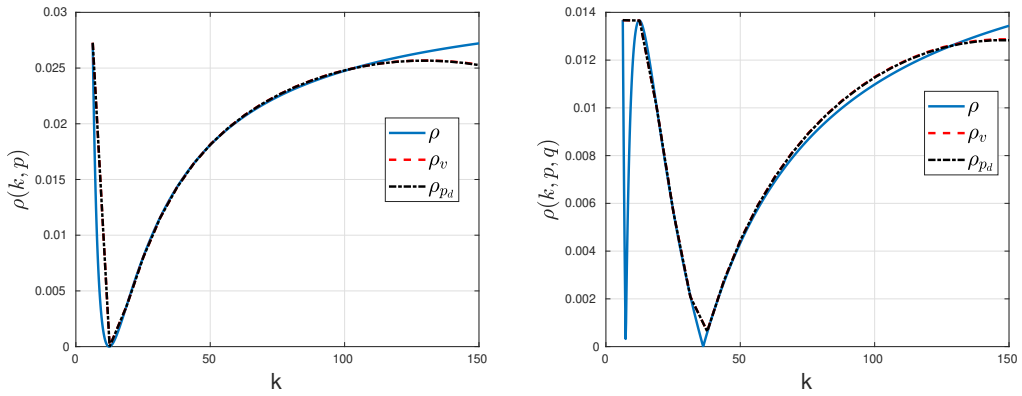


Figure 5.5: Comparison of the theoretical and numerical convergence factors. On the left for the single sided optimized parameter from Theorem 5.3.1 and on the right one for the double sided parameters of Theorem 5.3.2. The minimum frequency is now $k_{\min} = 2\pi$.

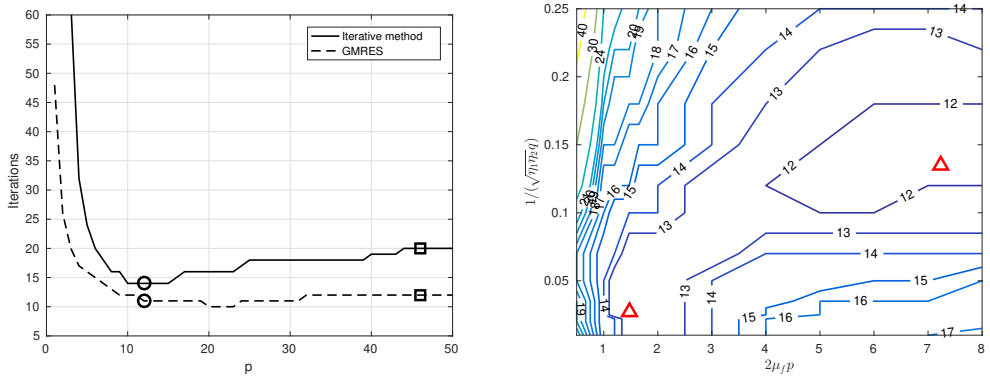


Figure 5.6: Number of iterations to reach the tolerance 10^{-9} for different optimized parameters. On the left, the circle represents the solution of Theorem 5.3.1, the square corresponds to the approach of (5.3.5). On the right the triangles correspond to the double solutions of Theorem 5.3.2 and the contour plot refers to the iterative method.

treated by the unbounded Fourier analysis and it is not present in our initial guess for the iterative method. Defining the functions $p_d^n = D^n(y+L)$ and $p_f = P^n$ with $P, D \in \mathbb{R}$ and L is the vertical length of the subdomains, and inserting them into (5.1.7), we obtain a convergence factor $\rho(k=0,p) := \frac{1-s_2}{1+s_1}$. Solving numerically the min-max problem $\min_p \max_{k\in\{0\}\cup[k_{\min},k_{\max}]} \rho(k,p)$ we obtain the equioscillation between $\rho(0,p)$ and $\rho(k_{\min},p)$ and a numerical value of $p \approx 48$. In the right panel of Fig. 5.7 we start the method with a totally random initial guess and this shows that taking into account the constant mode actually makes our analysis exact.

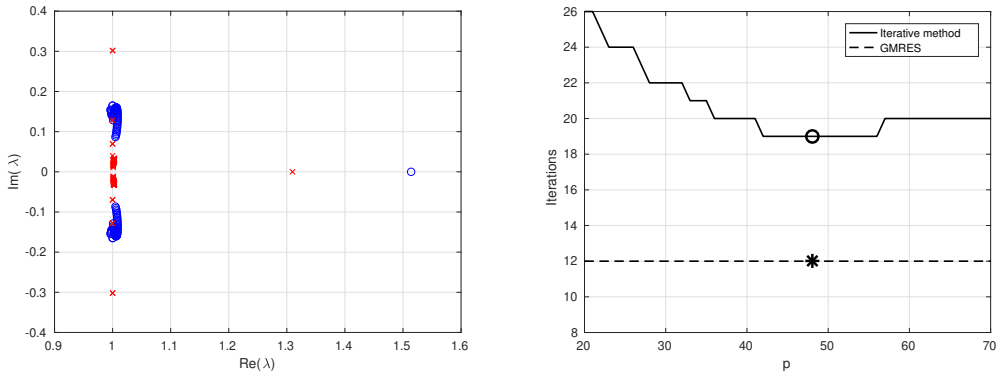


Figure 5.7: On the left panel, the blue circles correspond to first 100 eigenvalues of the preconditioned volume matrix in the case with the optimized parameter of Theorem 5.3.1 and the red crosses in the case using the solution of (5.3.5). On the right panel we show the number of iterations to reach convergence with periodic boundary conditions and with a random initial guess. The circle corresponds to the solution of Theorem 5.3.1 and the star to the value of p such that we have the minimal residual of GMRES.

5.3.1 Application of the probing technique

In this subsection, we apply the probing technique introduced in Section 2.5 to the Stokes-Darcy system. To do so, we need to reformulate the Stokes-Darcy system in a substructured form, that is, as an equation involving a Steklov-Poincaré operator over an interface variable. Then we apply the ADI method to solve the Steklov-Poincaré system similarly to (1.3.16). To the best of author's knowledge, a first substructured formulation of the coupled Stokes-Darcy system has been provided in [52]. We consider the geometrical setup described by Figure 5.1 and we set $\Gamma_d^N = \emptyset$ and $\epsilon = 0$, that is, we impose a zero tangential velocity and we introduce the space

$$H_s^\tau := \left\{ \mathbf{u}_s \in (H^1(\Omega_s))^d : \mathbf{u}_s = 0 \text{ on } \Gamma_s \text{ and } \mathbf{u}_s \cdot \tau_j = 0, j = 1, \dots, d-1 \text{ on } \Gamma \right\}.$$

The multidomain weak formulation of the Stokes-Darcy system, counterpart of (1.3.2) for the Laplace equation, is [52, Proposition 2.4.1]

$$\begin{array}{ll} \text{Find } \mathbf{u}_s \in H_s^\tau,\ p_s \in Q_s,\ p_d \in H_d \text{ such that:} & \\ \widetilde{a}(\mathbf{u}_s, \mathbf{v}) + b_s(\mathbf{v}, p_s) = \int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{v}, & \forall \mathbf{v} \in H_0^1(\Omega_s), \\ b_s(\mathbf{u}_s, q_s) = 0, & \forall q_s \in Q_s, \\ \widetilde{a}_d(p_d, q_d) = \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d, & \forall q_d \in H_0^1(\Omega_d), \\ \int_\Gamma p_d \eta = \int_{\Omega_s} \mathbf{f}_s \cdot E^s \eta - \widetilde{a}(\mathbf{u}_s, E^s \eta) - b_s(E^s \eta, p_s), & \forall \eta \in \Lambda. \\ \int_\Gamma (\mathbf{u}_s \cdot \mathbf{n}) \eta = \widetilde{a}_d(p_d, E^d \eta) - \int_{\Omega_d} \mathbf{g}_d \cdot \nabla E^d \eta, & \forall \eta \in \Lambda, \end{array} \tag{5.3.7}$$

where E^s and E^d are general continuous extension operators from Λ to respectively H_s^τ and H_d. Equations $(5.3.7)_{4,5}$ are the coupling conditions along the interface and they play the same role as the continuity of the traces and of the normal derivatives in $(1.3.2)_{3,4}$. To

obtain the Steklov-Poincaré equation for (1.3.2), we selected a primal interface function, the trace, so that we could decompose each subdomain solution as the sum of two functions, one which is the harmonic extension of the trace, and another one which takes into account the force term, that is $u_i = \mathscr{H}_i(u_{|\Gamma}) + \mathscr{G}_i(f_i)$. In order for the u_i to be solutions, they still need to satisfy the continuity of the dual variable, that is the normal derivative, and this is exactly what the Steklov-Poincaré equation imposes. We are now going to follow the same logical path for the Stokes-Darcy system. We remark it is possible to use two different interface variables, $\lambda := \mathbf{u}_s \cdot \mathbf{n} = (-K\nabla p_d + \mathbf{g}_d)\cdot\mathbf{n}$ and $\varphi := p_d = -\mathbf{n}\cdot T(\mathbf{u}_s, p_s)\cdot\mathbf{n}$. In contrast to the Laplace case, the function λ corresponds to the trace of the Stokes velocity and to the normal derivative of the Darcy pressure, while φ corresponds to the trace of the Darcy pressure and to the normal component of the normal stress for the Stokes domain. We could choose either λ or φ as primal variables, however choosing λ leads to some technical difficulties. First, we would need to solve a Dirichlet Stokes boundary value problem which, in order to have a unique solution, requires to deal with a quotient space for the pressure field. We would have then to compute the normal stress along Γ which depends on the pressure defined up to a constant. Moreover, for a complete Dirichlet Stokes problem, the boundary condition needs to satisfy the compatibility condition (5.3.6) which, supposing homogeneous boundary conditions along Γ_s, implies that $\int_\Gamma \lambda = 0$. Thus we have a further constraint on the interface variable. For these reasons, we choose as interface variable the Darcy pressure φ, and we refer the reader to [55, Section 5.1] for a detailed discussion concerning the interface variable λ.

We define the operator $\mathscr{H}_s : \Lambda \to H_s^\tau \times Q_s$ such that $\mathscr{H}_s(\varphi) = (\mathscr{H}_s^1(\varphi), \mathscr{H}_s^2(\varphi))$ satisfies

$$\begin{aligned} \widetilde{a}_s(\mathscr{H}_s^1(\varphi), \mathbf{v}) + b_s(\mathbf{v}, \mathscr{H}_s^2(\varphi)) &= -\int_\Gamma \varphi(\mathbf{v}\cdot\mathbf{n}), \qquad && \forall \mathbf{v} \in H_s^\tau, \\ b_s(\mathscr{H}_s^1(\varphi), q) &= 0, && \forall q \in Q_s, \end{aligned} \tag{5.3.8}$$

and the function $(\mathbf{u}_s^0, p_s^0) \in H_s^\tau \times Q_s$ solution of the boundary value problem with zero normal stress condition along Γ [3]

$$\begin{aligned} \widetilde{a}_s(\mathbf{u}_s^0, \mathbf{v}) + b_s(\mathbf{v}, p_s^0) &= \int_{\Omega_s} \mathbf{f}\cdot\mathbf{v}, \qquad && \forall \mathbf{v} \in H_s^\tau, \\ b_s(\mathbf{u}_s^0, q) &= 0, && \forall q \in Q_s. \end{aligned} \tag{5.3.9}$$

Concerning the Darcy domain, we define $\mathscr{H}_d : \Lambda \to H_d$ such that $\mathscr{H}_d(\varphi)$ satisfies

$$\begin{aligned} \widetilde{a}_d(\mathscr{H}_d(\varphi), q_d) = 0, \quad \forall q_d \in H_0^1(\Omega_d), \\ \mathscr{H}_d(\varphi) = \varphi, \quad \text{on } \Gamma, \end{aligned} \tag{5.3.10}$$

and the function $p_d^0 \in H_0^1(\Omega_d)$ solution of

$$\widetilde{a}_d(p_d^0, q_d) = \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d, \quad \forall q_d \in H_0^1(\Omega_d). \tag{5.3.11}$$

[3] Using the interface variable λ, we would have to solve a Dirichlet Stokes problem, introducing the quotient space $Q_s^0 := \left\{ q \in Q_s : \int_{\Omega_s} q = 0 \right\}$.

Knowing a priori $(\mathbf{u}_s^{\text{ex}}, p_s^{\text{ex}}, p_d^{\text{ex}})$ solution of (5.3.7) and defining $\varphi = p_{d,|\Gamma}^{\text{ex}}$, we would have that $(\mathbf{u}_s^{\text{ex}}, p_s^{\text{ex}}) = (\mathscr{H}_s^1(\varphi) + \mathbf{u}_s^0, \mathscr{H}_s^2(\varphi) + p_s^0)$ and $p_d^{\text{ex}} = \mathscr{H}_d(\varphi) + p_d$. However we do not know the solution of (5.3.7) a priori. Nevertheless, we have that $(\mathscr{H}_s^1(\varphi) + \mathbf{u}_s^0, \mathscr{H}_s^2(\varphi) + p_s^0)$ and $\mathscr{H}_d(\varphi) + p_d$ satisfy already the first four equations of (5.3.7). We use the fifth to obtain an equation for the unknown interface variable φ. Replacing $(\mathscr{H}_s^1(\varphi) + \mathbf{u}_s^0, \mathscr{H}_s^2(\varphi) + p_s^0)$ and $\mathscr{H}_d(\varphi) + p_d$ into $(5.3.7)_5$, we get

$$\langle \mathscr{S}\varphi, \eta \rangle = \langle \chi, \eta \rangle, \tag{5.3.12}$$

where $\mathscr{S} = \mathscr{S}_s + \mathscr{S}_d$ are defined as

$$\langle \mathscr{S}_s\varphi, \eta \rangle := -\int_\Gamma (\mathscr{H}_s^1(\varphi) \cdot \mathbf{n})\eta, \quad \langle \mathscr{S}_d\varphi, \eta \rangle := \widetilde{a}_d(\mathscr{H}_d(\varphi), \mathscr{H}_d(\eta)), \tag{5.3.13}$$

and $\chi \in \Lambda'$ is such that

$$\langle \chi, \eta \rangle = \int_\Gamma (\mathbf{u}_s^0 \cdot \mathbf{n})\eta + \int_{\Omega_d} \mathbf{g}_d \cdot \nabla \mathscr{H}_d(\eta) - \widetilde{a}_d(p_d^0, \mathscr{H}_d(\eta)).$$

We emphasise that we set the extension operator $\mathscr{E}_d = \mathscr{H}_d$. Solving equation (5.3.12) permits to obtain the exact trace pressure φ, from which we can then recover the exact local solutions performing subdomain solves. For a proof of existence and uniquess of the solution to (5.3.12) we refer to [52, Proposition 2.6.1]. Repeating the same calculations presented at the end of Section 1.3.4 for the Laplace equation, we can rewrite the transmission conditions of (5.1.7) at the weak level as

$$\begin{aligned} \langle (s_2\mathscr{S}_s + I)\lambda^n, \eta \rangle &= \langle (I - s_2\mathscr{S}_d)\lambda^{n-1}, \eta \rangle + \langle s_2\chi, \eta \rangle, \quad \forall \eta \in \Lambda, \\ \langle (s_1\mathscr{S}_d + I)\lambda^n, \eta \rangle &= \langle (I - s_1\mathscr{S}_s)\lambda^{n-1}, \eta \rangle + \langle s_1\chi, \eta \rangle, \quad \forall \eta \in \Lambda, \end{aligned} \tag{5.3.14}$$

which leads to the fixed point iteration for the error equation

$$\lambda^{n+1} = (s_2\mathscr{S}_s + I)^{-1}(I - s_2\mathscr{S}_d)(s_1\mathscr{S}_d + I)^{-1}(I - s_1\mathscr{S}_s)\lambda^{n-1} = T(s_1, s_2)\lambda^{n-1}.$$

We are now perfectly in the framework discussed in Section 2.5. We introduce finite dimensional approximations Σ_s, Σ_d of the operators $\mathscr{S}_s$ and $\mathscr{S}_d$. We thus choose a set of probing vectors $\{\mathbf{x}_j\}$ with j in some index set $\mathscr{K}$, and we aim to minimize numerically the ratio

$$\min_{s_1, s_2 \in \mathbb{R}} \max_{j \in \mathscr{I}} \frac{\|s_2\Sigma_s\mathbf{x}_j - M_\Gamma\mathbf{x}_j\|}{\|s_1\Sigma_d\mathbf{x}_j + M_\Gamma\mathbf{x}_j\|} \frac{\|s_1\Sigma_d\mathbf{x}_j - M_\Gamma\mathbf{x}_j\|}{\|s_2\Sigma_s\mathbf{x}_j + M_\Gamma\mathbf{x}_j\|}, \tag{5.3.15}$$

where M_Γ is the mass matrix over the interface defined as $(M_\Gamma)_{i,j} := \int_\Gamma \phi_i\phi_j$, and $\{\phi_k\}$ is a basis for the finite element approximation of Λ.

We set the physical parameters equal to $\mu = 0.1$, $\eta_1 = \eta_2 = 1$. We consider the probing vectors $\mathbf{x}_j$ such that $\mathbf{x}_j$ is an approximation of $\sin(\pi k x)$, and k belongs to a index set $\mathscr{K}$. The optimized parameters are equal to $s_2 = \frac{1}{\sqrt{\eta_1\eta_2 p}}$ and $s_1 = 2\mu p$, with $p \in \mathbb{R}$. On the left of Figure 5.8, we consider Dirichlet boundary conditions all along $\partial\Omega$. We consider two different sets, $\mathscr{K}_1 := \{1, \sqrt{N}, N\}$ and $\mathscr{K}_2 := \{1, 2, N\}$, where N is the number of degrees of

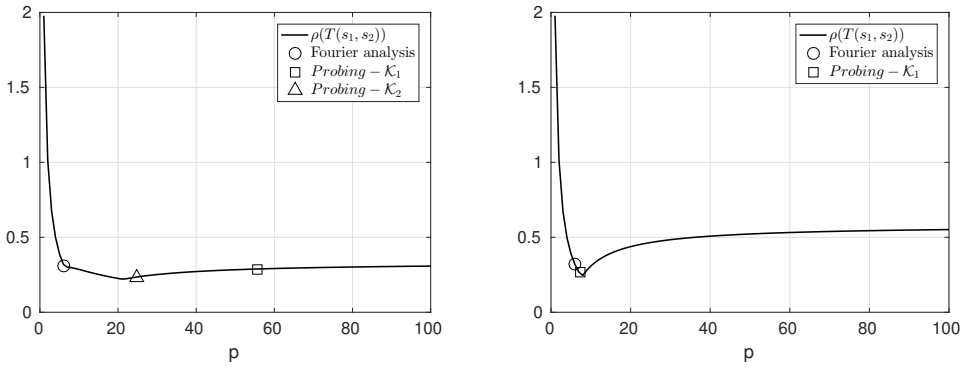


Figure 5.8: Comparison between the spectral radius of the iteration operator $\mathcal{T}(s_1,s_2)$ and several estimated parameters through the probing technique.

freedom along Γ. As we discussed, the Fourier analysis does not provide good estimates. However, also probing with the set index $\mathcal{K}_1$ does not provide good results. Guided by the left panel of Figure 5.4 where we saw that the initialisation with the first even frequency lead to the slowest numerical convergence factor, and since the first odd frequency does not satisfy (5.3.6), we decided to probe with the set $\mathcal{K}_2$. We remark that the addition of the first even frequency in the index set leads to a very precise estimate of the optimal parameter. The choice of the set of probing vector is thus not trivial for the case of Dirichlet boundary conditions. We report that we also tried the power method approach described in Section 2.5, but we did not observe any significant improvement with respect to $\mathcal{K}_1$. We then impose a zero normal stress condition on the upper horizontal edge of Ω_s. Due to this conditions, the velocity field does not need to satisfy (5.3.6). We remark that both the Fourier analysis and the probing technique with index set $\mathcal{K}_1$ permit to get excellent optimized parameters. This experiment corroborates our statement that the compatibility condition is the key element for the failure of the Fourier analysis for the Stokes-Darcy coupling.

5.4 Two-level optimized Schwarz methods

In this Section, we discuss how to use the two-level OSM framework introduced in Chapter 3 to design an efficient two-level solver for the Stokes-Darcy coupling. We consider the following geometrical and physical setting. A Newtonian fluid is flowing in a domain $\Omega_s = (0,1)\times(0,1)$ which interacts through an interface $\Gamma = [0,1]\times\{0\}$ with a porous medium in a domain $\Omega_d = (0,1)\times(-1,0)$, see Fig 5.9. We suppose the fluid enters in Ω_s from the left with a velocity profile $\mathbf{u}_s = (y^3,0)^\top$, we impose $\mathbf{u}_s = (1,0)^\top$ on the top boundary, while a zero normal stress condition is imposed on the right boundary, that is $-\mathbf{n}\cdot(2\mu\nabla^s\mathbf{u}_s - p_s\mathbf{I})\cdot\mathbf{n} = 0$. Concerning the porous medium domain, we set a homogeneous Dirichlet boundary condition along $\partial\Omega_d\setminus\Gamma$.

We now consider the one-level OSM (5.1.7), and we explicitly express the dependence of

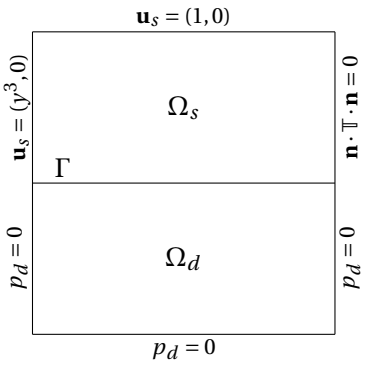


Figure 5.9: Geometry for the Stokes-Darcy problem.

λ_s and λ_d on the physical variables $\mathbf{u}_s, p_s, p_d$,

$$
\begin{aligned}
-\nabla \cdot (2\mu \nabla^s \mathbf{u}_s^n - p_s^n I) &= 0, \quad \text{in } \Omega_s, \\
\nabla \cdot \mathbf{u}_s^n &= 0, \quad \text{in } \Omega_s, \\
-\nabla \cdot K \nabla p_d^n &= 0, \quad \text{in } \Omega_d, \\
p_d^n - s_1 \left(K \nabla p_d^n \cdot \mathbf{n} \right) &= -\mathbf{n} \cdot (2\mu \nabla^s \mathbf{u}_s^{n-1} - p_s^{n-1} \mathbf{I}) \cdot \mathbf{n} + s_1 \mathbf{u}_s^{n-1} \cdot \mathbf{n} \quad \text{on } \Gamma, \\
-\mathbf{n} \cdot (2\mu \nabla^s \mathbf{u}_s^n - p_s^n \mathbf{I}) \cdot \mathbf{n} - s_2 \mathbf{u}_s^n \cdot \mathbf{n} &= p_d^{n-1} + s_2 \left(K \nabla p_d^{n-1} \cdot \mathbf{n} \right) \quad \text{on } \Gamma, \\
-\epsilon \tau_j \cdot (2\mu \nabla^s \mathbf{u}_s^n - p_s^n \mathbf{I}) \cdot \mathbf{n} &= \mu \mathbf{u}_s^n \cdot \tau_j \quad \text{on } \Gamma.
\end{aligned}
\tag{5.4.1}
$$

To obtain the enhanced matrix for the Stokes-Darcy coupling we consider the fixed point version of system (5.4.1) by letting $n \to \infty$. Then, using the bilinear forms introduced in (5.2.2), the weak formulation of system (5.4.1) is

$$
\begin{aligned}
a_s(\mathbf{u}_s, \mathbf{v}) + b_s(\mathbf{v}, p_s) - b_{SD}(p_d, \mathbf{v}) &= \langle \bar{f}, \mathbf{v} \rangle \qquad \forall \mathbf{v} \in H_s, \\
b_s(\mathbf{u}_s, q_s) &= 0 \qquad \forall q_s \in Q_s, \\
a_d(p_d, q_d) - b_{DS}(\mathbf{u}_s, p_s, q_d) &= 0 \qquad \forall q_d \in H_d,
\end{aligned}
\tag{5.4.2}
$$

where we have introduced the new coupling bilinear forms

$$
\begin{aligned}
b_{SD}(p_d, \mathbf{v}) &:= -\int_\Gamma (p_d + s_2 K \nabla p_d \cdot \mathbf{n}) \mathbf{v} \cdot \mathbf{n}, \\
b_{DS}(\mathbf{u}_s, p_s, q_d) &:= \int_\Gamma (\mathbf{u}_s \cdot \mathbf{n}) q_d - \frac{1}{s_1} \int_\Gamma \mathbf{n} \cdot (2\mu \nabla^s \mathbf{u}_s - p_s \mathbf{I}) \cdot \mathbf{n},
\end{aligned}
$$

and the functional $\bar{f}$ takes into account the non homogeneous Dirichlet conditions. A finite element discretization of (5.4.2) leads to the discretize system

$$
\begin{pmatrix}
\begin{pmatrix} A_s & B_s \\ B_s^\top & 0 \end{pmatrix} & -B_{SD} \\
-B_{DS} & A_d
\end{pmatrix}
\begin{pmatrix}
\begin{pmatrix} \mathbf{u}_s \\ \mathbf{p}_s \end{pmatrix} \\
\mathbf{p}_d
\end{pmatrix}
=
\begin{pmatrix}
\begin{pmatrix} \bar{\mathbf{f}} \\ \mathbf{0} \end{pmatrix} \\
\mathbf{0}
\end{pmatrix}.
\tag{5.4.3}
$$

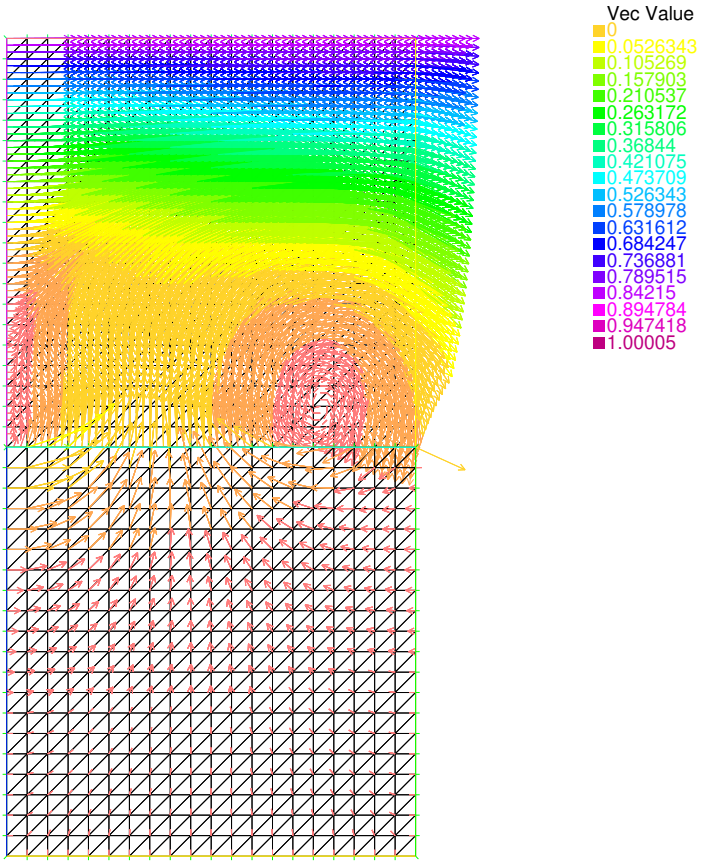


Figure 5.10: Plot of the velocity field solution to the problem described in Figure 5.9.

We implement the two-level OSM using the finite element software FreeFem++ [112] and we choose $\mathbb{P}_1$-bubble elements for the Stokes velocity and for the Darcy pressure and $\mathbb{P}_1$ elements for the Stokes pressure. We have 25404 degrees of freedom on the fine mesh and 6454 on the coarse one. The optimized parameter p is chosen to maximizing the smoothing property of the convergence factor (5.3.1). The parameters are $h = 0.05$, $\mu = 0.1$, $K = \mathrm{diag}(1,1)$. We first compute the exact discrete solution $(\bar{\mathbf{u}}_s, \bar{p}_s, \bar{p}_d)^\top$ by solving directly system (5.4.3) and then we count the number of iterations for the one-level OSM and the MOSM to reach a tolerance of $\mathrm{Tol} = 10^{-6}$, i.e.

$$\left(\|\mathbf{u}_s^n - \bar{\mathbf{u}}_s\|_{H_s} + \|p_s^n - \bar{p}_s\|_{Q_s} + \|p_d^n - \bar{p}_d\|_{H_d}\right) \leq \mathrm{Tol}. \tag{5.4.4}$$

For the two-level method, we used two pre-smoothing steps and no post-smoothing. The one-level OSM requires 14 iterations while the two-level OSM only 4. In Figure 5.10 we show the velocity fields in the two subdomains for the problem described by Figure 5.9.

5.5 Two-level and Multilevel substructured optimized Schwarz methods

In this section, we aim to apply two-level and multilevel substructured methods to the Stokes-Darcy system. To do so, we derive a substructured formulation for the one-level OSM method (5.1.7) since it is the starting point to formulate S2S and G2S methods. While doing so, we will generalize informally the computational framework described in Section 4 to nonoverlapping decompositions and to more general domain decomposition smoothers. A more formal derivation can be obtained immediately from our calculations for the Stokes-Darcy system, and we summarize the main steps for a two-subdomains decomposition in Remark 5.5.2.

We consider the geometry described in Section 5.1. The first step is to properly generalize the concept of substructures and substructured unknowns in the nonoverlapping case. In Section 4, the substructured iteration (4.1.11) involves the variables v_j defined over the substructures S_j, $j \in \mathscr{J}$. Recall that $S_j = \cup_{\ell \in \mathscr{N}_j} \left(\partial\Omega'_\ell \cap \Omega'_j\right)$. We emphasize that we can think of a substructured S_j as the curve where the neighbouring subdomains take the updated information about u_j at each iteration. For a nonoverlapping decomposition, we modify this definition to $S_j := \cup_{\ell \in \mathscr{N}_j} \left(\partial\Omega_\ell \cap \partial\Omega_j\right)$, hence S_j are now the portions of the boundary of Ω_j which are shared with neighbouring subdomains. For a nonoverlapping decomposition into two subdomains, this definition implies $S_1 = S_2 = \partial\Omega_1 \cap \partial\Omega_2 =: \Gamma$. We also need to identify some interface variables. For the Parallel Schwarz method, we used $v_j = \chi_j \tau_j u_j$, which represent the traces of the subdomains solutions u_j on S_j, that is the data that subdomain j is passing to the neighbouring subdomains (multiplied by a partition of unity function). Looking at the OSM iteration (5.1.7)-(5.1.8), it is then natural to use the two substructured variables λ_s, $\lambda_d \in \Lambda$, that is

$$\lambda_s = p_d + s_2\left(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}\right) \text{ and } \lambda_d = -\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} + s_1 \mathbf{u}_s \cdot \mathbf{n}.$$

Note that λ_s is exactly the data that the Darcy domain passes to the Stokes domain at each iteration, and λ_d coincides with the data passed from the Stokes to the Darcy domain. We now introduce two continuous trace operators $\tau_d : H_d \to \Lambda$ and $\tau_s : H_s \times Q_s \to \Lambda$ such that

$$\begin{aligned} \tau_d(q) &= q_{|\Gamma}, && \forall q \in H_d \\ \tau_s((\mathbf{v}, q_s)) &= (\mathbf{v} \cdot n)_{|\Gamma}, && \forall (\mathbf{v}, q_s) \in H_s \times Q_s. \end{aligned}$$

With these trace operators, we rewrite (5.1.8) as

$$\begin{aligned} \lambda_s^n &= p_d^n + s_2\left(K\nabla p_d^n \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}\right) = \left(1 + \frac{s_2}{s_1}\right) p_d^n - \frac{s_2}{s_1}\lambda_d^{n-1} = \left(1 + \frac{s_2}{s_1}\right)\tau_d(\mathscr{E}_d(\lambda_d^{n-1}, \mathbf{g}_d)) - \frac{s_2}{s_1}\lambda_d^{n-1}, \\ \lambda_d^n &= -\mathbf{n} \cdot T(\mathbf{u}_s^n, p_s^n) \cdot \mathbf{n} + s_1 \mathbf{u}_s^n \cdot \mathbf{n} = \lambda_s^{n-1} + (s_1 + s_2)\mathbf{u}_s^n \cdot \mathbf{n} = \lambda_s^{n-1} + (s_1 + s_2)\tau_s(\mathscr{E}_s(\lambda_s^{n-1}, \mathbf{f}_s)). \end{aligned} \tag{5.5.1}$$

Assuming that the iteration converges and taking the limit for $n \to \infty$, we obtain the system

$$\begin{aligned} \lambda_s &= \left(1 + \frac{s_2}{s_1}\right)\tau_d(\mathscr{E}_d(\lambda_d, \mathbf{g}_d)) - \frac{s_2}{s_1}\lambda_d, \\ \lambda_d &= \lambda_s + (s_1 + s_2)\tau_s(\mathscr{E}_s(\lambda_s, \mathbf{f}_s)). \end{aligned} \tag{5.5.2}$$

We now prove that system (5.5.2) is equivalent to the weak formulation (5.2.4).

Theorem 5.5.1 (Equivalence between volume and substructured formulations)**.** *Let the pair* $(\mathbf{u}_s, p_s, p_d) \in H_s \times Q_s \times H_d$ *solve* (5.2.4)*, then the pair* $(\lambda_s, \lambda_d) := (p_d + s_2(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}), -\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} + s_1 \mathbf{u}_s \cdot \mathbf{n})$ *satisfy* (5.5.2)*. On the other hand, if* (λ_s, λ_d) *satisfy* (5.5.2)*, then* $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s, \widetilde{p}_d) := (\mathscr{E}_s(\lambda_s, \mathbf{f}_s), \mathscr{E}_d(\lambda_d, \mathbf{g}_d))$ *satisfy* (5.2.4)*.*

Proof. Let us suppose $(\mathbf{u}_s, p_s, p_d)$ satisfies (5.2.4) and remark that, due to the derivation of the weak formulation, it holds that

$$-\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} = p_d \quad \text{and} \quad \mathbf{u}_s \cdot \mathbf{n} = -(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}) \text{ in } \Lambda. \tag{5.5.3}$$

Let us also define

$$\begin{aligned} \lambda_d &:= -\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} + s_1 \mathbf{u}_s \cdot \mathbf{n}, \\ \lambda_s &:= p_d + s_2(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}). \end{aligned}$$

It is clear that if $(\mathbf{u}_s, ps) = \mathscr{E}_s(\lambda_s, \mathbf{f}_s)$ and $p_d = \mathscr{E}_d(\lambda_d, \mathbf{g}_d)$, then (λ_s, λ_d) satisfies (5.5.2). Indeed, we would have

$$\begin{aligned} &(1 + \frac{s_2}{s_1})\tau_d(\mathscr{E}_d(\lambda_d, \mathbf{g}_d)) - \frac{s_2}{s_1}\lambda_d = (1 + \frac{s_2}{s_1})p_d - \frac{s_2}{s_1}(-\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} + s_1 \mathbf{u}_s \cdot \mathbf{n}) = \\ &= p_d + s_2(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}) = \lambda_s \end{aligned}$$

and

$$\begin{aligned} &\lambda_s + (s_1 + s_2)\tau_s(\mathscr{E}_s(\lambda_s, \mathbf{f}_s)) = p_d + s_2(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}) + (s_1 + s_2)(\mathbf{u}_s \cdot \mathbf{n}) = \\ &-\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} + s_1(\mathbf{u}_s \cdot \mathbf{n}) = \lambda_d. \end{aligned}$$

We are then left to show that $(\mathbf{u}_s, ps) = \mathscr{E}_s(\lambda_s, \mathbf{f}_s)$ and $p_d = \mathscr{E}_d(\lambda_d, \mathbf{g}_d)$. We start observing that equation (5.5.3) implies

$$\begin{aligned} \lambda_d &= -\mathbf{n} \cdot T(\mathbf{u}_s, p_s) \cdot \mathbf{n} + s_1 \mathbf{u}_s \cdot \mathbf{n} = p_d + s_1(\mathbf{u}_s \cdot \mathbf{n}), \\ \lambda_s &= p_d + s_2(K\nabla p_d \cdot \mathbf{n} - \mathbf{g}_d \cdot \mathbf{n}) = p_d - s_2(\mathbf{u}_s \cdot \mathbf{n}). \end{aligned} \tag{5.5.4}$$

Testing (5.2.4) against the test function $(\mathbf{0}, 0, q_d)$, we get p_d satisfies

$$\widetilde{a}_d(p_d, q_d) - \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d = \int_\Gamma (\mathbf{u}_s \cdot \mathbf{n}) q_d, \quad \forall q_d \in H_d. \tag{5.5.5}$$

On the other hand, $\widetilde{p}_d = \mathscr{E}_d(\lambda_d, \mathbf{g}_d)$ satisfy

$$a_d(\widetilde{p}_d, q_d) = \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d + \frac{1}{s_1} \int_\Gamma \lambda_d q_d, \quad \forall q_d \in H_d. \tag{5.5.6}$$

Replacing the expression of λ_d from (5.5.4), and using (5.5.5), we get

$$a_d(\widetilde{p}_d, q_d) = a_d(p_d, q_d), \quad \forall q_d \in H_d, \tag{5.5.7}$$

which implies $\widetilde{p}_d = \mathscr{E}_d(\lambda_d, \mathbf{g}_d) = p_d$. Similarly, testing (5.2.3) against test functions $(\mathbf{v}, q_s, 0)$ we obtain that $(\mathbf{u}_s, p_s)$ satisfy,

$$\begin{aligned} &\widetilde{a}_s(\mathbf{u}_s, \mathbf{v}) + b_s(\mathbf{v}, p_s) + C(p_d, \mathbf{v}) = \int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{v}, \\ &b_s(\mathbf{u}_s, q_s) = 0. \end{aligned} \tag{5.5.8}$$

On the other hand, using the expression of λ_s in (5.5.4), $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s) = \mathscr{E}_s(\lambda_s, \mathbf{f}_s)$ satisfies

$$\begin{aligned} &a_s(\widetilde{\mathbf{u}}_s, \mathbf{v}) + b_s(\mathbf{v}, \widetilde{p}_s) = \int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{v} - \int_\Gamma p_d(\mathbf{v}\cdot\mathbf{n}) + \int_\Gamma s_2(\mathbf{u}_s\cdot\mathbf{n})(\mathbf{v}\cdot\mathbf{n}), \qquad \forall \mathbf{v} \in H_s, \\ &\int_{\Omega_s} \nabla\cdot\widetilde{\mathbf{u}}_s \cdot q_s = 0, \qquad \forall q_s \in Q_s. \end{aligned} \tag{5.5.9}$$

Inserting (5.5.8) we obtain that the couple $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s) = \mathscr{E}_s(\lambda_s, \mathbf{f}_s)$ satisfies

$$\begin{aligned} &a_s(\widetilde{\mathbf{u}}_s, \mathbf{v}) + b_s(\mathbf{v}, \widetilde{p}_s) = a_s(\mathbf{u}_s, \mathbf{v}) + b_s(\mathbf{v}, p_s), \qquad \forall \mathbf{v} \in H_s \\ &\int_{\Omega_s} \nabla\cdot\widetilde{\mathbf{u}}_s \cdot q_s = 0, \qquad \forall q_s \in Q_s. \end{aligned} \tag{5.5.10}$$

and we deduce that $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s) = \mathscr{E}_s(\lambda_s, \mathbf{f}_s) = (\mathbf{u}_s, p_s)$. This concludes the first part of the proof.

We now suppose that (λ_s, λ_d) are solution of (5.5.2) and we define

$$(\widetilde{\mathbf{u}}_s, \widetilde{p}_s, \widetilde{p}_d) := (\mathscr{E}_s(\lambda_s, \mathbf{f}_s), \mathscr{E}_d(\lambda_d, \mathbf{g}_d)).$$

From (5.5.2) we obtain

$$\lambda_s = \widetilde{p}_d - s_2(\widetilde{\mathbf{u}}\cdot\mathbf{n}), \text{ and } \lambda_d = \widetilde{p}_d + s_1(\widetilde{\mathbf{u}}\cdot\mathbf{n}). \tag{5.5.11}$$

Adding problems (5.2.11) and (5.2.10) we obtain for all $(\mathbf{v}, q_s, q_d) \in H_s \times Q_s \times H_d$

$$\begin{aligned} &\widetilde{a}_s(\widetilde{\mathbf{u}}_s, \mathbf{v}) + b_s(\mathbf{v}, \widetilde{p}_s) + \int_\Gamma s_2(\widetilde{\mathbf{u}}_s\cdot\mathbf{n})(\mathbf{v}\cdot\mathbf{n}) + \widetilde{a}_d(\widetilde{p}_d, q_d) + \frac{1}{s_1}\int_\Gamma p_d q_d \\ &= \int_{\Omega_s} \mathbf{f}\cdot\mathbf{v} - \int_\Gamma \lambda_s(\mathbf{v}\cdot\mathbf{n}) + \int_{\Omega_d} \mathbf{g}_d \cdot \nabla q_d + \frac{1}{s_1}\int_\Gamma \lambda_d q_d, \quad \forall (\mathbf{v}, q_s, q_d) \in H_s \times Q_s \times H_d, \\ &b_s(\widetilde{\mathbf{u}}_s, q_s) = 0. \end{aligned} \tag{5.5.12}$$

Inserting (5.5.11) into (5.5.12), we obtain that $(\widetilde{\mathbf{u}}_s, \widetilde{p}_s, \widetilde{p}_d)$ are solutions of (5.2.4), that is solutions of the original coupled problem which concludes the proof. □

We now define $G_s : \Lambda \to \Lambda$ as $G_s(\lambda) := \tau_s \mathscr{E}_s(\lambda, 0)$ and $G_d : \Lambda \to \Lambda$ as $G_d(\lambda) := \tau_d \mathscr{E}_d(\lambda, 0)$. Furthermore we set $\chi_s := \tau_s(\mathscr{E}_s(0, \mathbf{f}_s))$ and $\chi_d := \tau_d(\mathscr{E}_d(0, \mathbf{g}_d))$. Using the linearity of the operators $\mathscr{E}_s$ and $\mathscr{E}_d$, (5.5.2) can be reformulated as[4]

$$\begin{pmatrix} I & \frac{s_2}{s_1} I - (1 + \frac{s_2}{s_1}) G_d(\cdot) \\ -I - (s_1 + s_2) G_s(\cdot) & I \end{pmatrix} \begin{pmatrix} \lambda_s \\ \lambda_d \end{pmatrix} = \begin{pmatrix} \chi_s \\ \chi_d \end{pmatrix}, \tag{5.5.13}$$

[4]Working in a pure algebraic setting, one could get a slightly modified version of (5.5.13), see (4.6) in [53], which differs in the sign of the term $(s_1 + s_2)G_s(\cdot)$ and in a parameter s_1 multiplying G_d(G_s and G_d are called S_f and S_d in [53]). These terms are already included in the definition of the infinite dimensional operators G_s and G_d.

which is exactly the counterpart of (4.1.13) for the Stokes-Darcy system. The substructured version of (5.1.7) is then

$$\begin{pmatrix}\lambda_s^n\\ \lambda_d^n\end{pmatrix}=\begin{pmatrix}0 & \frac{s_2}{s_1}I-(1+\frac{s_2}{s_1})G_d(\cdot)\\ -I-(s_1+s_2)G_s(\cdot) & 0\end{pmatrix}\begin{pmatrix}\lambda_s^{n-1}\\ \lambda_d^{n-1}\end{pmatrix}+\begin{pmatrix}\chi_s\\ \chi_d\end{pmatrix}. \tag{5.5.14}$$

Defining the errors $e_s^n=\lambda_s-\lambda_s^n$ and $e_d^n=\lambda_d-\lambda_d^n$, the substructured OSM (5.5.14) reads in the error form

$$\begin{pmatrix}e_s^n\\ e_d^n\end{pmatrix}=\begin{pmatrix}0 & \frac{s_2}{s_1}I-(1+\frac{s_2}{s_1})G_d(\cdot)\\ -I-(s_1+s_2)G_s(\cdot) & 0\end{pmatrix}\begin{pmatrix}e_s^{n-1}\\ e_d^{n-1}\end{pmatrix}. \tag{5.5.15}$$

Remark 5.5.2 (Substructured OSM for the Laplace equation). Considering the OSM for the Laplace equation (1.3.14), we define two interface variables as $\lambda_1^n:=\frac{\partial u_1^n}{\partial \mathbf{n}_2}+s_2u_1^n$ and $\lambda_2^n:=\frac{\partial u_2^n}{\partial \mathbf{n}_1}+s_1u_2^n$. It is possible to show, [61, Chapter 2], that the OSM iterates satisfy

$$\lambda_1^n=-\lambda_2^{n-1}+2s_1u_2^n \text{ and } \lambda_2^n=-\lambda_1^{n-1}+2s_2u_1^n.$$

Expressing now $u_j^n:=G_j(\lambda_j^{n-1},f)$, where G_j are extensions operators of a Robin trace on Γ with some right hand side f, we obtain the substructured iteration

$$\begin{pmatrix}\lambda_1^n\\ \lambda_2^n\end{pmatrix}=\begin{pmatrix}0 & -I+G_2(\cdot,0)\\ -I+G_1(\cdot,0) & 0\end{pmatrix}\begin{pmatrix}\lambda_1^{n-1}\\ \lambda_2^{n-1}\end{pmatrix}+\begin{pmatrix}G_2(0,f)\\ G_1(0,f)\end{pmatrix}.$$

5.5.1 Numerical experiments

We present numerical experiments to study the convergence properties of the S2S and G2S methods applied to the Stokes-Darcy system. We consider the geometry of Fig 5.9 with a mesh of regular elements whose mesh size is h. We present tables with the number of iterations to reach a tolerance Tol = 10^{-8}. We consider the one-level OSM (G), the S2S method with a coarse space made of N_c eigenfunctions of G (S2S-G(N_c)), the S2S method with a coarse space made of N_c random functions obtained through PCA (S2S-PCA(N_c)), and the G2S method. Concerning the two-level methods, we emphasize that the one-level OSM is not a natural smoother, see also the discussion in Chapter 3, and thus we need to choose properly the optimized parameters to have good smoothing properties. A wrong estimation of these parameters could destroy the excellent convergence properties of the two-level methods. Hence, we aim to study both how fast the two-level methods are if the exact optimized parameters are available, and how the convergence deteriorates if one uses either Fourier or probing techniques to estimate them.

5.5.1.1 Robustness with respect to the mesh size

We first study the robustness of the methods with respect to the mesh size h. We consider two different settings. On the left of Table 5.1, we consider homogeneous Dirichlet boundary conditions except on the top horizontal edge of Ω_s, where a zero normal stress condition is imposed. We have observed that the Fourier analysis is precise in this settings,

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
G	12(12)	13(13)	14(14)	15(14)
S2S-G(5)	11(7)	13(8)	14(8)	15(8)
S2S-PCA(5)	9(6)	14(8)	16(10)	16(10)
S2S-G(10)	7(6)	13(6)	14(8)	14(8)
S2S-PCA(10)	6(5)	12(6)	16(8)	16(8)
G2S	6(6)	6(6)	5(5)	4(4)

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
G	13(12)	12(12)	12(12)	12(12)
S2S-G(5)	10(7)	10(8)	10(9)	9(9)
S2S-PCA(5)	7(7)	7(9)	8(10)	9(10)
S2S-G(10)	6(5)	6(6)	8(6)	8(7)
S2S-PCA(10)	5(4)	6(6)	8(7)	8(8)
G2S	6(6)	6(6)	5(5)	4(4)

Table 5.1: Number of iterations to reach a tolerance of Tol = 10^{-8} for the one-level OSM (G), the S2S method with coarse space made of N_c eigenfunctions of G (S2S-G(N_c)), the S2S method with N_c random functions obtained through PCA (S2S-PCA(N_c)), and the G2S method. The physical parameters are $\mu = 0.1$, $\eta_1 = \eta_2 = 1$.

see Section 5.3.1. On the right of Table 5.1, we impose homogeneous boundary conditions all along $\partial\Omega$, and we use probing to estimate the parameters. For each method we present a couple of numbers. The first number indicates the iterations required to reach the tolerance with a parameter obtained through Fourier analysis (left table) and probing technique (right table). When using the Fourier approach, we solve the optimization problem (5.3.1) with a range of frequencies $k \in [\pi, N\pi]$, where N is the number of unknowns on Γ. The resulting parameters are used for G and the S2S methods. For the G2S method we use frequencies in the range $k \in [\frac{N}{2}\pi, N\pi]$ to optimized the smoothing property. Concerning probing, we use the set $\mathcal{K}_2$ defined in Section 5.3.1 for the G and S2S methods, while $\mathcal{K}_3 := \{\frac{N}{2}, N\}$ for the G2S method. The second number in bracket is obtained using the parameter $p = \text{argmin}\rho(X(p))$, where X is either the one-level OSM (G), the S2S-G method and the G2S method. These exact optimized parameters are found through a brute force optimization. For the S2S-PCA, we use the same parameters of the S2S-G method.

We remark that the one-level OSM has mesh independent convergence, as shown in [56]. We stress that in the spectral case it is absolutely not trivial how to choose the parameters. Setting $p = \text{argmin}\rho(G(p))$ is not necessarily the best choice, as it guaranties that the spectral radius is minimized, but we do not have any control on the remaining part of the spectrum. As Table 5.1 shows, it could be better to choose a different p which has few large eigenvalues, whose corresponding eigenvectors are inserted into the coarse space, and the remaining spectrum contains only very small eigenvalues. In our numerical experiments, it happens that $p = \text{argmin}\rho(G(p))$ leads to a very large plateau of eigenvalues, even though $\rho(G(p))$ is minimized, and thus increasing the dimension of the coarse space does not lead to a significant improvement of the convergence. The S2S methods and G2S are faster than the one-level OSM, even tough the improvement is not so significant as for the second order elliptic equation discussed in Chapter 4. Nevertheless, we note that both methods are roughly twice faster than the one-level OSM, and the G2S method becomes even faster as the mesh size decreases since the coarse problem size increases. Since each iteration of the one-level OSM is quite expensive, as it requires to solve a Stokes problem, it is really promising to see this reduction in terms of iteration numbers with only

μ	10^{-2}	10^{-3}	10^{-4}
G(p)	29(29)	64(61)	90(30)
G(p,q)	17(15)	25(22)	80(16)
S2S-G(5)	19(13)	36(19)	14(13)
S2S-PCA(5)	12(12)	34(21)	10(14)
S2S-G(10)	9(8)	27(11)	11(10)
S2S-PCA(10)	10(8)	26(12)	10(10)
G2S	7(7)	9(9)	11(11)

Table 5.2: Number of iterations to reach a tolerance of Tol = 10^{-8} for the one-level OSM (G), the S2S method with coarse space made of N_c eigenfunctions of G (S2S-G(N_c)), the S2S method with N_c random functions obtained through PCA (S2S-PCA(N_c)) and the G2S method. The mesh size is $h = \frac{1}{8}$ and $\eta_1 = \eta_2 = 1$.

the additional cost of a small coarse problem defined over the interface. For instance, considering a mesh size $h = \frac{1}{128}$, the global problem has 214788 unknowns (148739 for the Stokes equation and 66049 for the Darcy equation), while the dimension of the coarse problem for the G2S method is 254.

5.5.1.2 Robustness with respect to physical parameters

In this subsection we study the robustness of the two-level methods with respect to the physical parameters. The first Robin-Robin method for the Stokes-Darcy coupling was indeed introduced in order to overcome the deterioration of the convergence of the Dirichlet-Neumann method as the physical parameters become smaller. In addition to the methods already considered, we add a one-level double sided OSM denoted with $G(p,q)$. Table 5.2 reports the number of iteration to reach a tolerance of Tol $< 10^{-8}$ for the different methods as μ become smaller. We remark that the one-level methods are not very robust, and the estimation of the optimized parameters is not trivial as it fails for both the one-level methods when $\mu = 10^{-4}$. Finally, the addition of a coarse space makes the iterative methods less sensitive to the choice of the optimized parameters.

CHAPTER 6

Substructured Nonlinear Preconditioning

"The subject of preconditioning (...) is already decades old. Yet the design and study of preconditioners seems constantly fresh as it continues to address problems from new application domains, adapts to new computer architectures, and incorporates ideas from new fields"

— E. Chow, K. Vulk, *Preconditioning in the new decade*, SIAM News, Issue 2, 2020.

Among the new ideas which will give life to the preconditioning field in the next decade, the authors of the epigraph included nonlinear preconditioning. At a first glance, stating that we aim to precondition a nonlinear system seems a nonsense. As a matter of fact, preconditioning is traditionally associated with linear systems as the same word preconditioning automatically induces the reader to think about techniques to better "condition" a matrix operator A. Suppose we aim to solve a linear system $Au = b$. Then to precondition the linear system on the left or on the right means to replace the original linear system with

$$M^{-1}Au = M^{-1}b, \quad \text{left preconditioning},$$
$$AP^{-1}y = b, \quad Pu = y, \quad \text{right preconditioning},$$

where M^{-1} and P^{-1} are called respectively left and right preconditioners.
Let us now consider a nonlinear system $F(u) = 0$. By preconditioning a nonlinear system, we mean that we aim to replace the original nonlinear system with a new nonlinear system, still having the same solution, but for which the nonlinearities are more balanced and Newton's method converges faster [18, 77]. Thus, to precondition a nonlinear system on the left or on right means to replace the original nonlinear system with

$$G(F(u)) = 0, \quad \text{left preconditioning},$$
$$F(H(y)), \quad u = H^{-1}(y), \quad \text{right preconditioning},$$

where G and H are nonlinear functions called respectively left and right preconditioners. Seminal contributions in nonlinear preconditioning have been made by Cai and Keyes

in [18, 19], where they introduced the ASPIN method (Additive Schwarz Preconditioned Inexact Newton), which is a left preconditioner. The development of efficient preconditioners is not an easy task even in the linear case. One useful strategy is to study efficient iterative methods, and then to use the associated preconditioners (see equation (1.1.7)) in combination with Krylov methods [77]. The same logical path paved the way to the development of the RASPEN method (Restricted Additive Schwarz Preconditioned Exact Newton) in [59] which will be explained in the next sections. For the sake of completeness, we remark that domain decomposition methods can also be applied as either nonlinear iterative methods, that is by just solving nonlinear problems in each subdomain and then exchanging information between subdomains as in the linear case [128, 129, 17, 117], or as preconditioners to solve the Jacobian linear system inside a Newton's iteration. In the latter case, the term Newton-Krylov-DD is employed, where DD is replaced by the domain decomposition preconditioner used [17, 117].

This chapter is based on ongoing work with Faycal Chaouqui, Martin Gander and Pratik Kumbhar. Starting from the RASPEN method introduced in [59], we define a similar method at the substructured level and we called it "SRASPEN" (Substructured Restricted Additive Schwarz Preconditioned Exact Newton) method. Considering one-level variants, we prove that substructuring does not modify the convergence behaviour of Newton's method, that is, RASPEN and SRASPEN methods produce the same iterates on the substructures. However, we will discuss the advantages of applying substructuring from a computational point of view. Considering instead two-level variants, we will show that the SRASPEN method exhibits faster convergence. Since here we will discuss preliminary results, we refer the reader to [29] for further details.

We emphasize that substructuring has received very little attention in the nonlinear case compared to the linear case. Actually, the term nonlinear elimination is usually employed instead of substructuring, and this concept is tightly linked with the development of right preconditioners. In the next section we provide a brief review of two methods where substructuring ideas are used in the nonlinear case.

6.1 Nonlinear Elimination

We consider the nonlinear system

$$F(u) = 0, \tag{6.1.1}$$

where $F = (f_1, f_2, \ldots, f_n)^\top$ and $u = (u_1, u_2, \ldots, u_n)^\top$. A standard technique to solve (6.1.1) is Newton's method which, starting from an initial guess u^0, generates a sequence of iterates $\{u^k\}$, $k = 1, 2, \ldots$, defined as

$$u^k = u^{k-1} + r_k, \quad r_k = -(J_F(u^{k-1}))^{-1} F(u^{k-1}), \tag{6.1.2}$$

where $J_F(u^{k-1})$ is the Jacobian of F evaluated at u^{k-1}. It is well known that Newton's method converges quadratically sufficiently close to the exact solution u. However, the choice of the initial guess u^0 has a tremendous influence and global convergence is not

guaranteed. Another popular method to solve (6.1.1) is the inexact Newton method which differs from Newton's method in two aspects. First, it allows one to solve the Jacobian system approximately, which can be of great interest for large problems. Second, it introduces a relaxation parameter $t_k \in (0,1]$ multiplying r_k. The goal of t_k is to enlarge the convergence basin of Newton's method, as the residual is certainly decreasing along the direction r_k in a neighbourhood of u^k, but sometimes Newton's method could take too large steps in this direction, ending up increasing the residual. The inexact Newton method is based on the recurrence relation

$$u^k = u^{k-1} + t_k r_k, \text{ where } r_k \text{ sastifies } \alpha_k = \frac{||J_F(u^{k-1})r_k + F(u^{k-1})||}{||F(u^{k-1})||} < 1, \tag{6.1.3}$$

and $t_k \in (0,1]$ is chosen such that $||F(u^k)|| < \delta||F(u^{k-1})||$, $\delta \in (0,1)$. We emphasise that α_k is a measure of how well we solve approximately the Jacobian system and usually $\alpha_k < 1$ is sufficient for convergence. We refer the interested reader to the monograph [51] for a comprehensive theory of convergence of Newton's method.

To the best of the author's knowledge, the first nonlinear elimination algorithm has been proposed in [120] based on the following idea. Let us suppose that F can be partitioned into

$$F(x_1,x_2) = (F_1(x_1,x_2), F_2(x_1,x_2)), \text{ with } J_F(x_1,x_2) = \begin{pmatrix} J_{F_{11}} & J_{F_{12}} \\ J_{F_{21}} & J_{F_{22}} \end{pmatrix}, \tag{6.1.4}$$

where $J_{F_{ij}} := \frac{\partial F_i}{\partial x_j}$. The functions F_j have as many equations as the degrees of freedom of x_j, $j = 1,2$. We assume that $F_1(x_1,x_2)$, regarded as a function of x_1 given x_2, has some components which are highly nonlinear, with large residual, and that lead to very small values of t_k. For smooth functions F_1, we can use the Implicit Function Theorem to find formally a function $h(x_2)$ such that $F_1(h(x_2),x_2) = 0$, that is, for every choice of x_2 there exists an x_1 function of x_2, such that $F_1(x_1(x_2),x_2) = 0$. Assuming that $h(\cdot)$ is known, then one solves the smaller nonlinear system

$$g(x_2) := F_2(h(x_2),x_2) = 0. \tag{6.1.5}$$

From the computational point of view, it is straightforward to apply Newton's method to (6.1.5). Computing the Jacobian of g and using implicit differentiation for h', one finds

$$g'(x_2) = (F_2(h(x_2),x_2))' = J_{F_{22}} + J_{F_{21}}h'(x_2) = J_{F_{22}} - J_{F_{21}}J_{F_{11}}^{-1}J_{F_{12}},$$

that is, the Schur complement of J_F. Thus, solving $g'(x_2^{k-1})r_k = -g(x_2^{k-1})$ is equivalent to solve

$$\begin{pmatrix} J_{F_{11}} & J_{F_{12}} \\ J_{F_{21}} & J_{F_{22}} \end{pmatrix}\begin{pmatrix} \Delta x_1 \\ r_k \end{pmatrix} = \begin{pmatrix} 0 \\ -F_2(h(x_2),x_2) \end{pmatrix}, \tag{6.1.6}$$

i.e., the original global Jacobian problem evaluated at $(h(x_2),x_2)$. In Algorithm 7 we summarize this method proposed in [120]. We stress that Step 5 of Algorithm 7 requires to evaluate $h(\cdot)$ in $x_2^k + t_k r_k$ for every t_k. Since generally we do not have a closed form expression for $h(\cdot)$, we need to solve $F_1(\Delta x_1, x_2 + t_k r_k) = 0$ and set $\Delta x_1 = h(x_2 + t_k r_k)$ for every

Algorithm 7: Nonlinear Elimination Algorithm [120]

Require: x^0 and F.

1: Split the nonlinear system into $F = (F_1, F_2)$ and the unknowns into $x^0 = (x_1^0, x_2^0)$.
2: Solve $F_1(x_1^0, x_2^0) = 0$ and set $x_1 = h(x_2^0)$, $k = 0$.
3: Repeat from 4 to 6 until convergence
4: Solve equation (6.1.6) with the current approximation $(h(x_2^k), x_2^k)$.
5: Find t_k such that $||F_2(h(x_2^k + t_k r_k), x_2^k + t_k r_k)|| \leq \delta ||F(h(x_2^k), x_2^k)||$.
6: Set $x_2^{k+1} = x_2^k + t_k r_k$, $h(x_2^{k+1}) = h(x_2^k + t_k r_k)$ and $k = k + 1$.

value of t_k. This implies a potentially large number of solves for F_1. Indeed, we can think of Algorithm 7 as an efficient way to spread the computational costs. If F_1 slows down the convergence of Newton's method, instead of performing several solves on the global and large nonlinear system, Algorithm 7 performs very few iterations of the inexact Newton method applied to F_2, while focusing on solving the harder nonlinear problem F_1. Algorithm 7 can also be interpreted as a Newton's method applied to the right preconditioned system $F(H(x)) = 0$. To see this, let us ignore Step 5 so that we are dealing with an exact Newton's method and define $H(x_1, x_2) := (h(x_2), x_2)$. Then starting from $x^0 = \left(x_1^0, x_2^0\right)$, Step 2 computes $y^0 = H(x^0)$, while Step 3 applies Newton's method to $F(H(x^0))$ to get an approximation x^1. Finally Step 6 computes $y^1 = H(x^1)$ for the next Newton iteration.

The decomposition (6.1.4) assumed by the nonlinear elimination algorithm proposed in [120] is very general and does not necessarily share any link with a domain decomposition method. Cai and collaborators have recently introduced domain decomposition variants of this algorithm and applied them to several problems such as CFD [20] and hyperelasticity with application to arteries' deformation [103, 104]. The motivations of the work proposed in [20] lie in the observations that certain local strong nonlinearities can be handled locally and for instance, local subdomain solves allow one to have a zero residual inside each subdomain, but still quite large residuals are present along the interfaces between subdomains. Thus, whenever the residual of a global Newton's iteration is too large in some part of the domain, they propose to add a RAS step inside a Newton-Krylov-RAS iteration. As the authors explicitely stated in [20], "its (The RAS step) purpose is to provide a better initial guess for the next outer Newton iteration". To define mathematically the method we follow the notation of [20].
Let us introduce the set of indices $S := \{1, 2, ..., n\}$, and decompose it into several subset S_i, $i = 1, ..., N$, such that $\bigcup_{i=1}^{N} S_i = S$ and $S_i \cap S_j = \emptyset$ if $i \neq j$. In addition to S_i, we consider the set S_i^δ such that $S_i \subset S_i^\delta$ for every i. The sets S_i and S_i^δ can be thought as an algebraic nonoverlapping and overlapping decomposition of S [1]. We also need to define the spaces $V_i := \left\{x \in \mathbb{R}^n : x_j = 0, \text{if } j \notin S_i\right\}$, $V_i^\delta := \left\{x \in \mathbb{R}^n : x_j = 0, \text{if } j \notin S_i^\delta\right\}$ and the matrices $R_i^0 : \mathbb{R}^n \to V_i$ and $R_i^\delta : \mathbb{R}^n \to V_i^\delta$. Moreover, we define the restriction of F onto V_i^δ as

[1] The sets of indices S_i can be defined dynamically at each outer Newton iteration as the sets of indices of equations f_i for which the residual is too large according to some criteria. In this case, the S_i do not share any link with a geometric decomposition of the domain, see [104].

Algorithm 8: Newton-Krylov-Schwarz-RAS [20]

Require: x^k and F.

1: If global condition is satisfied, stop.
2: If local conditions are not satisfied go to Step 3. Otherwise go to Step 4 and set $\widetilde{x}^k = x^k$.
3: Compute one step of the RAS method, $\widetilde{x}^k = G(x^k)$.
4: Compute the next approximation x^{k+1}, performing one step of the Newton method on $F(x) = 0$ with the approximate solution $\widetilde{x}^k$ and use a Krylov method preconditioned by a Schwarz method to solve the Jacobian system.
5: Repeat from 1 to 4 until convergence.

$F_i^\delta := R_i^\delta F(x)$, $i = 1, ..., N$. Then, as we will discuss in detail in Section (6.2), one step of the nonlinear RAS method can be written as

$$w = G(x) = x + \sum_{i=1}^{N} R_0^j x_i^\delta, \text{ where } x_i^\delta \text{ are solution of } F_i^\delta(x + x_i^\delta) = 0.$$

With these ingredients, the algorithm proposed in [20] is described by Algorithm 8. The aim of the local nonlinearity conditions is to check that the residual is balanced among the sets S_i. Thus for every $i = 1, ..., N$, we claim that the nonlinearity is not balanced if there exits an $m \in \{1, ..., N\}$ such that $||F_m^\delta(x^k)|| > \rho||F(x^k)||$, where $\rho \in (0, 1)$ is a parameter to be tuned. For further details we refer the interested reader to [104]. Algorithm 8 can be seen as a special case of Algorithm 7, where the function H is now the Schwarz operator G, and the unknowns in the interior of each subdomain are treated implicitly, as the genereric variable x_1 was eliminated in Algorithm 7. Indeed, the Schwarz operator G guarantees that, given some particular values on the interface between subdomains, the residual is zero in the interior of the subdomains, that is, the set of the interior nonlinear equations is satisfied.

In other words, Algorithm 8 uses the RAS method in an outer Newton iteration to eliminate the interior degrees of freedom, leaving to Newton's method the duty to take care of the remaining unknowns. In the next section we are going to adopt a different point of view. We will derive a substructured formulation of the Schwarz method for nonlinear problems, that is the counterpart of (4.1.11). Once we have the iterative method, we will use Newton's method to solve the corresponding fixed-point equation. The method we propose is therefore naturally defined over the substructured unknowns, and it does not require to explicitly eliminate the interior unknowns at each outer Newton iteration.

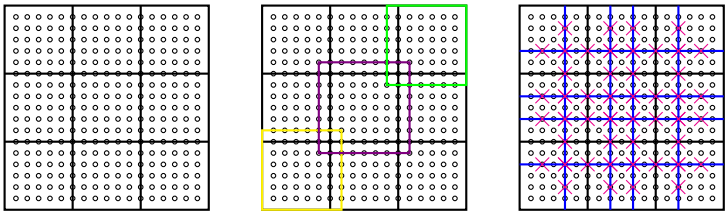


Figure 6.1: The domain Ω is divided into nine nonoverlapping subdomains (left). The center panel shows how the diagonal nonoverlapping subdomains are enlarged to form overlapping subdomains. On the right, we denote the unknowns represented in V^s (blue line) and the unknowns of a coarse space of V^s (red crosses).

6.2 Definition of the SRASPEN method

Let us consider the boundary value problem posed in a Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$,

$$\begin{aligned} \mathscr{L}(u) &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \partial\Omega. \end{aligned} \tag{6.2.1}$$

We assume that (6.2.1) admits a unique solution in some Hilbert space V. If the boundary value problem is linear, a discretization of (6.2.1) leads to the linear system

$$Au = f, \tag{6.2.2}$$

while if the boundary value problem is nonlinear, we solve the nonlinear system

$$F(u) = 0. \tag{6.2.3}$$

In this section we introduce a new one-level nonlinear solver based on domain decomposition for the nonlinear system (6.2.3) called the SRASPEN (Substructured Restricted Additive Schwarz Preconditining Exact Newton) method. To introduce this method and to better specify the relations between SRASPEN and other existing linear and nonlinear solvers, we take a brief excursus on domain decomposition methods to solve (6.2.2) and (6.2.3).
Let us decompose the domain Ω into N overlapping subdomains Ω'_j, that is $\Omega = \bigcup_{j\in\mathscr{J}} \Omega'_j$ with $\mathscr{J} := \{1,2,\dots,N\}$ (see Fig. 6.1). For each subdomain Ω'_j, we define V_j as the restriction of V onto Ω'_j. Further, we introduce the classical restriction and prolongation operators $R_j : V \to V_j$, $P_j : V_j \to V$, and the restricted prolongation operators $\widetilde{P}_j : V_j \to V$. We assume that these operators satisfy

$$R_j P_j = I_{V_i}, \quad \sum_{j\in\mathscr{J}} \widetilde{P}_j R_j = I, \tag{6.2.4}$$

where I_{V_i} is the identity on V_i and I is the identity on V. A classical domain decomposition method to solve a linear equation (6.2.2) is the RAS method, which starting from an

approximation u^0 computes for $n = 1, 2, \ldots$

$$u^n = u^{n-1} + \sum_{j \in \mathscr{J}} \widetilde{P}_j A_j^{-1} R_j (f - A u^{n-1}), \tag{6.2.5}$$

where A_j are defined as $A_j := R_j A P_j$. Let us now rewrite the iteration (6.2.5) in an equivalent form using the hypothesis in (6.2.4) and the definition of A_j,

$$\begin{aligned} u^n &= \sum_{j \in \mathscr{J}} \widetilde{P}_j R_j u^{n-1} + \sum_{j \in \mathscr{J}} \widetilde{P}_j A_j^{-1} R_j \left(f - A u^{n-1}\right) = \sum_{j \in \mathscr{J}} \widetilde{P}_j A_j^{-1} \left(A_j R_j u^{n-1} + R_j \left(f - A u^{n-1}\right)\right) \\ &= \sum_{j \in \mathscr{J}} \widetilde{P}_j A_j^{-1} \left(R_j A P_j R_j u^{n-1} + R_j \left(f - A u^{n-1}\right)\right) = \sum_{j \in \mathscr{J}} \widetilde{P}_j A_j^{-1} R_j \left(f - A\left(I - P_j R_j\right) u^{n-1}\right) \\ &=: G^{\mathrm{RAS}}(u^{n-1}). \end{aligned} \tag{6.2.6}$$

Similarly, the RAS method can be used to solve the nonlinear equation (6.2.3). To show this, we introduce the solution operators G_j which are defined through

$$R_j F(P_j G_j(u) + (I - P_j R_j) u) = 0. \tag{6.2.7}$$

The nonlinear RAS method then reads

$$u^n = \sum_{j \in \mathscr{J}} \widetilde{P}_j G_j(u^{n-1}). \tag{6.2.8}$$

It is possible to show that (6.2.8) reduces to (6.2.6) if $F(u)$ is a linear function. Infact assuming $F(u) = Au - f$, equation (6.2.7) becomes,

$$\begin{aligned} R_j F\left(P_j G_j\left(u^{n-1}\right) + \left(I - P_j R_j\right) u^{n-1}\right) &= R_j \left(A\left(P_j G_j\left(u^{n-1}\right) + \left(I - P_j R_j\right) u^{n-1}\right) - f\right) \\ &= A_j G_j\left(u^{n-1}\right) + R_j \left(A\left(I - P_j R_j\right) u^{n-1} - f\right) = 0, \end{aligned} \tag{6.2.9}$$

which implies $G_j\left(u^{n-1}\right) = A_j^{-1} R_j \left(f - A\left(I - P_j R_j\right) u^{n-1}\right)$, and thus (6.2.8) reduces to (6.2.6). We remark that $\left(P_j R_j - I\right) u^{n-1}$ contains non-zero elements only outside subdomain Ω_j, and in particular $A\left(P_j R_j - I\right) u^{n-1}$ represents precisely the boundary condition for Ω_j given the old approximation u^{n-1}. This observation suggests that the RAS method, like most domain decomposition methods, can be written in a substructured formulation. Infact, despite the iteration (6.2.6) is written in volume form, that is, it involves the whole vector u^{n-1}, only very few elements of u^{n-1} are needed to compute the new approximation u^n. For further details about the classical parallel Schwarz method in a substructured formulation we refer to [80] for the two subdomain case, and [41] for a general decomposition into several subdomains with cross points at the continuous level.

We now define a substructured formulation for the RAS method both for the linear and the nonlinear case. In the following we use the notation introduced in [40]. For any $j \in \mathscr{J}$, we define the set of neighbouring indices $\mathscr{N}_j := \{\ell \in \mathscr{J} : \Omega_j \cap \partial\Omega_\ell \neq \emptyset\}$. Given a $j \in \mathscr{J}$, we introduce the substructure of Ω_j defined as $S_j := \bigcup_{\ell \in \mathscr{N}_j} \left(\partial\Omega_\ell \cap \Omega_j\right)$, that is the union of all the portions of $\partial\Omega_\ell$ with $\ell \in \mathscr{N}_j$. The substructure of the whole domain Ω is defined

as $S := \bigcup_{j\in\mathcal{J}} \overline{S_j}$. We now introduce the space V^S which can be interpreted as the space V restricted onto the substructure S. We can define it in two ways. Either $V^S := V_{|S}$, or $V^S := \otimes_{j\in S_j} V_{|S_j}$. In the following sections we have used the first definition since we prefer to have full rank operators R_S and P_S. Remark that in Chapter 4, we used the second definition, $V^S := \otimes_{j\in S_j} V_{|S_j}$, which doubles the unknowns in the overlap between interfaces and leads to operators R_S and P_S which are not full rank, and may cause problems while solving the Jacobian system inside a Newton's iteration. Associated to V^S, we consider the restriction operator $R_S : V \to V^S$ and a prolongation operator $P_S : V^S \to V$. The restriction operator R_S takes an element $v \in V$ and restricts it onto the skeleton S. The prolongation operator P_S extends an element $v \in V^S$ to the global space V. How this extension is done is not crucial as we will use P_S inside a domain decomposition algorithm, and thus only the values on the skeleton S will play a role. Hence, we only require that $R_S P_S = I_S$, where I_S is the identity operator on V^S. In the following, P_S extends an element $v_S \in V^S$ to zero in $\Omega \setminus S$, but the same analysis can be adapted to any other choice of P_S.

Given a substructured approximation $v^0 \in V^S$, for $n = 1, 2, \ldots$ we define the Substructured RAS (SRAS) method as

$$v^n = G^{\mathrm{SRAS}}(v^{n-1}), \tag{6.2.10}$$

where $G^{\mathrm{SRAS}}(v) := R_S G^{\mathrm{RAS}}(P_S v)$. The RAS method and SRAS method are obviously tightly linked, but when are they equivalent? We must impose some conditions on P_S and R_S. It is sufficient to assume that the restriction and prolongation operators satisfy

$$R_S G^{\mathrm{RAS}}(u) = R_S G^{\mathrm{RAS}}(P_S R_S u), \quad \forall u \in V. \tag{6.2.11}$$

Heuristically, we need that the operator $P_S R_S$ preserves all the information used by G^{RAS} to compute the new iterate. The formal equivalence between RAS and SRAS is shown in the following theorem.

Theorem 6.2.1 (Equivalence between RAS and SRAS)**.** *Assume that the operators R_S and P_S satisfy Assumption* (6.2.11)*. Then given an initial guess $u^0 \in V$ and its substructured restriction $v^0 := R_S u^0 \in V^S$, define the sequences $\{u^n\}$ and $\{v^n\}$ such that*

$$u^n = G^{RAS}(u^{n-1}), \quad v^n = G^{SRAS}(v^{n-1}).$$

Then for every $n \geq 1$, $R_S u^n = v^n$.

Proof. We prove the statement for $n = 1$ through a direct calculation. Taking the restriction of u^1 we have

$$R_S u^1 = R_S G^{\mathrm{RAS}}(u^0) = R_S G^{\mathrm{RAS}}(P_S R_S u^0) = R_S G^{\mathrm{RAS}}(P_S v^0) = G^{\mathrm{SRAS}}(v^0) = v^1,$$

where we used assumption (6.2.11) and the definition of v^0. The other cases follow by induction. □

Similarly to the linear case, we can define a substructured RAS method in the nonlinear case. Defining

$$G_j^S(v^{n-1}) := R_S \widetilde{P}_j G_j\left(P_s v^{n-1}\right), \tag{6.2.12}$$

we obtain the nonlinear substructured iteration,

$$v^n = R_S \sum_{j\in\mathscr{J}} \widetilde{P}_j G_j(P_S v^{n-1}) = \sum_{j\in\mathscr{J}} G_j^S(v^{n-1}). \tag{6.2.13}$$

The same identical calculations of Theorem 6.2.1 allow one to obtain an equivalence result between nonlinear RAS and nonlinear substructured RAS.

Theorem 6.2.2 (Equivalence nonlinear RAS and SRAS)**.** *Assume that the operators R_S and P_S satisfy $R_S\sum_{j\in\mathscr{J}}\widetilde{P}_j G_j(u) = R_S\sum_{j\in\mathscr{J}}\widetilde{P}_j G_j(P_S R_S u)$. Then given an initial guess $u^0 \in V$ and its substructured restriction $v^0 := R_S u^0 \in V^S$, define the sequences $\{u^n\}$ and $\{v^n\}$ such that*

$$u^n = \sum_{j\in\mathscr{J}} \widetilde{P}_j G_j(u^{n-1}), \quad v^n = \sum_{j\in\mathscr{J}} G_j^S(v^{n-1}).$$

Then for every $n \geq 1$, $R_S u^n = v^n$.

In the manuscript [59], it has been proposed to use the fixed point equation of the nonlinear RAS method as a preconditioner for Newton's method, in a spirit that goes back to [19, 18]. This method has been called RASPEN (Restricted Additive Schwarz Preconditioning Exact Newton) and it consists in applying the Newton method to the fixed point equation

$$\mathscr{F}(u) = u - \sum_{j\in\mathscr{J}} \widetilde{P}_j G_j(u) = 0. \tag{6.2.14}$$

Here and in the article in preparation [29], we analyze a substructured version of the RASPEN method thus called SRASPEN. It consists in applying Newton's method to the fixed point equation

$$\mathscr{F}^S(v) = R_S\mathscr{F}(P_S v) = R_S P_S v - \sum_{j\in\mathscr{J}} R_S\widetilde{P}_j G_j(P_S v) = v - \sum_{j\in\mathscr{J}} G_j^S(v) = 0. \tag{6.2.15}$$

6.2.1 Computation of the Jacobian and implementation details

To apply Newton's method, we need to compute the Jacobian of SRASPEN. Since SRASPEN and RASPEN methods are closely related, indeed $\mathscr{F}^S(v) = R_S\mathscr{F}(P_S v)$, we can immediately compute the Jacobian of $\mathscr{F}^S$ once we have the Jacobian of $\mathscr{F}$, through the chain rule $J_{\mathscr{F}^S}(v) = R_S J_{\mathscr{F}}(P_S v)P_S$. The Jacobian of $\mathscr{F}$ has been derived in [59] and we report the main steps for the sake of completeness. Differentiating equation (6.2.14) with respect to u leads to

$$J_{\mathscr{F}}(u) := \frac{d\mathscr{F}}{du}(u) = I - \sum_{j\in\mathscr{J}} \widetilde{P}_j \frac{dG_j}{du}(u), \tag{6.2.16}$$

where $J_{\mathscr{F}}(w)$ denotes the action of the Jacobian of RASPEN on a vector w. Recall that the local inverse operators $G_j : V \to V_j$ are defined in equation (6.2.7) as the solutions of $R_j F(P_j G_j(u) + (I - P_j R_j)u) = 0$. Differentiating this relation yields

$$\frac{dG_j}{du}(u) = R_j - \left(R_j J\left(u^{(j)}\right) P_j\right)^{-1} R_j J\left(u^{(j)}\right), \tag{6.2.17}$$

where $u^{(j)} := P_j G_j(u) + (I - P_j R_j)u$ is the volume solution vector in subdomain j and J is the Jacobian of the original nonlinear function F. Combining the above equations (6.2.16)-(6.2.17) and defining $\widetilde{u}^{(j)} := P_j G_j(P_S v) + (I - P_j R_j)P_S v$, we get

$$J_{\mathscr{F}}(u) = \left(\sum_{j\in\mathscr{J}} \tilde{P}_j \left(R_j J\left(u^{(j)}\right) P_j\right)^{-1} R_j J\left(u^{(j)}\right)\right), \tag{6.2.18}$$

$$J_{\mathscr{F}^S}(v) = R_S\left(\sum_{j\in\mathscr{J}} \tilde{P}_j \left(R_j J\left(\widetilde{u}^{(j)}\right) P_j\right)^{-1} R_j J(\widetilde{u}^{(j)})\right) P_S, \tag{6.2.19}$$

where we used the assumptions $\sum_{j\in\mathscr{J}} \widetilde{P}_j R_j = I$ and $R_S P_S = I_S$. We remark that to assemble $J_{\mathscr{F}}(u)$ or to compute its action on a given vector, one needs to calculate $J\left(u^{(j)}\right)$, that is evaluating the Jacobian of the original nonlinear function F on the subdomain solutions u^j. The subdomain solutions u^j are obtained evaluating $\mathscr{F}(u)$, that is performing one step of the RAS method with initial guess equal to u. A smart implementation can use that the local Jacobian matrices $R_j J\left(u^{(j)}\right) P_j$ are already computed by the inner Newton solvers while solving the nonlinear problem on each subdomain, and hence no extra cost is required to assemble this term. Further, the matrices $R_j J\left(u^{(j)}\right)$ are different from the local Jacobian matrices at very few columns corresponding to the degrees of freedom on the interfaces and thus one could only modify those specific entries. In a lazier implementation, one can directly evaluate the Jacobian of F on the subdomain solutions u^j, without relying on already computed quantities. Concerning $J_{\mathscr{F}^S}(v)$, we emphasize that $\widetilde{u}^{(j)}$ is the volume subdomain solution obtained by the substructured RAS method starting from a substructured function v. Thus, as $u^{(j)}$, $\widetilde{u}^{(j)}$ is readily available in a Newton's iteration after evaluating the function $\mathscr{F}^S$.

From the computational point of view, (6.2.18) has several implications. First, the substructured Jacobian $J_{\mathscr{F}^S}$ is a matrix of dimension $N_S \times N_S$ where N_S is the number of unknowns on S, and thus is a much smaller matrix than $J_{\mathscr{F}}$, whose size is $N_v \times N_v$, with N_v the number of unknowns in volume. Hence, at each Newton iteration, the SRASPEN method must solve a much smaller system compared to the RASPEN method. This is even more important if one does not rely on some Krylov method, but prefers to use a direct solver as the assembly of $J_{\mathscr{F}^S}$ is dramatically cheaper than for $J_{\mathscr{F}}$. Further implementation details and a more extensive comparison are available in the numerical section 6.5.

6.3 Convergence analysis of RASPEN and SRASPEN

Theorem (6.2.2) states an equivalence between the nonlinear iterative RAS method and the nonlinear iterative substructured RAS method. Are the RASPEN and SRASPEN meth-

ods equivalent? Does the Newton method behave differently if applied to the volume or to the substructured fixed point equation? In this section we aim to answer these questions, discussing the convergence properties of the exact Newton method applied to $\mathscr{F}$ and $\mathscr{F}^S$. Let us remember that, given two approximations u^0 and v^0, the exact Newton method computes for $n \geq 1$,

$$u^n = u^{n-1} - \left(J_{\mathscr{F}}\left(u^{n-1}\right)\right)^{-1} \mathscr{F}\left(u^{n-1}\right), \quad v^n = v^{n-1} - \left(J_{\mathscr{F}^S}\left(v^{n-1}\right)\right)^{-1} \mathscr{F}^S(v^{n-1}), \tag{6.3.1}$$

where $J_{\mathscr{F}}\left(u^{n-1}\right)$ and $J_{\mathscr{F}^S}\left(v^{n-1}\right)$ are the Jacobian matrices respectively of $\mathscr{F}$ and $\mathscr{F}^S$ evaluated at u^{n-1} and v^{n-1}. In this paragraph we do not need a precise expression for $J_{\mathscr{F}}$ and $J_{\mathscr{F}^S}$. However we recall that, by definition $\mathscr{F}^S(v) = R_S\mathscr{F}(P_S v)$, so that the chain rule derivation leads to $J_{\mathscr{F}^S}(v) = R_S J_{\mathscr{F}}(P_S v)P_S$. If the operators R_S and P_S were square matrices, we would immediately obtain that the RASPEN and SRASPEN methods are equivalent, due to the affine invariance theory for Newton's method [51]. However, in our case R_S and P_S are rectangular matrices, mapping between spaces of different dimensions. Nevertheless, in the following theorem we show that the RASPEN and SRASPEN methods provide the same iterates restricted on the interfaces under further assumptions on R_S and P_S.

Theorem 6.3.1 (Equivalence RASPEN and SRASPEN)**.** *Assume that the operators R_S and P_S satisfy*

$$R_S\mathscr{F}(u) = R_S\mathscr{F}(P_S R_S u) = \mathscr{F}^S(R_S u). \tag{6.3.2}$$

Given an initial guess $u^0 \in V$ and its substructured restriction $v^0 := R_S u^0 \in V^S$, define the sequences $\{u^n\}$ and $\{v^n\}$ such that

$$u^n = u^{n-1} - \left(J_{\mathscr{F}}\left(u^{n-1}\right)\right)^{-1} \mathscr{F}\left(u^{n-1}\right), \quad v^n = v^{n-1} - \left(J_{\mathscr{F}^S}\left(v^{n-1}\right)\right)^{-1} \mathscr{F}^S\left(v^{n-1}\right).$$

Then for every $n \geq 1$, $R_S u^n = v^n$.

Proof. We prove the equality $R_S u^1 = v^1$ through direct calculations and the general case is obtained by induction. Taking the restriction of the RASPEN iteration

$$R_S u^1 = R_S u^0 - R_S\left(J_{\mathscr{F}}\left(u^0\right)\right)^{-1} \mathscr{F}\left(u^0\right) = v^0 - R_S\left(J_{\mathscr{F}}\left(u^0\right)\right)^{-1} \mathscr{F}\left(u^0\right),$$

and we deduce that to prove $R_S u^1 = v^1$ we need to show that

$$R_S\left(J_{\mathscr{F}}\left(u^0\right)\right)^{-1} \mathscr{F}\left(u^0\right) = \left(J_{\mathscr{F}^S}\left(v^0\right)\right)^{-1} \mathscr{F}^S\left(v^0\right). \tag{6.3.3}$$

Due to the definition of $\mathscr{F}^S$ and of v^0, and to the assumption (6.3.2), we have

$$\mathscr{F}^S\left(v^0\right) = R_S\mathscr{F}\left(P_S v^0\right) = R_S\mathscr{F}\left(P_S R_S u^0\right) = R_S\mathscr{F}\left(u^0\right),$$

which substituted into (6.3.3) leads to

$$R_S\left(J_{\mathscr{F}}\left(u^0\right)\right)^{-1} \mathscr{F}\left(u^0\right) = \left(J_{\mathscr{F}^S}\left(v^0\right)\right)^{-1} R_S\mathscr{F}\left(u^0\right),$$

Algorithm 9: Two-level iterative RAS

1: Solve the coarse problem $F_0\left(y\right) = F_0\left(R_0 u^k\right) - R_0 F\left(u^k\right)$ and set $C_0\left(u^k\right) = y - R_0 u^k$.
2: Add the coarse correction to the current iterate $u^{k+\frac{1}{2}} = u^k + P_0 C_0\left(u^k\right)$.
3: Compute one step of the RAS method, $u^{k+1} = \sum_{j\in\mathscr{J}} \widetilde{P}_j G_j\left(u^{k+\frac{1}{2}}\right)$.
4: Repeat from 1 to 3 until convergence.

which holds true if $R_S\left(J_{\mathscr{F}}\left(u^0\right)\right)^{-1} = \left(J_{\mathscr{F}^S}\left(v^0\right)\right)^{-1} R_S$. Multiplying on the left and on the right by the Jacobians, and since $J_{\mathscr{F}^S}\left(v^0\right) = R_S J_{\mathscr{F}}\left(P_S R_S u^0\right) P_S$, we are left to show that

$$R_S J_{\mathscr{F}}\left(P_S R_S u^0\right) P_S R_S = R_S J_{\mathscr{F}}\left(u^0\right),$$

which is trivially true taking the Jacobian of the assumption (6.3.2). □

6.4 Two-level methods

In this section we consider two level versions of the iterative RAS and SRAS methods as well as of the RASPEN and SRASPEN methods. To define a two-level method, we introduce a coarse space $V_0 \subset V$, a restriction operator $R_0 : V \to V_0$ and an interpolation operator $P_0 : V_0 \to V$. The nonlinear system F can be projected onto the coarse space V_0, defining the coarse nonlinear function $F_0(u_0) := R_0 F(P_0 u_0)$, for every $u_0 \in V_0$. Due to the definition, it follows immediately that $J_0(u_0) = R_0 J(P_0 u_0) P_0$, $\forall\, u_0 \in V_0$. To compute a coarse correction we rely on the FAS approach [15]. Given a current approximation u, the coarse correction $C_0(u)$ is computed as the solution of

$$F_0(C_0(u) + R_0 u) = F_0(R_0 u) - R_0 F(u). \tag{6.4.1}$$

The two-level RAS method is described by Algorithm 9 and it consists of a coarse correction followed by one iteration of the RAS method.

We now focus on a substructured counterpart. We introduce a coarse substructured space $V_0^S \subset V^S$, a restriction operator $\overline{R}_0 : V^S \to V_0^S$ and a prolongation operator $\overline{P}_0 : V_0^S \to V^S$. We define the coarse substructured function as

$$\mathscr{F}_0^S(v_0) := \overline{R}_0 \mathscr{F}^S(\overline{P}_0(v_0)), \quad \forall\, v_0 \in V_0^S. \tag{6.4.2}$$

From the definition it follows that $J_{\mathscr{F}_0^S}(v_0) = \overline{R}_0 J_{\mathscr{F}^S}(\overline{P}_0 v_0)\overline{P}_0$, $\forall\, v_0 \in V_0^S$. There is a profound difference between the two-level volume RAS and the two-level substructured RAS: in the first one (Algorithm 9), the coarse function is obtained restricting the original nonlinear system $F(u) = 0$ onto a coarse mesh. In the substructured version, the coarse substructured function is defined restricting onto V_0^S the substructured fixed point equation. That is, the coarse substructured function corresponds to a coarse version of the SRASPEN method. Hence, we remark that this algorithm is exactly the nonlinear countepart of the linear 2-level algorithm described in Chapter 4, where, similarly, the coarse

Algorithm 10: Two-level substructured iterative RAS

1: Solve the coarse problem $\mathscr{F}_0^S\left(y\right) = \mathscr{F}_0^S\left(\overline{R}_0 v^k\right) - \overline{R}_0 \mathscr{F}^S\left(v^k\right)$ and set $C_0^S\left(v^k\right) = y - \overline{R}_0 v^k$.

2: Add the coarse correction to the current iterate $v^{k+\frac{1}{2}} = v^k + \overline{P}_0 C_0^S\left(v^k\right)$.

3: Compute one-step of the RAS method, $v^{k+1} = \sum_{j\in\mathscr{J}} G_j^S\left(v^{k+\frac{1}{2}}\right)$.

4: Repeat from 1 to 3 until convergence.

substructured matrix did not involve the original volume matrix, but the fixed point equation (4.1.13). The substructured two-level iterative RAS is then defined in Algorithm 10. As in the linear case, numerical experiments will show that the substructured two-level method exhibits a faster convergence in terms of iteration counts compared to the two-level volume RAS method. However, we remark that evaluating F_0 is rather cheap, while evaluating $\mathscr{F}_0^S$ could be quite expensive as it requires to perform subdomain solves on the fine mesh. One possible improvement is to approximate $\mathscr{F}_0^S$ replacing $\mathscr{F}^S$ in its definition with another function which performs subdomain solves on a coarse mesh, much in the spirit of the coarse matrix $\widetilde{A}_c$ in (4.4.1) for the linear case. Further, we emphasize that a prerequisite of any domain decomposition method is that the subdomain solves are cheap to compute in a high performance parallel implementation, so that in such a setting evaluating $\mathscr{F}_0^S$ should be cheap as well.

Once we have defined the two-level iterative methods, we are ready to introduce the two-level versions of the RASPEN and SRASPEN methods. The fixed point equation of the two-level RAS method is

$$\mathscr{F}_{2L}(u) := u - \sum_{j\in\mathscr{J}} \widetilde{P}_j G_j(u + P_0 C_0(u)) = -P_0 C_0(u) - \sum_{j\in\mathscr{J}} \widetilde{P}_j C_j(u + P_0 C_0(u)) = 0, \quad (6.4.3)$$

where we have introduce the correction operators $C_j(u) := G_j(u) - R_j u$. Thus, the two-level RASPEN method, defined in [59], consists of applying Newton's method to the fixed point equation (6.4.3).
Similarly, the fixed point equation of the substructured two-level RAS method is

$$\mathscr{F}_{2L}^S(v) := v - \sum_{j\in\mathscr{J}} G_j^S\left(v + \overline{P}_0 C_0^S(v)\right) = -\overline{P}_0 C_0^S(v) - \sum_{j\in\mathscr{J}} C_j^S\left(v + \overline{P}_0 C_0^S(v)\right) = 0, \quad (6.4.4)$$

where we have introduced the correction operators $C_j^S(v) := G_j^S(v) - R_S \widetilde{P}_j R_j P_S v$. The two-level SRASPEN method consists in applying Newton's method to the fixed point equation (6.4.4).

6.4.1 Computation of the Jacobian and implementation details

In this subsection, we discuss how to compute the Jacobian matrices for the two-level versions of the RASPEN and SRASPEN methods. First, we consider the RASPEN method and we explain the calculations already reported in [59]. To compute the Jacobian of $\mathscr{F}_{2L}(u)$

we need $\frac{dC_j}{du}$, $j = 1, ..., N$ and $\frac{dC_0}{du}$. Recalling that $C_j(u) = G_j(u) - R_j u$ and (6.2.17), it follows that

$$\frac{dC_j}{du}(u) = \frac{dG_j}{du}(u) - R_j = -(R_j J\left(u^j\right) P_j)^{-1} R_j J\left(u^j\right),$$

where $u^j := P_j G_j(u) + (I - P_j R_j)u = u + P_j C_j(u)$. To compute the term $\frac{dC_0}{du}$ we differentiate (6.4.1) and we get

$$J_0(R_0 u + C_0(u))\left(R_0 + \frac{dC_0}{du}\right) = J_0(R_0 u)R_0 - R_0 J(u),$$

which implies

$$\frac{dC_0}{du}(u) = -R_0 + \widetilde{J}_0^{-1}(\widehat{J}_0 R_0 - R_0 J(u)),$$

with $\widetilde{J}_0 = J_0(R_0 u + C_0(u))$ and $\widehat{J}_0 = J_0(R_0 u)$, that is the same Jacobian but evaluated on different functions. Finally, using the chain rule while differentiating (6.4.3), we obtain

$$\frac{d\mathscr{F}_{2L}}{du}(u) = -P_0 \frac{dC_0}{du}(u) + \sum_{j \in \mathscr{J}} \widetilde{P}_j (R_j J(u^j) P_j)^{-1} R_j J(u^j)\left(I + P_0 \frac{dC_0}{du}(u)\right), \tag{6.4.5}$$

where $u^j = P_j G_j(u + P_0 C_0(u)) + (I - P_j R_j)(u + P_0 C_0(u))$. We remark once more that these quantities are readily available at each Newton iteration, since $u + P_0 C_0(u)$ and $G_j(u + P_0 C_0(u))$ are computed when evaluating $\mathscr{F}_{2L}(u)$.

We now focus on the substructured two-level function $\mathscr{F}^S_{2L}$. Differentiating (6.4.2) yields

$$J_{\mathscr{F}^S_0}(\overline{R}_0 v + C^S_0(v))\left(R_0 + \frac{dC^S_0}{dv}(v)\right) = J_{\mathscr{F}^S_0}(\overline{R}_0 v)\overline{R}_0 - \overline{R}_0 J_{\mathscr{F}^S}(v),$$

which leads to

$$\frac{dC^S_0}{dv}(v) = -\overline{R}_0 + \widetilde{J}^{-1}\left(\widehat{J}\overline{R}_0 - \overline{R}_0 J_{\mathscr{F}^S}(v)\right), \tag{6.4.6}$$

where $\widetilde{J} := J_{\mathscr{F}^S_0}(\overline{R}_0 v + C^S_0(v))$ and $\widehat{J} := J_{\mathscr{F}^S_0}(\overline{R}_0 v)$. The Jacobian $J_{\mathscr{F}^S_0}$ can be computed combining equations (6.4.2) and (6.2.18). Furthermore, we need the subdomain extensions of $\overline{R}_0 v$ and $\overline{R}_0 v + C^S_0(v)$ but again, these terms are already computed when evaluating the right hand side of the FAS equation.

Moreover, since $C^S_j(v) = G^S_j(v) - R_S \widetilde{P}_j R_j P_S(v)$ and $G^S_j(v) = R_S \widetilde{P}_j G_j(P_s v)$, it holds that

$$\begin{aligned}\frac{dC^S_j}{dv}(v) &= \frac{dG^S_j}{dv}(v) - R_S \widetilde{P}_j R_j P_S = R_S \widetilde{P}_j \left[(R_j J_F(\widetilde{u}^j) P_j)^{-1} R_j J_F(\widetilde{u}^j) + R_j\right] P_S - R_S \widetilde{P}_j R_j P_S \\ &= -R_S \widetilde{P}_j (R_j J_F(\widetilde{u}^j) P_j)^{-1} R_j J_F(\widetilde{u}^j).\end{aligned} \tag{6.4.7}$$

Using (6.4.6) and (6.4.7), we finally obtain

$$\frac{d\mathscr{F}^S_{2L}}{dv}(v) = -P_0 \frac{dC^S_0}{dv}(v) + \sum_{j \in \mathscr{J}} R_S \widetilde{P}_j (R_j J(\widetilde{u}^j) P_j)^{-1} R_j J(\widetilde{u}^j)\left(I + P_0 \frac{dC^S_0}{dv}(v)\right), \tag{6.4.8}$$

where $\widetilde{u}^j = P_j G_j(P_S(v + \overline{P}_0 C^S_0(v)) + (I - P_j R_j)(P_S(v + \overline{P}_0 C^S_0(v)))$.

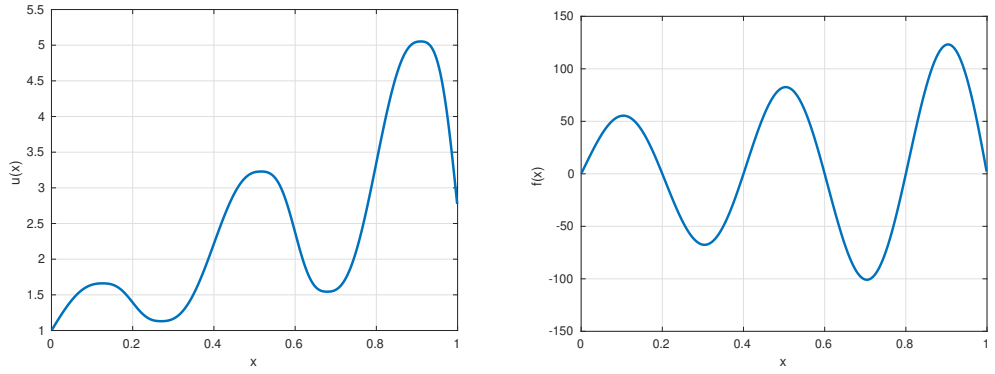


Figure 6.2: Solution field $u(x)$ of the Forchheimer equation (left panel) and force term $f(x)$ (right panel).

6.5 Numerical results

In this section we present numerical results in order to compare the Newton method, the nonlinear RAS method, the RASPEN method and the SRASPEN method for the solution of a one dimensional Forchheimer's equation and of a two dimensional nonlinear diffusion equation.

6.5.1 Forchheimer's equation in 1D

Forchheimer's equation is an extension of the Darcy equation for high flow rates, where the linear relation between the flow velocity and the gradient flow does not hold anymore. In a one dimensional domain $\Omega := (0,1)$, the Forchheimer model is

$$\begin{aligned} &q(-\lambda(x)u(x)'))' = f(x) \quad \text{in } \Omega, \\ &u(0) = u_L \quad \text{and} \quad u(1) = u_R, \end{aligned} \tag{6.5.1}$$

where $u_L, u_R \in \mathbb{R}$, $\lambda(x)$ is a positive and bounded permeability field and $q(y) := \text{sign}(y)\frac{-1+\sqrt{1+4\gamma|y|}}{2\gamma}$, with $\gamma > 0$. To discretize (6.5.1), we use the finite volume scheme described in detail in [59]. In our numerical experiments, we set $\lambda(x) = 2+\cos(5\pi x)$, $f(x) = 50\sin(5\pi x)e^x$, $\gamma = 1$, $u(0) = 1$ and $u(1) = e$. The solution field $u(x)$ and the force field $f(x)$ are shown in Figure 6.2. We then study the convergence behaviour of the different methods. Figure 6.3 shows how the relative error decays for the different methods and for a decomposition into 20 subdomains (left panel) and 50 subdomains (right panel). The initial guess is equal to zero for all the methods. From Figure 6.3, it seems that the convergence of the RASPEN and SRASPEN methods is not affected by the number of subdomains. However, Figure 6.3 does not tell the whole story, as one should focus not only on the number of iterations, but also on the cost of each iteration. To compare the cost of an iteration of the RASPEN and SRASPEN methods, we have to distinguish two cases, that is if one solves the Jacobian system directly or with some Krylov methods, e.g GMRES. First suppose that we want to

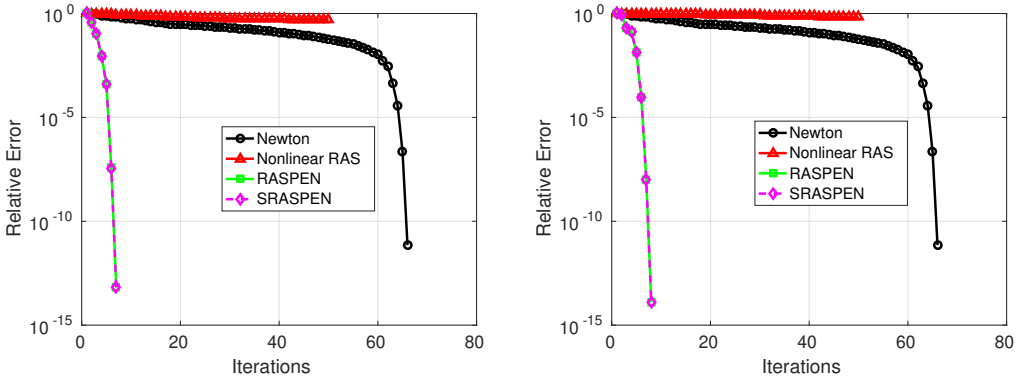


Figure 6.3: Convergence behaviour for the Newton method, the nonlinear RAS method, the RASPEN method and the SRASPEN method applied to the Forchheimer's equation. On the left, the simulation refers to a decomposition into 20 subdomains while on the right we consider 50 subdomains. The mesh size is $h = 10^{-3}$ and the overlap is $8h$.

solve the Jacobian system with a direct method and thus we need to assemble and store the Jacobians. From the expressions in equation (6.2.18) we remark that the assembly of the Jacobian of the RASPEN method requires $N \times N_v$ subdomain solves, where N is the number of subdomains and N_v is the number of unknowns in volume. On the other hand, the assembly of the Jacobian of SRASPEN method requires $N \times N_S$ solves, where N_S is the number of unknowns on the substructures and $N_S \ll N_v$. Thus, while the assembly of $J_{\mathscr{F}}$ is prohibitive, it can still be affordable to assemble $J_{\mathscr{F}^S}$. Further, the direct solution of the Jacobian system is feasible and cheap as $J_{\mathscr{F}^S}$ has size $N_S \times N_S$. Suppose now that we solve the Jacobian systems with GMRES and we indicate with $I(k)$ and $I^S(k)$ the number of GMRES iterations to solve the volume and substructured Jacobian systems at the k-th outer Newton's iteration. Each GMRES iteration requires N subdomain solves which can be performed in parallel. In our numerical experiment we have observed that generally $I^S(k) \leq I(k)$, with $I(k) - I^S(k) \approx 0, 1, 2$, that is GMRES requires the same number of iterations or slightly less to solve the substructured Jacobian system compared to the volume one.

To compare the two methods fairly, we follow [59] and introduce the quantity $L(n)$ which counts the number of subdomain solves performed by the two methods up to iteration n, taking into account the advantages of a parallel implementation. We set $L(n) = \sum_{k=1}^{n} L_{in}^k + I(k)$, where L_{in}^k is the maximum over the subdomains of the number of Newton's iterations required to solve the local subdomain problems at iteration k. The number of linear solves performed by GMRES should be $I(k) \cdot N$, but as the N linear solves can be performed in parallel, the total cost of GMRES corresponds approximately to $I(k)$ linear solves. Figure 6.4 shows the error decay with respect to $L(n)$. We note that the two methods require approximately the same computational cost and SRASPEN is slightly faster. For the decomposition into 50 subdomains, the RASPEN method requires on average 91.5 GMRES iterations per Newton iteration, while the SRASPEN method requires an average

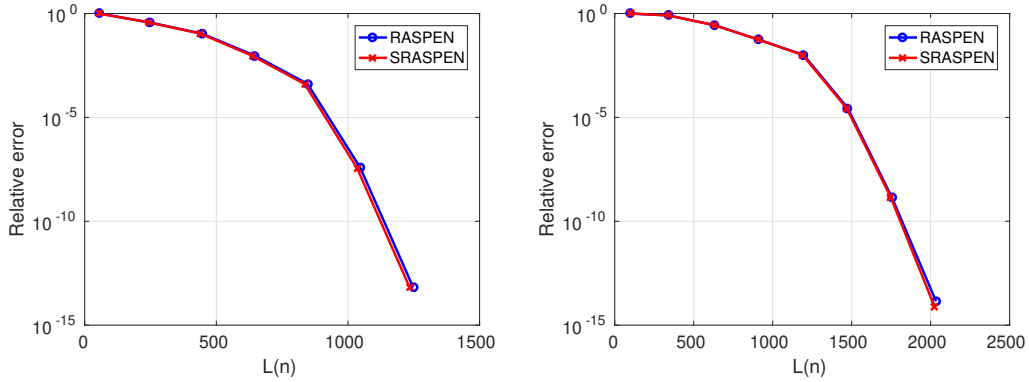


Figure 6.4: Relative error decay for the RASPEN method and the SRASPEN method applied to the Forchheimer equation with respect to the number of linear solves. On the left, the simulation refers to a decomposition into 20 subdomains while on the right we consider 50 subdomains. The mesh size is $h = 10^{-3}$.

of 90.87 iterations. The size of the substructured space V^S is $N_S = 98$. For the decomposition into 20 subdomains, the RASPEN method requires an average of 40 GMRES iterations per Newton's iteration, while the SRASPEN method needs 38 iterations. The size of V^S is $N_S = 38$, which means that GMRES reaches the given tolerance of 10^{-12} after exactly N_S steps which is the size of the substructured Jacobian. Under these circumstances, it can be convenient to actually assemble $J_{\mathcal{F}^S}$, as it requires $N_S \cdot N$ subdomain solves which is the total cost of GMRES. Furthermore, the $N_S \cdot N$ subdomain solves are embarrassingly parallel, while the $N_S \cdot N$ solves of GMRES can be parallelized in the spatial direction, but not in the iterative one. As a future work, we believe it will be interesting to study the convergence of a Quasi-Newton method based on the SRASPEN method, where one assembles the Jacobian substructured matrix after every few outer Newton iterations, reducing the overall computational cost.

As a final remark, we specify that Figure 6.4 has been obtained setting a zero initial guess for the nonlinear subdomain problems. However, at the iteration k of the RASPEN method one can use the subdomain restriction of the updated volume solution, that is $R_j u^{k-1}$, which has been obtained by solving the volume Jacobian system at iteration $k-1$ and is thus generally a better initial guess for the next iteration. On the other hand in the SRASPEN method, one could use the subdomain solutions computed at iteration $k-1$, i.e. u_i^{k-1}, as initial guesses for the nonlinear subdomain problems, as the substructured Jacobian system corrects only the substructured values. Numerical experiments showed that with this particular choice of initial guesses for the nonlinear subdomain problems, the SRASPEN method requires generally more Newton iterations to solve the local problems. In this setting, there is not a method which is constantly faster than the other as it depends on a delicate trade-off between the better GMRES performance and the need to perform more Newton iterations for the nonlinear local problems in the SRASPEN method.

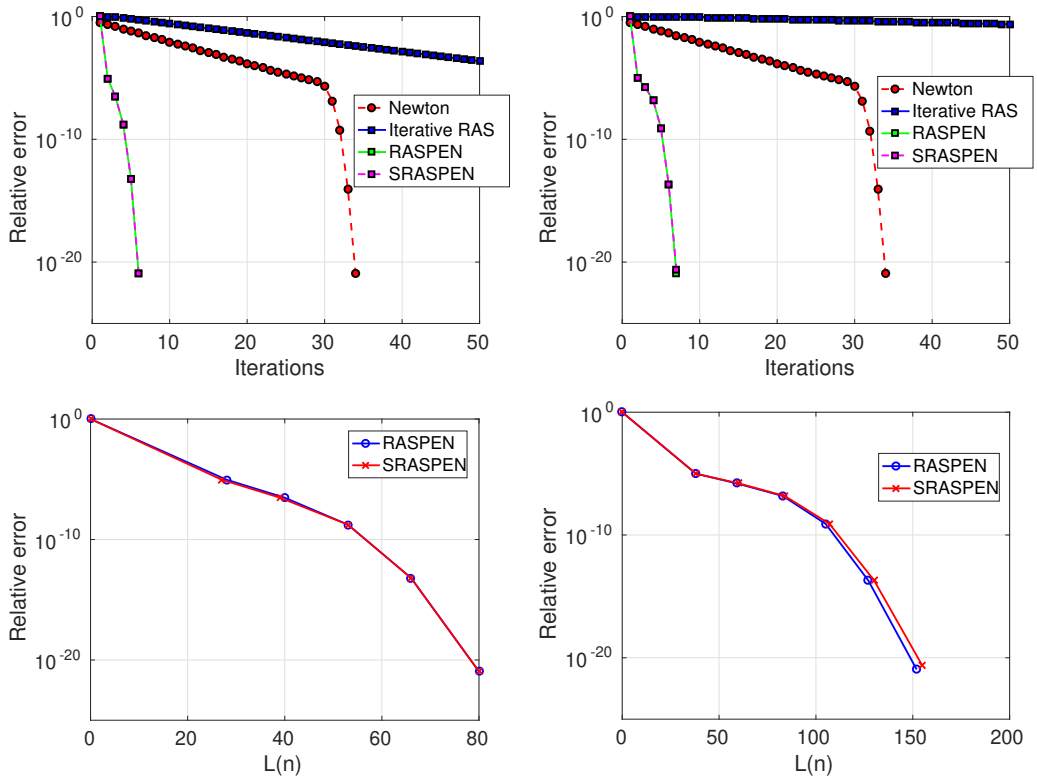


Figure 6.5: Relative error decay versus the number of iterations (top row) and error decay versus number of linear solves (bottom row). The left figures refer to a decomposition into 4 subdomains, while the right figures to a decomposition into 25 subdomains. The mesh size is $h = 0.012$ and the overlap is $8h$.

6.5.2 Nonlinear Diffusion

In this subsection we consider the nonlinear diffusion problem

$$\begin{aligned} &-\nabla\cdot\left(1+u(x)^2\right)\nabla u(x) = f, \quad \text{in } \Omega\subset\mathbb{R}^2,\\ &u(x) = g(x) \quad \text{on } \partial\Omega, \end{aligned} \tag{6.5.2}$$

where Ω is a square domain and the right hand side f is such that $u(x) = \sin(\pi x)\sin(\pi y)$ is the exact solution. We start all the methods with the initial guess $u^0 = 10^5$, so that we start far away from the exact solution, and hence Newton's method exhibits a long plateau before quadratic convergence begins. Figure 6.5 shows the convergence behaviour for the different methods with respect to the number of iterations and the number of linear solves. The average number of GMRES iterations is 8.1667 for both the RASPEN and SRASPEN methods for the 4 subdomain decomposition. For a decomposition into 25 subdomains, the average number of GMRES iterations is 19.14 for the RASPEN method and 19.57 for the SRASPEN method. We remark that as the number of subdomains increases, GMRES needs more iterations to solve the Jacobian system. This is consistent

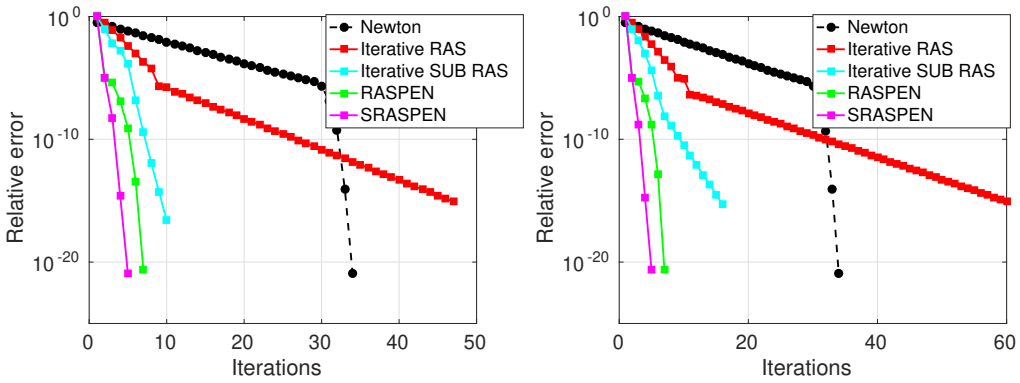


Figure 6.6: Relative error decay versus the number of iterations for the Newton method, the iterative two-level methods RAS and SRAS and the two-level variants of the RASPEN and SRASPEN methods. The left figure refers to a decomposition into 4 subdomains, while the right figure refers to a decomposition into 16 subdomains. The mesh size is $h = 0.012$ and the overlap is $4h$.

with the interpretation of (6.2.18) as a Jacobian matrix $J\left(u^{(j)}\right)$ preconditioned by the additive operator $\sum_{j\in\mathcal{J}}\left(R_j J\left(u^{(j)}\right)P_j\right)^{-1}$; We expect this preconditioner not to be scalable since it does not involve a coarse correction.

We conclude this chapter showing the convergence behaviour for the two-level variants of the RASPEN and SRASPEN methods. We use a coarse grid in volume taking half of the points in x and y, and a coarse substructured grid taking half of the unknowns as depicted in Figure 6.1. The interpolation and restriction operators $P_0, R_0, \overline{P}_0$ and $\overline{R}_0$ are the classical linear interpolation and full weighting restriction operators defined in Chapter 4. From Figure 6.6, we note that the substructured iterative method is much faster than the classical two-level RAS method, and this observation is in agreement with the linear case treated in Chapter 4. Since the two-level iterative methods are not equivalent, we also remark that the two-level SRASPEN methods shows a better performance than the two-level RASPEN method in terms of iteration number. As the one-level smoother is the same in all methods, the better convergence of the substructured methods implies that the coarse equation involving $\mathcal{F}_0^S$ provides a much better coarse correction than the classical volume one involving F_0.
Even though the two-level substructured methods are faster in terms of iteration number, the solution of the FAS problem involving $\mathcal{F}_0^S = \overline{R}_0\mathcal{F}^S(\overline{P}_0(v_0))$ is rather expensive as it requires to evaluate twice the substructured function $\mathcal{F}^S$ (each evaluation requires subdomain solves) to compute the right hand side, to solve a Jacobian system involving $J_{\mathcal{F}_0^S}$, and to evaluate $\mathcal{F}^S$ on the iterates, which again require the solution of subdomain problems. Unless one has a full parallel implementation available, the coarse correction involving $\mathcal{F}_0^S$ is doomed to represent a bottleneck. Future efforts will be in the direction of approximating $\mathcal{F}_0^S$, by replacing the function $\mathcal{F}^S$, which is defined on a fine mesh, with an

approximation on a very coarse grid, thus reducing the overall cost of the substructured coarse correction.

List of Figures

List of Tables

Bibliography

[1] J. Aarnes and T. Y. Hou. Multiscale domain decomposition methods for elliptic problems with high aspect ratios. *Acta Math. Appl. Sin.*, 18(1):63–76, 2002.

[2] O. Axelsson. *Iterative solution methods*. Cambridge university press, 1996.

[3] I. M. Babuska and S. A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Journal on numerical analysis*, 34(6):2392–2423, 1997.

[4] J. Bear. *Hydraulics of Groundwater*. Dover Books on Engineering. Dover Publications, 2012.

[5] J. Bear and A. Cheng. *Modeling Groundwater Flow and Contaminant Transport*. Theory and Applications of Transport in Porous Media. Springer Netherlands, 2010.

[6] G. S. Beavers and D. Joseph. Boundary conditions at a naturally permeable wall. *Journal of Fluid Mechanics*, 30(1):197–207, 1967.

[7] M. F. Benedetto, S. Berrone, S. Pieraccini, and S. Scialò. The virtual element method for discrete fracture network simulations. *Computer Methods in Applied Mechanics and Engineering*, 280:135–156, 2014.

[8] S. Berrone, S. Pieraccini, and S. Scialò. On simulations of discrete fracture network flows with an optimization-based extended finite element method. *SIAM Journal on Scientific Computing*, 35(2):A908–A935, 2013.

[9] S. Berrone, S. Pieraccini, and S. Scialò. A PDE-constrained optimization formulation for discrete fracture network flows. *SIAM Journal on Scientific Computing*, 35(2):B487–B510, 2013.

[10] N. Birgle, R. Masson, and L. Trenty. A domain decomposition method to couple nonisothermal compositional gas liquid Darcy and free gas flows. *Journal of Computational Physics*, 368:210–235, 2018.

[11] P. Bjorstad, M. J. Gander, A. Loneland, and T. Rahman. Does SHEM for additive Schwarz work better than predicted by its condition number estimate? *Domain Decomposition Methods in Science and Engineering XXIV, LNCSE, Springer*, pages 129–138, 2018.

[12] E. Blayo, D. Cherel, and A. Rousseau. Towards optimized Schwarz methods for the Navier-Stokes equations. *Journal of Scientific Computing*, 66:275–295, 2016.

[13] M. Bonazzoli, V. Dolean, I. G. Graham, E. A. Spence, and P. Tournier. Two-level preconditioners for the Helmholtz equation. *Domain Decomposition Methods in Science and Engineering XXIV*, pages 139–147, 2018.

[14] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Mathematics of Computation*, 55(191):1–22, 1990.

[15] A. Brandt and O. E. Livne. *Multigrid Techniques*. Society for Industrial and Applied Mathematics, 2011.

[16] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer New York, 2007.

[17] X.-C. Cai and M. Dryja. Domain decomposition methods for monotone nonlinear elliptic problems. *Contemporary Mathematics*, 180, 1994.

[18] X.-C. Cai and D. E. Keyes. Nonlinearly preconditioned inexact Newton algorithms. *SIAM Journal on Scientific Computing*, 24(1):183–200, 2002.

[19] X.-C. Cai, D. E. Keyes, and D. P. Young. A nonlinear additive Schwarz preconditioned inexact Newton method for shocked duct flow. In *Proceedings of the 13th International Conference on Domain Decomposition Methods*. Citeseer, 2001.

[20] X.-C. Cai and X. Li. Inexact Newton methods with restricted additive Schwarz based nonlinear elimination for problems with high local nonlinearity. *SIAM journal on scientific computing*, 33(2):746–762, 2011.

[21] X.-C. Cai and M. Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM journal on scientific computing*, 21(2):792–797, 1999.

[22] E. Cancès, Y. Maday, and B. Stamm. Domain decomposition for implicit solvation models. *The Journal of chemical physics*, 139(5):054111, 2013.

[23] Y. Cao, M. Gunzburger, X. He, and X. Wang. Robin–Robin domain decomposition methods for the steady-state Stokes–Darcy system with the Beavers–Joseph interface condition. *Numerische Mathematik*, 117(4):601–629, 2011.

[24] Y. Cao, M. Gunzburger, X. He, and X. Wang. Parallel, non-iterative, multi-physics domain decomposition methods for time-dependent Stokes-Darcy systems. *Mathematics of Computation*, 83(288):1617–1644, 2014.

[25] Y. Cao, M. Gunzburger, F. Hua, and X. Wang. Coupled Stokes-Darcy model with Beavers-Joseph interface boundary condition. *Communications in Mathematical Sciences*, 8(1):1–25, 2010.

[26] F. Chaouqui. *Optimal Coarse Space Correction for Domain Decomposition Methods*. PhD thesis, Université de Genève, 09/18 2018. ID: unige:121801.

[27] F. Chaouqui, E. T. Chow, and D. B. Szyld. Asynchronous domain decomposition methods for nonlinear pdes. *In preparation 2020*.

[28] F. Chaouqui, G. Ciaramella, M. J. Gander, and T. Vanzan. On the scalability of classical one-level domain-decomposition methods. *Vietnam Journal of Mathematics*, 46(4):1053–1088, Dec 2018.

[29] F. Chaouqui, M. J. Gander, P. M. Kumbhar, and T. Vanzan. Substructured iterative and preconditioner RAS method: the linear and nonlinear case. *in preparation*, 2020.

[30] F. Chaouqui, M. J. Gander, and K. Santugini-Répiquet. A continuous analysis of Neumann-Neumann methods: Scalability and New Coarse Spaces. *to appear in SIAM Journal of Scientific Computing, 2020*.

[31] F. Chaouqui, M. J. Gander, and K. Santugini-Repiquet. A coarse space to remove the logarithmic dependency in Neumann–Neumann methods. In *Domain Decomposition Methods in Science and Engineering XXIV*, pages 159–167. Springer, 2017.

[32] F. Chaouqui, M. J. Gander, and K. Santugini-Repiquet. On nilpotent subdomain iterations. In *Domain Decomposition Methods in Science and Engineering XXIII*, pages 125–133, Cham, 2017. Springer International Publishing.

[33] F. Chaouqui, M. J. Gander, and K. Santugini-Repiquet. A local coarse space correction leading to a well-posed continous Neumman-Neumann method in the presence of cross points. In *Domain Decomposition Methods in Science and Engineering XXV*, 2020.

[34] F. Chaouqui and D. B. Szyld. On optimal coarse space correction for restricted and optimized additive Schwarz method. *submitted, 2020*.

[35] W. Chen, M. Gunzburger, F. Hua, and X. Wang. A parallel Robin–Robin domain decomposition method for the Stokes–Darcy system. *SIAM Journal on Numerical Analysis*, 49(3):1064–1084, 2011.

[36] G. Ciaramella and M. J. Gander. Analysis of the parallel Schwarz method for growing chains of fixed-sized subdomains: Part I. *SIAM Journal on Numerical Analysis*, 55(3):1330–1356, 2017.

[37] G. Ciaramella and M. J. Gander. Analysis of the parallel Schwarz method for growing chains of fixed-sized subdomains: Part II. *SIAM Journal on Numerical Analysis*, 56(3):1498–1524, 2018.

[38] G. Ciaramella and M. J. Gander. Analysis of the parallel Schwarz method for growing chains of fixed-sized subdomains: Part III. *Electronic Transactions on Numerical Analysis*, 49:210–244, 2018.

[39] G. Ciaramella, M. J. Gander, L. Halpern, and J. Salomon. Methods of reflections: relations with Schwarz methods and classical stationary iterations, scalability and preconditioning. *The SMAI journal of computational mathematics*, 5:161–193, 2019.

[40] G. Ciaramella, M. Hassan, and B. Stamm. On the scalability of the Schwarz method. *The SMAI journal of computational mathematics*, 6:33–68, 2020.

[41] G. Ciaramella and T. Vanzan. Substructured two-level and multilevel domain decomposition methods. *arXiv preprint arXiv:1908.05537*, 2019.

[42] G. Ciaramella and T. Vanzan. Spectral substructured two-level domain decomposition methods. submitted 2020.

[43] G. Ciaramella and T. Vanzan. Substructured two-grid and multi-grid domain decomposition methods. submitted 2020.

[44] P. G. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications:*. Applied mathematics. Society for Industrial and Applied Mathematics, 2013.

[45] P. Cocquet, M. J. Gander, and X. Xiang. A finite difference method with optimized dispersion correction for the Helmholtz equation. *Domain Decomposition Methods in Science and Engineering XXIV*, pages 205–213, 2018.

[46] L. Conen, V. Dolean, R. Krause, and F. Nataf. A coarse space for heterogeneous Helmholtz problems based on the Dirichlet-to-Neumann operator. *Journal of Computational and Applied Mathematics*, 271:83 – 99, 2014.

[47] C. D'Angelo and P. Zunino. Robust numerical approximation of coupled Stokes' and Darcy's flows applied to vascular hemodynamics and biochemical transport. *ESAIM: M2AN*, 45(3):447–476, 2011.

[48] H. Darcy. *Les fontaines publiques de la ville de Dijon: Exposition et application des principes a suivre et des formules a employer dans les questions de distribution d'eau; ouvrage terminé par un appendice relatif aux fournitures d'eau de plusieurs villes au filtrage des eaux et a la fabrication des tuyaux de fonte, de plomb, de tole et de bitume*. Victor Dalmont, Libraire des Corps imperiaux des ponts et chaussées et des mines, 1856.

[49] A. K. Datta. Porous media approaches to studying simultaneous heat and mass transfer in food processes. i: Problem formulations. *Journal of food engineering*, 80(1):80–95, 2007.

[50] B. Deprès. *Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régimes harmoniques*. PhD thesis, Université Dauphine-Parix IX, 1991.

[51] P. Deuflhard. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2010.

[52] M. Discacciati. *Domain decomposition methods for the coupling of surface and groundwater flows*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2004.

[53] M. Discacciati and L. Gerardo-Giorda. Optimized Schwarz methods for the Stokes–Darcy coupling. *IMA Journal of Numerical Analysis*, 38(4):1959–1983, 2017.

[54] M. Discacciati and L. Gerardo-Giorda. Is minimising the convergence rate a good choice for efficient optimized Schwarz preconditioning in heterogeneous coupling? The Stokes-Darcy case. In *Domain Decomposition Methods in Science and Engineering XXIV*, pages 233–241, Cham, 2018. Springer International Publishing.

[55] M. Discacciati and A. Quarteroni. Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. *Rev. Mat. Complut*, 22(2):315–426, 2009.

[56] M. Discacciati, A. Quarteroni, and A. Valli. Robin–Robin domain decomposition methods for the Stokes–Darcy coupling. *SIAM Journal on Numerical Analysis*, 45(3):1246–1268, 2007.

[57] C. R. Dohrmann, A. Klawonn, and O. B. Widlund. A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners. In *Domain Decomposition Methods in Science and Engineering XVII*, pages 247–254, 2008.

[58] V. Dolean, M. J. Gander, and L. Gerardo-Giorda. Optimized Schwarz methods for Maxwell's equations. *SIAM Journal on Scientific Computing*, 31(3):2193–2213, 2009.

[59] V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson. Nonlinear preconditioning: How to use a nonlinear Schwarz method to precondition Newton's method. *SIAM Journal on Scientific Computing*, 38(6):A3357–A3380, 2016.

[60] V. Dolean, M. J. Gander, and E. Veneros. Optimized Schwarz methods for Maxwell equations with discontinuous coefficients. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 517–525. Springer, Cham, 2014.

[61] V. Dolean, P. Jolivet, and F. Nataf. *An introduction to domain decomposition methods: algorithms, theory, and parallel implementation*, volume 144. SIAM, 2015.

[62] V. Dolean and F. Nataf. A new domain decomposition method for the compressible Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 40(4):689–703, 2006.

[63] 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. Meth. in Appl. Math.*, 12(4):391–414, 2012.

[64] O. Dubois. *Optimized Schwarz methods for the advection-diffusion equation*. PhD thesis, Mcgill University, 2007.

[65] O. Dubois. Optimized Schwarz methods with Robin conditions for the advection-diffusion equation. *Lecture Notes in Computational Science and Engineering*, 55:181, 2007.

[66] O. Dubois and M. J. Gander. Convergence behavior of a two-level optimized Schwarz preconditioner. In Michel Bercovier, Martin J. Gander, Ralf Kornhuber, and Olof Widlund, editors, *Domain Decomposition Methods in Science and Engineering XVIII*, pages 177–184, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[67] O. Dubois, M. J. Gander, S. Loisel, A. St-Cyr, and D. B. Szyld. The optimized Schwarz method with a coarse grid correction. *SIAM Journal on Scientific Computing*, 34(1):A421–A458, 2012.

[68] 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.

[69] E. Efstathiou and M. J. Gander. Why restricted additive Schwarz converges faster than additive Schwarz. *BIT Numerical Mathematics*, 43(5):945–959, 2003.

[70] M. El Bouajaji, V. Dolean, M. J. Gander, and S. Lanteri. Comparison of a one and two parameter family of transmission conditions for Maxwell's equations with damping. In *Domain Decomposition Methods in Science and Engineering XX*, pages 271–278. Springer, 2013.

[71] O. G. Ernst and M. J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In *Numerical analysis of multiscale problems*, pages 325–363. Springer Berlin Heidelberg, 2012.

[72] J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Model. Sim.*, 8(4):1461–1483, 2010.

[73] J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high contrast media: Reduced dimension coarse spaces. *Multiscale Model. Sim.*, 8(5):1621–1644, 2010.

[74] M. J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006.

[75] M. J. Gander. Schwarz methods over the course of time. *Electron. Trans. Numer. Anal*, 31(5):228–255, 2008.

[76] M. J. Gander. On the influence of geometry on optimized Schwarz methods. *SeMA Journal*, 53(1):71–78, 2011.

[77] M. J. Gander. On the origins of linear and non-linear preconditioning. In Chang-Ock Lee, Xiao-Chuan Cai, David E. Keyes, Hyea Hyun Kim, Axel Klawonn, Eun-Jae Park, and Olof B. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXIII*, pages 153–161, Cham, 2017. Springer International Publishing.

[78] M. J. Gander and O. Dubois. Optimized Schwarz methods for a diffusion problem with discontinuous coefficient. *Numerical Algorithms*, 69(1):109–144, 2015.

[79] M. J. Gander and G. H. Golub. A non-overlapping optimized Schwarz method which converges with arbitrarily weak dependence on h. *Domain Decomposition Methods in Science and Engineering XIV*, pages 281–288, 2003.

[80] M. J. Gander and L. Halpern. Méthodes de décomposition de domaines-notions de base. 2012.

[81] M. J. Gander, L. Halpern, and F. Magoules. An optimized Schwarz method with two-sided Robin transmission conditions for the Helmholtz equation. *International journal for numerical methods in fluids*, 55(2):163–175, 2007.

[82] M. J. Gander, L. Halpern, and V. Martin. A new algorithm based on factorization for heterogeneous domain decomposition. *Numerical Algorithms*, 73(1):167–195, Sep 2016.

[83] M. J. Gander, L. Halpern, and K. Santugini-Repiquet. Discontinuous coarse spaces for dd-methods with discontinuous iterates. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 607–615. Springer, 2014.

[84] M. J. Gander, L. Halpern, and K. Santugini-Repiquet. A new coarse grid correction for RAS/AS. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 275–283. Springer, 2014.

[85] M. J. Gander, P. M. Kumbhar, and A. E. Ruehli. Analysis of overlap in waveform relaxation methods for RC circuits. In *Domain Decomposition Methods in Science and Engineering XXIV*, pages 281–289, Cham, 2018. Springer International Publishing.

[86] M. J. Gander, P. M. Kumbhar, and A. E. Ruehli. Asymptotic analysis for different partitionings of RLC transmission lines. In *accepted for Domain Decomposition Methods in Science and Engineering XXV, LNCSE, Springer-Verlag*, 2019.

[87] M. J. Gander, P. M. Kumbhar, and A. E. Ruehli. Asymptotic analysis for overlap in waveform relaxation methods for RC type circuits. *Journal of Scientific Computing*, 84(24), 2020.

[88] M. J. Gander and F. Kwok. Optimal interface conditions for an arbitrary decomposition into subdomains. In *Domain Decomposition Methods in Science and Engineering XIX*, pages 101–108. Springer, 2011.

[89] M. J. Gander and A. Loneland. SHEM: An optimal coarse space for RAS and its multiscale approximation. In Chang-Ock Lee, Xiao-Chuan Cai, David E. Keyes, Hyea Hyun Kim, Axel Klawonn, Eun-Jae Park, and Olof B. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXIII*, pages 313–321, Cham, 2017. Springer International Publishing.

[90] M. J. Gander, A. Loneland, and T. Rahman. Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. *arXiv preprint arXiv:1512.05285*, 2015.

[91] M. J. Gander, F. Magoules, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM Journal on Scientific Computing*, 24(1):38–60, 2002.

[92] M. J. Gander and F. Nataf. AILU: a preconditioner based on the analytic factorization of the elliptic operator. *Numerical linear algebra with applications*, 7(7-8):505–526, 2000.

[93] M. J. Gander and F. Nataf. AILU for Helmholtz problems: a new preconditioner based on the analytic parabolic factorization. *Journal of Computational Acoustics*, 9(04):1499–1506, 2001.

[94] M. J. Gander and B. Song. Complete, optimal and optimized coarse spaces for additive Schwarz. In *Domain Decomposition Methods in Science and Engineering XXIV*, pages 301–309, Cham, 2018. Springer International Publishing.

[95] M. J. Gander and T. Vanzan. Optimized Schwarz methods for advection diffusion equations in bounded domains. In *European Conference on Numerical Mathematics and Advanced Applications*, pages 921–929. Springer, 2017.

[96] M. J. Gander and T. Vanzan. Heterogeneous optimized Schwarz methods for coupling Helmholtz and Laplace equations. In *Domain Decomposition Methods in Science and Engineering XXIV*, pages 311–320, Cham, 2018. Springer International Publishing.

[97] M. J. Gander and T. Vanzan. Heterogeneous optimized Schwarz methods for second order elliptic PDEs. *SIAM Journal on Scientific Computing*, 41(4):A2329–A2354, 2019.

[98] M. J. Gander and T. Vanzan. On the derivation of optimized transmission conditions for the Stokes-Darcy coupling. *accepted in Domain Decomposition in Science end Engineering XXV*, 2019.

[99] M. J. Gander and T. Vanzan. Multilevel optimized Schwarz methods. *to appear in SIAM Journal on Scientific Computing*, 2020.

[100] M. J. Gander and Y. Xu. Optimized Schwarz methods for model problems with continuously variable coefficients. *SIAM Journal on Scientific Computing*, 38(5):A2964–A2986, 2016.

[101] M. J. Gander and Y. Xu. Optimized Schwarz methods with nonoverlapping circular domain decomposition. *Mathematics of Computation*, 86(304):637–660, 2017.

[102] L. Gerardo-Giorda, F. Nobile, and C. Vergara. Analysis and optimization of Robin–Robin partitioned procedures in fluid-structure interaction problems. *SIAM Journal on Numerical Analysis*, 48(6):2091–2116, 2010.

[103] S. Gong and X.-C. Cai. A nonlinear elimination preconditioned Newton method with applications in arterial wall simulation. In *International Conference on Domain Decomposition Methods*, pages 353–361. Springer, 2017.

[104] S. Gong and X.-C. Cai. A nonlinear elimination preconditioned inexact Newton method for heterogeneous hyperelasticity. *SIAM Journal on Scientific Computing*, 41(5):S390–S408, 2019.

[105] I. G. Graham, E. A. Spence, and E. Vainikko. Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption. *Math. Comput.*, 86:2089–2127, 2017.

[106] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1985.

[107] C. Grüninger. *Numerical coupling of Navier-Stokes and Darcy flow for soil-water evaporation*. Stuttgart: Eigenverlag des Instituts für Wasser-und Umweltsystemmodellierung, 2017.

[108] Y. Gu. *Nonlinear optimized Schwarz preconditioning for heterogeneous elliptic problems*. PhD thesis, Hong Kong Baptist University, 2019.

[109] W. Hackbusch. *Multi-grid methods and applications*, volume 4. Springer Science & Business Media, 2013.

[110] R. Haferssas, P. Jolivet, and F. Nataf. An additive Schwarz method type theory for Lions's algorithm and a symmetrized optimized restricted additive Schwarz method. *SIAM Journal on Scientific Computing*, 39(4):A1345–A1365, 2017.

[111] X. He, J. Li, Y. Lin, and J. Ming. A domain decomposition method for the steady-state Navier–Stokes–Darcy model with Beavers–Joseph interface condition. *SIAM Journal on Scientific Computing*, 37(5):S264–S290, 2015.

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

[113] A. Heinlein, A. Klawonn, J. Knepper, and O Rheinbach. Multiscale coarse spaces for overlapping Schwarz methods based on the ACMS space in 2D. *Electron. Trans. Numer. Anal.*, 48:156–182, 2018.

[114] M. R Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

[115] C. Japhet. *Domain decomposition methods and artificial boundary conditions in fluid dynamics: Optimized Order 2 method.* Theses, Université Paris-Nord - Paris XIII, July 1998.

[116] C. Japhet, F. Nataf, and F. Rogier. The optimized order 2 method: application to convection–diffusion problems. *Future generation computer systems*, 18(1):17–30, 2001.

[117] A. Klawonn, M. Lanser, and O. Rheinbach. Nonlinear FETI-DP and BDDC methods. *SIAM Journal on Scientific Computing*, 36(2):A737–A765, 2014.

[118] A. Klawonn, P. Radtke, and O. Rheinbach. FETI-DP methods with an adaptive coarse space. *SIAM J. Numer. Anal.*, 53(1):297–320, 2015.

[119] P. M. Kumbhar. *Asymptotic analysis of optimized waveform relaxation methods for RC circuits and RLCG transmission lines.* PhD thesis, Université de Genève, 01/28 2020. ID: unige:136729.

[120] P. J. Lanzkron, D. J. Rose, and J. T. Wilkes. An analysis of approximate nonlinear elimination. *SIAM Journal on Scientific Computing*, 17(2):538–559, 1996.

[121] S.-C. Lee, M. N. Vouvakis, and J. F. Lee. A non-overlapping domain decomposition method with non-matching grids for modeling large finite antenna arrays. *Journal of Computational Physics*, 203(1):1 – 21, 2005.

[122] F. Lemarié, L. Debreu, and E. Blayo. Toward an Optimized Global-in-Time Schwarz Algorithm for Diffusion Equations with Discontinuous and Spatially Variable Coefficients, Part 1: The Constant Coefficients Case. *Electronic Transactions on Numerical Analysis*, 40:148–169, 2013.

[123] F. Lemarié, L. Debreu, and E. Blayo. Toward an Optimized Global-in-Time Schwarz Algorithm for Diffusion Equations with Discontinuous and Spatially Variable Coefficients, Part 2: the Variable Coefficients Case. *Electronic Transactions on Numerical Analysis*, 40:170–186, 2013.

[124] J.L. Lions and E. Magenes. *Non-homogeneous Boundary Value Problems and Applications (Vol I)*. Die Grundlehren der mathematischen Wissenschaften. Springer-Verlag Berlin Heidelberg, 1972.

[125] P.-L. Lions. On the Schwarz alternating method. I. In *First international symposium on domain decomposition methods for partial differential equations*, volume 1, page 42. Paris, France, 1988.

[126] 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*, volume 6, pages 202–223. SIAM Philadelphia, PA, 1990.

[127] F. Lipparini, B. Stamm, E. Cances, Y. Maday, and B. Mennucci. Fast domain decomposition algorithm for continuum solvation models: Energy and first derivatives. *Journal of chemical theory and computation*, 9(8):3637–3648, 2013.

[128] S.-H. Lui. On Schwarz alternating methods for nonlinear elliptic pdes. *SIAM Journal on Scientific Computing*, 21(4):1506–1523, 1999.

[129] S.-H. Lui. On monotone iteration and Schwarz methods for nonlinear parabolic pdes. *Journal of computational and applied mathematics*, 161(2):449–468, 2003.

[130] S.-H. Lui. A Lions non-overlapping domain decomposition method for domains with an arbitrary interface. *IMA Journal of Numerical Analsysis*, 29(2):332–349, 2009.

[131] Y. Maday and F. Magoulès. Optimized Schwarz methods without overlap for highly heterogeneous media. *Computer Methods in Applied Mechanics and Engineering*, 196(8):1541 – 1553, 2007. Domain Decomposition Methods: recent advances and new challenges in engineering.

[132] F. Magoules, D. B. Szyld, and C. Venet. Asynchronous optimized Schwarz methods with and without overlap. *Numerische Mathematik*, 137(1):199–227, 2017.

[133] V. Martin. Schwarz waveform relaxation method for the viscous shallow water equations. In *Domain Decomposition Methods in Science and Engineering*, pages 653–660, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[134] A. Mikelic and W. Jäger. On the interface boundary condition of Beavers, Joseph, and Saffman. *SIAM Journal on Applied Mathematics*, 60(4):1111–1127, 2000.

[135] F. Nataf. Absorbing boundary conditions and perfectly matched layers in wave propagation problems. In *Direct and Inverse problems in Wave Propagation and Applications*, volume 14 of *Radon Ser. Comput. Appl. Math.*, pages 219–231. de Gruyter, 2013.

[136] F. Nataf, F. Rogier, and E. De Sturler. Optimal interface conditions for domain decomposition methods. Technical report, École Polytechnique de Paris, 1994.

[137] Z. Peng and J. F. Lee. A scalable nonoverlapping and nonconformal domain decomposition method for solving time-harmonic Maxwell's equations in Rˆ3. *SIAM Journal on Scientific Computing*, 34(3):A1266–A1295, 2012.

[138] J. Przemieniecki. Matrix structural analysis of substructures. *AIAA Journal*, 1(1):138–147, 1963.

[139] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Numerical mathematics and scientific computation. Clarendon Press, 1999.

[140] B. Rummler. The eigenfunctions of the Stokes operator in special domains. II. *ZAMM - Journal of Applied Mathematics and Mechanics*, 77(9):669–675, 1997.

[141] P. G. Saffman. On the boundary condition at the surface of a porous medium. *Studies in Applied Mathematics*, 50(2):93–101, 1971.

[142] H. A. Schwarz. *Ueber einen Grenzübergang durch alternierendes Verfahren*. 1870.

[143] B. Smith, P. Bjorstad, and W. Gropp. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 2004.

[144] 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.*, 349(23):1255 – 1259, 2011.

[145] 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. *Numer. Math.*, 126(4):741–770, 2014.

[146] A. St-Cyr, M. J. Gander, and S. J. Thomas. Optimized multiplicative, additive, and restricted additive Schwarz preconditioning. *SIAM Journal on Scientific Computing*, 29(6):2402–2425, 2007.

[147] C. Stolk, A. Mostak, and K. B. Samir. A multigrid method for the Helmholtz equation with optimized coarse grid corrections. *SIAM J. Scientific Computing*, 36, 2014.

[148] K. Stüben. A review of algebraic multigrid. In *Numerical Analysis: Historical Developments in the 20th Century*, pages 331–359. Elsevier, 2001.

[149] L. Tartar. *An Introduction to Sobolev Spaces and Interpolation Spaces*. Lecture Notes of the Unione Matematica Italiana. Springer Berlin Heidelberg, 2007.

[150] P.L. Tchebychev. *Théorie des mécanismes connus sous le nom de parallélogrammes*. Imprimerie de l'Académie impériale des sciences, 1853.

[151] A. Toselli and O. Widlund. *Domain Decomposition Methods - Algorithms and Theory*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2004.

[152] U. Trottenberg, C. W. Oosterlee, and A. Schuller. *Multigrid*. Elsevier Science, 2001.

[153] R. S Varga. *Matrix Iterative Analysis*, volume 27. Springer Science & Business Media, 2009.

[154] E. L. Wachspress. Optimum alternating-direction-implicit iteration parameters for a model problem. *Journal of the Society for Industrial and Applied Mathematics*, 10(2):339–350, 1962.

[155] E. L. Wachspress. Extended application of alternating direction implicit iteration model problem theory. *Journal of the Society for Industrial and Applied Mathematics*, 11(4):994–1016, 1963.

[156] C. Wagner. Tangential frequency filtering decompositions for symmetric matrices. *Numerische Mathematik*, 78(1):119–142, Nov 1997.

[157] W. Weiler and G. Wittum. Parallel frequency filtering. *Computing*, 58(4):303–316, Dec 1997.

[158] S.-L. Wu. Optimized overlapping Schwarz waveform relaxation for a class of time-fractional diffusion problems. *Journal of Scientific Computing*, 72(2):842–862, 2017.

[159] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM review*, 34(4):581–613, 1992.

[160] X. Xu and L. Qin. Spectral analysis of Dirichlet–Neumann operators and optimized Schwarz methods with Robin transmission conditions. *SIAM journal on Numerical Analysis*, 47(6):4540–4568, 2010.

[161] Harry Yserentant. On the multi-level splitting of finite element spaces. *Numerische Mathematik*, 49(4):379–412, 1986.