Master thesis

Czech Technical University in Prague

F1

Faculty of Civil Engineering
Department of Mechanics

Adaptive coarse spaces in FETI-DP method for highly heterogeneous problems

Tomáš Medřický

Supervisor: Ing. Martin Doškář, Ph.D.
May 2024

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Medřický** Jméno: **Tomáš** Osobní číslo: **484634**

Fakulta/ústav: **Fakulta stavební**

Zadávající katedra/ústav: **Katedra mechaniky**

Studijní program: **Stavební inženýrství**

Studijní obor: **Konstrukce a dopravní stavby**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Adaptivní hrubé prostory pro řešení vysoce heterogenních úloh metodou FETI-DP

Název diplomové práce anglicky:

Adaptive coarse spaces in FETI-DP method for highly heterogeneous problems

Pokyny pro vypracování:

Cílem práce je studium, implementace a porovnání technik zvýšení robustnosti jedno- a dvouúrovňové metody FETI-DP pro řešení úloh s vysokým kontrastem v materiálových parametrech.
Vypracování by tak mělo zahrnovat:
- řešerši literatury z oblasti heuristických i spektrálních přístupů k obohacování hrubého prostoru FETI-DP
- implementaci těchto obohacení pomocí deflačních technik a/nebo přístupů založených na transformaci báze
- návrh výpočetně efektivnějších alternativ k těmto obohacujícím přístupům.

Seznam doporučené literatury:

[1] J. Mandel a B. Sousedík, „Adaptive selection of face coarse degrees of freedom in the BDDC and the FETI-DP iterative substructuring methods", Computer Methods in Applied Mechanics and Engineering, roč. 196, č. 8, s. 1389–1399, led. 2007
[2] A. Heinlein, A. Klawonn, M. Lanser, a J. Weber, „A frugal FETI-DP and BDDC coarse space for heterogeneous problems", etna, roč. 53, s. 562–591, 2020
[3] A. Toselli a O. B. Widlund, Domain Decomposition Methods — Algorithms and Theory, roč. 34. in Springer Series in Computational Mathematics, vol. 34. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Martin Doškář, Ph.D. katedra mechaniky FSv

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **26.02.2024** Termín odevzdání diplomové práce: **20.05.2024**

Platnost zadání diplomové práce: _____

Ing. Martin Doškář, Ph.D.	prof. Ing. Jiří Máca, CSc.	prof. Ing. Jiří Máca, CSc.
podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

. _____ Datum převzetí zadání

_____ Podpis studenta

Acknowledgements

At this point, I would like to express my heartfelt thanks to everyone who supported me throughout the time of writing this thesis.

First and foremost, I would like to thank my supervisor Martin Doškář for his time, patience, and dedication to students during seminars, which he conducts beyond his duties. Your gift for seeing simple explanations in every problem is truly inspiring. Trying to adopt this perspective has made the time spent studying domain decomposition methods a genuinely enjoyable experience.

Special thanks go to Dr. Alexander Heinlein for providing me the opportunity to spend several months on an internship at TU Delft. Your experience, mentorship, and willingness to share your knowledge have been incredibly inspiring, and I have grown a lot from it both professionally and personally. Not to mention, your lightning-fast email responses are something I should definitely learn from!

During my time in Delft, I met many inspiring people in a fantastic team. Above all, I am grateful to Dr. Artur Castiel Ries de Souza. Artur, you made sure I felt welcome, and your passion for work combined with your always cheerful attitude is something I still find motivating. I am grateful to everyone I had the chance to get to know there.

I would also like to extend my gratitude to Professor Jan Zeman for his invaluable support and the time spent on consultations, which greatly aided me in writing this thesis. The same applies to Dr. Ivana Pultarová for her consultations and the time she dedicated to me. I greatly appreciate it.

To my family: thank you for your endless support and patience. Your encouragement kept me going through all ups and downs, and I truly appreciate it.

And finally, to my wonderful girlfriend Veronika: thank you for your saintly patience and for always thinking of me. I'll never forget how we explored Rotterdam by bike, despite your dislike for them, every time you visited me. Your support and understanding made this journey so much more bearable and enjoyable.

Declaration

Herewith, I declare that the submitted thesis is my own work, written under the supervision of Martin Doškář and with additional consultations provided by Dr. Alexander Heinlein, Dr. Ivana Pultarová, and Prof. Jan Zeman. I also certify that all the resources and data used are cited and properly referenced.

This research has been supported by the Czech Science Foundation through project no. 24-11845S and the Grant Agency of the Czech Technical University in Prague, grant No. SGS24/038/OHK1/1T/11. In addition, the author acknowledges the support from project No. 19-26143X of the Czech Science Foundation for his summer research internships during his studies and co-funding of his research stay at TU Delft in 2023; both activities significantly contributed to the formulation of problems and development of computational strategies addressed in this thesis.

In Prague, 27. May 2024

Abstract

This thesis investigates approaches to improve performance of the Finite Element Tearing and Interconnecting Dual Primal (FETI-DP) method in problems with heterogeneously distributed high-contrast coefficients, in which the classical domain decomposition methods struggle. In particular, we focus on adaptive and heuristic approaches for coarse-space enhancement, and we investigate the robustness and stability of the standard strategies based on projections and transformation of basis for incorporating these enhancements. We propose two modifications of existing coarse-space enrichment techniques within FETI-DP, which result in better convergence and improved robustness while maintaining the complexity of the orignal techniques. As the main outcome of the thesis, we introduce two novel techniques: (i) a reduced-basis strategy for the generalized eigenvalue problems appearing in the adaptive approaches and (ii) a heuristic for selecting degrees of freedom to be added to the coarse space. Both techniques allow to identify the ill-posed solution modes while significantly reducing computation cost otherwise pertinent to eigenproblem-based adaptive approaches. The effectiveness of these modifications is illustrated on numerical test problems designed to expose limitations of traditional coarse-space constructions and further tested on systems of equations arising in modular-topology optimization tasks.

Keywords: FETI-DP, domain decomposition, heterogeneity, coarse space, adaptive techniques

Supervisor: Ing. Martin Doškář, Ph.D.

Abstrakt

Tato práce se zabývá způsoby pro urychlení konvergence a zlepšení robustnosti metody Finite Element Tearing and Interconnecting Dual Primal (FETI-DP) u problémů s heterogenně rozloženými koeficienty s vysokým kontrastem, u kterých klasické metody rozkladu oblasti selhávají. Práce je zaměřena zejména na adaptivní a heuristické přístupy pro obohacení hrubého prostoru této metody a detailně zkoumá robustnost a stabilitu standardních postupů založených na projekcích a transformaci báze pro zahrnutí těchto obohacujících podmínek do výpočtu. V rámci práce jsou navrženy dvě modifikace stávajících přístupů pro obohacení hrubého prostoru, které při zachování výpočetní náročnosti vedou k významnému zlepšení robustnosti a konvergence. Klíčovým výsledkem práce jsou pak dvě nově navržené techniky pro obohacení hrubého prostoru: (i) využití redukované báze pro výpočet problému zobecněných vlastních čísel a (ii) heuristické pravidlo pro výběr vhodných stupňů volnosti, které jsou přidány do hrubého prostoru. Tyto techniky jsou schopné identifikovat většinu módů způsobujících problémy s konvergencí a zároveň jsou výpočetně úspornější v porovnání s adaptivními technikami využívajícími plného řešení zobecněné problému vlastních čísel. Účinnost těchto modifikací a nově navržených strategií je demonstrována na testovacích úlohách navržených tak, aby odhalily omezení tradičních konstrukcí prostoru hrubých proměnných, a dále testována na úlohách modulární topologické optimalizace.

Klíčová slova: FETI-DP, doménová dekompozice, heterogenní úlohy, hrubý prostor, adaptivní techniky

Překlad názvu: Adaptivní hrubé prostory pro řešení vysoce heterogenních úloh metodou FETI-DP

Contents

Chapter 1
Introduction

Numerical simulations have become indispensable in both engineering design and research. As computational power continues to grow and becomes more cost-effective, the complexity of numerical models has increased significantly, often involving millions of unknowns. This advancement necessitates the development of efficient solvers capable of leveraging the parallel processing capabilities available in modern computers. Domain-decomposition strategies exemplify methods designed specifically for parallel computing.

The recently proposed modular topology optimization framework [56] has provoked a research question: which domain decomposition strategy, if any, is suitable for problems that feature predefined partitioning into subdomains and exhibit high contrast in coefficients within a domain. Recall that in the most common topology optimization approach [4], the whole available space is discretized, resulting in simulations that model almost empty space with a very low-coefficient material model.

Our preliminary results presented in [37] and the author's bachelor's thesis [36, in Czech] indicated that these problems pose significant challenges for classical domain decomposition methods, underscoring the need for more advanced strategies. Among the investigated methods, the Finite Element Tearing and Interconnecting Dual Primal (FETI-DP) method [11] proved to be the most robust, although not yet entirely adequate. The method's ability to stabilize the solution with chosen degrees of freedom that are not solved subdomain-wise paved the way for our quest for an enrichment of this coarse space of unknowns, yielding an enhanced robustness of the solver. This thesis addresses this open research question by introducing two modifications of existing approaches and two novel approaches to coarse-space enhancement (a reduced-basis generalized eigenvalue problem and an adaptive heuristic).

This thesis begins with an overview of the Finite Element Tearing and Interconnecting Dual Primal (FETI-DP) method, a well-established technique that has seen significant advancements in both theory and implementation over the past two decades. Key enhancements from the literature are reviewed, with a focus on their applicability to problems with highly heterogeneous distributions of high-contrast coefficients. These enhancements include (i) scaling of the binding constraints across subdomains' interfaces and (ii) ways of enforcing additional coarse-space constraints within the FETI-DP framework (with particular emphasis on the comparison of projector preconditioning and transformation of basis).

The next part of the thesis is dedicated to the identification and construction of appropriate *coarse spaces*. Despite theoretical proofs on condition number bounds being established for some adaptive techniques, which typically involve identifying solution modes harmful to the method's convergence using generalized eigenvalue problems restricted to subdomains' interfaces, it is often not straightforward to determine the most computationally efficient approach. Starting with weighted averages, the focus then moves to adaptive approaches: from the original approach by Mandel and Sousedík [35] based on the generalized eigenvalue problem to its frugal heuristic counterpart [18] that mimics the original adaptive method. Here, motivated by ρ scaling, we propose the first modification which improves the performance of the original Frugal method. We then proceed with the introduction of reduced-dimensional strategy for obtaining adaptive constraints, with the reduced basis being constructed heuristically,

following our observations of method's behavior on many elliptic problems. Finally, we return to our original starting point, and we try to develop improved heuristics for the selection of nodal degrees of freedom to enrich the coarse space, given the experience we gained while working on the aforementioned strategies. Throughout this section, we illustrate the impact of individual modifications on scalar and elasticity elliptic problems devised to reveal drawbacks of practical coarse-space constructions. In the last section of the thesis, we provide numerical tests of performance of the above-mentioned enrichments and their enforcement in the topology optimization problems.

Throughout the thesis, we focus solely on linear problems described by elliptic partial differential equations. In particular, we consider only the scalar problem of steady-state heat conduction and the problem of linear statics. In both cases, only isotropic material models were assumed. All discrete systems of linear equations considered in this work were obtained using the finite element method with the standard conforming linear triangular and bilinear quadrilateral elements.

Chapter 2

FETI Dual-Primal

Finite **E**lement **T**earing and **I**nterconnecting **D**ual **P**rimal (henceforth, FETI-DP) method, first proposed by Farhat et al. [11], is a modified version of the original FETI [14]. Along with its primal counterpart, the **B**oundary **D**omain **D**ecomposition by **C**onstraints (BDDC) [7, 33], which features a similar numerical performance with a potential difference in multiplicities of 0 and 1 eigenvalues in the spectra [6, 34], it is regarded as one of the most effective numerically scalable nonoverlapping domain decomposition methods available [50]. In the early stages, the method was primarily developed for discretised second- and fourth-order elliptic PDE, aiming at the needs of the structural mechanics community. FETI-DP and BDDC methods were later successfully adopted in various problems, such as crack propagation [2] and on isogeometric analysis of compressible [40] and almost incompressible [58] linear elasticity problems, to name a few.

The FETI-DP emerged as a way to improve the original one-level FETI method (also known as FETI-1), which at the time of its inception lacked a sufficiently robust coarse space, a necessary factor for a numerical scalability of the method. The key difference between FETI-DP and FETI lies in the fact that FETI-DP directly preserves solution continuity in a set of few selected primal variables, which are coupled at the global level by partial assembly. Consequently, it represents a compromise between the originally dual nature of the FETI method and primal character of methods such as Schur Complements, as the substructures are no longer completely decoupled. This coupling in primal variables, intentionally chosen such that all the local subproblems remain invertible, serves as a coarse grid, which acts as a natural coarse space traditionally used in dual methods, catching contributions of kernels of the local stiffness matrices on the floating subdomains. Thus, at the cost of constructing a small global problem, we can avoid pseudo-inverses of local stiffness matrices and, at the same time, bypass the use of projections in a conjugate gradient solver.

Driven by rapid developments in many fields including numerical modelling and scientific parallel computing together with a consistent increase in accessibility of computational resources, the method and its applicability has been thoroughly investigated: from the theoretical standpoint, which included addressing provable upper limits on effective condition number [28, 32, 55] and, hand in hand, the convergence rate, and from the practical implementation perspective, where the scalability up to hundreds of thousands of computational cores has already been confirmed on parallel supercomputers [1].

The method itself has been evolving over time, taking inspiration in the development in other domain decomposition (DD) methods. In fact, the evolution of many DD methods has been closely intertwined. Nowadays, the classical DD methods are almost always combined with certain enhancements not inherent to their original variants. These enhancements, e.g. block conjugate gradients [5, 16, 38], are incorporated to overcome certain limitations of the particular method by making use of advantageous techniques from various areas, effectively combining their strengths to build a robust method reflecting specifics of the problem under consideration. Another example can be a recursive application of the method itself, e.g. for a solution of large coarse problems, giving rise to multilevel methods [23, 50, 54].

The last step towards perfection in the evolution is to let the algorithm itself identify the

harmful part of the problem being solved, i.e. allowing the solver to automatically identify the bad modes in the system responsible for slow convergence and to concentrate computational power there. Ideally, such an algorithm would enable a user to effortlessly, i.e. without any required in-depth knowledge or personal experience with a problem at hand, find a balance between optimal expected rate of convergence and the low dimensionality of the coarse problem, which is typically directly factorised. This desired behavior, which beneficially exploits the robustness of direct solvers and the ease of parallelization of iterative solvers, is addressed with the so-called adaptive techniques; see the work of Spillane et al. [51, 52] for strictly dual FETI and e.g. [35, 43] for FETI-DP in the context of non-overlapping DD methods.
The most sophisticated of the adaptive techniques give a user a control over the condition number bound with a single user-defined threshold, thus making the solver easy to use as a black-box. Moreover, the value of this threshold is typically of the same magnitude as the obtained condition number, hence it avoids the need for tedious parameter tuning. The only drawback of adaptive techniques is that they usually rely on a solution of local generalized eigenvalue problems (GEVP). By local we mean that each GEVP is restricted to a single subdomain (for a GenEO type [51, 52]) or an edge/face shared by a pair of adjacent subdomains, which is typical for the dual-primal FETI variant. While this locality of GEVPs enables parallel handling, the data-transfers, set-up, and solution of GEVPs still pose a computationally intensive part of the algorithm, particularly if adaptive constraints eventually on a very limited number of interfaces are required.
On the other side of complexity of adaptivity stand heuristic approaches. These approaches are commonly based upon geometric or physical expectations, reflecting certain observations of a problem behavior. To the best of author's knowledge, none of the approaches classified as heuristic is provably robust for arbitrary material distribution and contrast. Therefore, it is evident that these approaches have limitations in their applicability. Yet, for many realistic problems, these heuristics are a perfect choice: they are cheap to set up compared to adaptive approaches and they provide a reliable approximation of the coarse modes necessary for restoring solver's robustness. However, in some synthetic as well as real-world applications (particularly those with continuously varying coefficients and high contrast in material properties), these heuristics fail to deliver desired performance.

In the following chapter, we briefly overview the most widely used heuristics and adaptive approaches and comment on their limitations. In addition, we propose subtle modifications to selected heuristics that will, in certain cases, ensure better numerical behavior. Finally, we develop a reduced-dimensional approach for solution of localized GEVPs.

2.1 Decomposition of the original problem

Let us consider a given polygonal domain $\Omega \subset \mathbb{R}^d$ with d denoting the dimension of the problem. Typically $d \in \{2, 3\}$. In this thesis, however, we will limit our attention to two-dimensional problems. We decompose Ω into $N_\mathrm{s} \in \mathbb{N}$ non-overlapping subdomains $\Omega^{(i)}$, i.e.

$$\overline{\Omega} = \bigcup_{i=1}^{N_\mathrm{s}} \overline{\Omega}^{(i)} \qquad \text{with} \qquad \Omega^{(i)} \cap \Omega^{(j)} = \emptyset \quad i \neq j \qquad \forall\, i, j \in 1, \ldots, N_\mathrm{s} \tag{2.1}$$

with the global interface boundary Γ which is – inspired by the the method of Schur complements – occasionally referred to as the *skeleton* of the decomposition,

$$\Gamma = \bigcup_{i=1}^{N_\mathrm{s}} \Gamma^{(i)} \qquad \text{where} \qquad \Gamma^{(i)} = \partial\Omega^{(i)} \setminus \partial\Omega. \tag{2.2}$$

Moreover, we let $\Gamma^{(ij)}$ denote a part of Γ shared by subdomains i and j, i.e. $\Gamma^{(ij)} = \Gamma^{(i)} \cap \Gamma^{(j)}$.

2.2 Description of geometry

For description of geometry, we follow the notation and definition introduced in [29] for three-dimensional problems. First, for any nodal point x of the discretization of $\overline{\Omega}_h$ we let $\mathcal{N}_x$ denote the set of parent substructures' indices of x

$$\mathcal{N}_x := \left\{k \in \{1,\dots,N_\mathrm{s}\} : x \in \overline{\Omega}_h^{(k)}\right\} \tag{2.3}$$

and, according to [29], define equivalence relations $\sim$ for interface nodal points

$$x \sim y \iff \mathcal{N}_x = \mathcal{N}_y \quad \text{s.t.} \quad y \in \mathcal{C}_\mathrm{con}(x) \ \wedge \ x, y \in \Gamma_h$$

and

$$z \sim x \sim y \iff \mathcal{N}_z = \mathcal{N}_x = \mathcal{N}_y \quad \text{s.t.} \quad y, z \in \mathcal{C}_\mathrm{con}(x) \ \wedge \ y \neq z \ \wedge \ x, y, z \in \Gamma_h \ ,$$

where $\mathcal{C}_\mathrm{con}(x)$ signifies the index set of neighboring nodal points of x within the connectivity graph. Then, since we restrict ourselves to 2D problems only, we distinguish the following three **equivalence classes**, with each nodal point $x \in \overline{\Omega}_h$ belonging to exactly one of these groups:

- **Interiors**: $\mathcal{I}$ denote the set of all nodal points that lie in the union of interiors of substructures, i.e.

  $$x \in \mathcal{I} \iff |\mathcal{N}_x| = 1.$$

- **Edges**: $\mathcal{E}$ denotes the set of all nodal points that lie on the interface edges

  $$x \in \mathcal{E} \iff |\mathcal{N}_x| = 2 \wedge \exists\, y, z : z \sim x \sim y$$

  Individual edges E_{ij} are defined as open sets in Γ

  $$E_{ij} := \left\{\bigcup x \in \mathcal{E} : \mathcal{N}_x = \{i, j\} \ \wedge \ \forall y \in E_{ij}, \ \exists z \in E_{ij} : y \in \mathcal{C}_\mathrm{con}(z)\right\} \tag{2.4}$$

  Therefore, we omit the case when two subdomains share two individual edges

- **Vertices**: $\mathcal{V}$ denotes the set of all nodal points that lie on the vertices defined as endpoints of all $\Gamma^{(ij)}$

  $$x \in \mathcal{V} \iff |\mathcal{N}_x| \geq 2 \ \wedge \ \nexists y, z : z \sim x \sim y$$

2.3 Original FETI-DP formulation

In the following section, we will establish most of the necessary notation and briefly outline the standard FETI-DP method in a rather algebraic framework. The notation will be strongly adopted from the work of Klawonn and Widlund [29] and their collaborators.

Let's suppose we wish to solve the system of linear equations

$$\mathsf{K}_\Omega \mathsf{u}_\Omega = \mathsf{f}_\Omega \, , \tag{2.5}$$

which arises from numerically discretized elliptic partial differential equations on a polygonal domain Ω. Using decomposition into N_s non-overlapping subdomains accordingly to Section 2.1, problem (2.5) can be posed as a series of completely decoupled locally defined subdomain-wise problems

$$\mathsf{K}^{(s)} \mathsf{u}^{(s)} = \mathsf{f}^{(s)} \tag{2.6}$$

subjected to additional constraint ensuring continuity of $\{\mathsf{u}^{(s)}\}_{s=1\dots N_\mathrm{s}}$ across subdomain boundaries.

Based on the character of subdomain degrees of freedom (DOFs), we recognise :

1. **Interior DOFs** I, which correspond to nodal points $x \in \mathcal{I}$. These DOFs in $\overline{\Omega} \setminus \Gamma$ are usually condensed out first. Despite its name, nodes on $\partial\Omega_N \setminus \Gamma$ are contained in this set as well.

2. **Interface DOFs** Γ, which can be further attributed to one of the following groups:

   a. **Primal DOFs** Π are coupled at the global level through assembly operators. Each subdomain $\Omega^{(s)}$ should have a sufficient number of primal DOFs $\Pi^{(s)}$ to yield the local system (2.6) invertible. The choice of DOFs in Π determines the resulting a priori coarse space. In most cases, the endpoints of all nonempty parts $\Gamma^{(ij)}$ are adopted. Even though different options are availabe, we will exclusively start with a Π containing all vertices as outlined in Section 2.2, which generally does not lead to the smallest number of primal DOFs needed to retain invertibility of local subproblems, but brings many benefits for adaptive approaches, i.e. satisfies [35, Sec. 4, Assumption 8]. Slight variations in the definitions of vertex constraints, such as excluding subdomain vertices with a multiplicity less than 3 in three-dimensional cases, are also feasible.

   b. **Dual DOFs** Δ are located at Γ and not contained in Π. Continuity in the dual DOFs is enforced in an iterative manner by imposed continuity constraints. The term *dual* emphasizes that, due to the applied principle of Lagrange multipliers, the quantity enforcing the continuity in these DOFs is of the dual character, i.e., for a solution in displacements, we iterate in forces and vice versa.

Assuming that ordering of degrees of freedom follows the above-mentioned classification of variables, we arrive at the following partitioning of the matrix and vectors of the local problem (2.6)

$$\mathsf{K}^{(s)} = \begin{bmatrix} \mathsf{K}^{(s)}_{\mathrm{II}} & \mathsf{K}^{(s)}_{\mathrm{I}\Delta} & \mathsf{K}^{(s)}_{\mathrm{I}\Pi} \\ \mathsf{K}^{(s)}_{\Delta\mathrm{I}} & \mathsf{K}^{(s)}_{\Delta\Delta} & \mathsf{K}^{(s)}_{\Delta\Pi} \\ \mathsf{K}^{(s)}_{\Pi\mathrm{I}} & \mathsf{K}^{(s)}_{\Pi\Delta} & \mathsf{K}^{(s)}_{\Pi\Pi} \end{bmatrix}, \qquad \mathsf{u}^{(s)} = \begin{bmatrix} \mathsf{u}^{(s)}_{\mathrm{I}} \\ \mathsf{u}^{(s)}_{\Delta} \\ \mathsf{u}^{(s)}_{\Pi} \end{bmatrix}, \qquad \text{and} \qquad \mathsf{f}^{(s)} = \begin{bmatrix} \mathsf{f}^{(s)}_{\mathrm{I}} \\ \mathsf{f}^{(s)}_{\Delta} \\ \mathsf{f}^{(s)}_{\Pi} \end{bmatrix},$$

operating on a local finite element space commonly denoted by W_s, i.e. $\mathsf{u}^{(s)}, \mathsf{f}^{(s)} \in W_s$ and $\mathsf{K}^{(s)} \in W_s \to W_s$ [57]. To simplify the notation, this splitting can be further restated by grouping into

$$\mathsf{K}^{(s)}_{\mathrm{RR}} = \begin{bmatrix} \mathsf{K}^{(s)}_{\mathrm{II}} & \mathsf{K}^{(s)}_{\mathrm{I}\Delta} \\ \mathsf{K}^{(s)}_{\Delta\mathrm{I}} & \mathsf{K}^{(s)}_{\Delta\Delta} \end{bmatrix} \quad \mathsf{K}^{(s)}_{\mathrm{R}\Pi} = \begin{bmatrix} \mathsf{K}^{(s)}_{\mathrm{I}\,\Pi} \\ \mathsf{K}^{(s)}_{\Delta\Pi} \end{bmatrix} \quad \mathsf{u}^{(s)}_{\mathrm{R}} = \begin{bmatrix} \mathsf{u}^{(s)}_{\mathrm{I}} \\ \mathsf{u}^{(s)}_{\Delta} \end{bmatrix} \quad \text{and} \quad \mathsf{f}^{(s)}_{\mathrm{R}} = \begin{bmatrix} \mathsf{f}^{(s)}_{\mathrm{I}} \\ \mathsf{f}^{(s)}_{\Delta} \end{bmatrix}, \tag{2.7}$$

$$\mathsf{K}^{(s)}_{\Gamma\Gamma} = \begin{bmatrix} \mathsf{K}^{(s)}_{\Delta\Delta} & \mathsf{K}^{(s)}_{\Delta\Pi} \\ \mathsf{K}^{(s)}_{\Pi\Delta} & \mathsf{K}^{(s)}_{\Pi\Pi} \end{bmatrix} \quad \mathsf{K}^{(s)}_{\Gamma\mathrm{I}} = \begin{bmatrix} \mathsf{K}^{(s)}_{\Delta\mathrm{I}} \\ \mathsf{K}^{(s)}_{\Pi\mathrm{I}} \end{bmatrix} \quad \mathsf{u}^{(s)}_{\Gamma} = \begin{bmatrix} \mathsf{u}^{(s)}_{\Delta} \\ \mathsf{u}^{(s)}_{\Pi} \end{bmatrix} \quad \text{and} \quad \mathsf{f}^{(s)}_{\Gamma} = \begin{bmatrix} \mathsf{f}^{(s)}_{\Delta} \\ \mathsf{f}^{(s)}_{\Pi} \end{bmatrix}. \tag{2.8}$$

The R symbol unifies the remaining DOFs that are not part of the primal set Π, i.e. $\mathrm{R} = \mathrm{I} \cap \Delta$, whereas subscript Γ specifically refers to those DOFs associated with the physical boundary Γ, grouping both dual and primal DOFs together. Apart from individual completely local terms, we introduce block-diagonal matrices

$$\mathsf{K}_{\mathrm{II}} = \operatorname{diag}_{i=1}^{N_\mathrm{s}} \mathsf{K}^{(i)}_{\mathrm{II}} \qquad \mathsf{K}_{\Delta\Delta} = \operatorname{diag}_{i=1}^{N_\mathrm{s}} \mathsf{K}^{(i)}_{\Delta\Delta}$$
$$\mathsf{K}_{\Pi\,\Pi} = \operatorname{diag}_{i=1}^{N_\mathrm{s}} \mathsf{K}^{(i)}_{\Pi\,\Pi} \qquad \text{and} \qquad \mathsf{K}_{\mathrm{RR}} = \operatorname{diag}_{i=1}^{N_\mathrm{s}} \mathsf{K}^{(i)}_{\mathrm{RR}}$$

and vectors

$$\mathsf{u}_{\mathrm{I}} = \begin{bmatrix} \mathsf{u}_{\mathrm{I}}{}^{(1)} \\ \mathsf{u}_{\mathrm{I}}{}^{(2)} \\ \vdots \\ \mathsf{u}_{\mathrm{I}}{}^{(N_\mathrm{s})} \end{bmatrix} \quad \mathsf{f}_{\mathrm{I}} = \begin{bmatrix} \mathsf{f}_{\mathrm{I}}{}^{(1)} \\ \mathsf{f}_{\mathrm{I}}{}^{(2)} \\ \vdots \\ \mathsf{f}_{\mathrm{I}}{}^{(N_\mathrm{s})} \end{bmatrix} \quad \mathsf{u}_{\Delta} = \begin{bmatrix} \mathsf{u}^{(1)}_{\Delta} \\ \mathsf{u}^{(2)}_{\Delta} \\ \vdots \\ \mathsf{u}^{(N_\mathrm{s})}_{\Delta} \end{bmatrix} \quad \mathsf{f}_{\Delta} = \begin{bmatrix} \mathsf{f}^{(1)}_{\Delta} \\ \mathsf{f}^{(2)}_{\Delta} \\ \vdots \\ \mathsf{f}^{(N_\mathrm{s})}_{\Delta} \end{bmatrix} \quad \mathsf{u}_{\Pi} = \begin{bmatrix} \mathsf{u}^{(1)}_{\Pi} \\ \mathsf{u}^{(2)}_{\Pi} \\ \vdots \\ \mathsf{u}^{(N_\mathrm{s})}_{\Pi} \end{bmatrix} \quad \mathsf{f}_{\Pi} = \begin{bmatrix} \mathsf{f}^{(1)}_{\Pi} \\ \mathsf{f}^{(2)}_{\Pi} \\ \vdots \\ \mathsf{f}^{(N_\mathrm{s})}_{\Pi} \end{bmatrix}.$$

The diag operation refers to the construction of a block diagonal matrix from a given set of matrices.
As already mentioned, FETI-DP keeps the set of primal variables Π partially assembled. For this assembly, simple restriction/prolongation operators are defined. Specifically, we have a matrix $\mathsf{R}^{(s),\mathsf{T}}$ with values $\{0,1\}$ that prolongates an element from $\Pi^{(s)}$ to the appropriate position in the assembled global representation of elements of Π. The role of $\mathsf{R}^{(s)}$ is then completely opposite; it restricts an appropriate element from Π to subdomain-specific $\Pi^{(s)}$. Consequently, the primally assembled terms, denoted as $\widetilde{\bullet}$, read

$$\widetilde{\mathsf{K}}_{\Pi\Pi} = \sum_{i=1}^{N_\mathrm{s}} \mathsf{R}_\Pi^{(i)\mathsf{T}} \mathsf{K}_{\Pi\Pi}^{(i)} \mathsf{R}_\Pi^{(i)} = \mathsf{R}_\Pi^\mathsf{T} \mathsf{K}_{\Pi\Pi} \mathsf{R}_\Pi \qquad \widetilde{\mathsf{f}}_\Pi = \sum_{i=1}^{N_\mathrm{s}} \mathsf{R}_\Pi^{(i)\mathsf{T}} \mathsf{f}_\Pi^{(i)}, \tag{2.9}$$

$$\widetilde{\mathsf{K}}_{\Pi\mathrm{R}} = \begin{bmatrix} \mathsf{R}_\Pi^{(1)\mathsf{T}} \mathsf{K}_{\Pi\mathrm{R}}^{(1)} & \dots & \mathsf{R}_\Pi^{(N_\mathrm{s})\mathsf{T}} \mathsf{K}_{\Pi\mathrm{R}}^{(N_\mathrm{s})} \end{bmatrix} = \widetilde{\mathsf{K}}_{\mathrm{R}\Pi}^\mathsf{T} \quad \text{where} \quad \mathsf{R}_\Pi^\mathsf{T} = \begin{bmatrix} \mathsf{R}_\Pi^{(1),\mathsf{T}} & \mathsf{R}_\Pi^{(2),\mathsf{T}} & \dots & \mathsf{R}_\Pi^{(N_\mathrm{s}),\mathsf{T}} \end{bmatrix}$$

and values stored in $\mathsf{u}_\Pi = \mathsf{R}_\Pi \widetilde{\mathsf{u}}_\Pi$ coincide at appropriate positions by explicit subassembly.

For dual variables Δ, we follow the steps of the other FETI-based methods and use a jump operator B. In the past, it was experimented with various ways of introducing continuity constraints; a brief overview on possible forms of jump operator and their influence on the solver has already been given in the bachelor thesis of the author [36]. In two-dimensional problems with a vertex-based coarse space, there is generally no need to incorporate redundant Lagrange multipliers because the maximum number of substructures sharing a dual degree of freedom is deliberately limited to two. However, sometimes it might be beneficial to introduce a few redundant constraints, for instance, when the transformation of basis is adopted. Currently, a fully redundant set of Lagrange multipliers enjoys privileged status among all options, particularly when combined with advanced techniques such as adaptive approaches. This is due to the fact that the full set of Lagrange multipliers efficiently addresses the necessity of introducing a scaling. Moreover, the consistency of scaling is easily ensured with the redundant set of Lagrange multipliers, which streamlines the implementation process. We wish to emphasize that the redundancy does not pose a problem; the condensed system remains positive definite on Range(B), thus the solution vector is uniquely determined. The Lagrange multipliers are, on the other hand, uniquely determined only up to an element in Kernel(B^T). Consequently, we construct matrix B using values $\{-1,0,1\}$ with each row consisting exactly one $+1$ and one -1 value. All the remaining continuity conditions across Γ not handled by Π are now resolved by

$$\mathsf{B}_\Delta \mathsf{u}_\Delta = \mathsf{0}\,. \tag{2.10}$$

The matrix B_Δ is known as a jump operator, and it is again assembled from domain-wise contributions. In what follows, we will often deal with the restriction of all quantities to a certain part of the interface, and thus we will frequently refer to these domain-wise contributions. In an effort to unify the nomenclature used in this thesis, we slightly deviate from the commonly used definition of the local contributions of the jump operator. In particular, we consider that all local $\mathsf{B}_\Delta^{(s)}$ comprise only non-zero rows, i.e., each $\mathsf{B}_\Delta^{(s)}$ contains as many rows as there are conditions defined on substructure $\Omega^{(s)}$. For a formal mapping to the global set of constraints in Eq. (2.10), we introduce auxiliary prolongation operators $\mathsf{R}_{\mathsf{B}_\Delta}^{(s)} \in \mathbb{R}^{n_\Delta \times n_\Delta^{(s)}}$, with n_Δ and $n_\Delta^{(s)}$ being the total and subdomain-wise number of Lagrange multipliers, determined by the number of nontrivial rows of B_Δ and $\mathsf{B}_\Delta^{(s)}$, respectively.

Matrix $\mathsf{R}_\mathsf{B}^{(s)}$ maps rows in $\mathsf{B}_\Delta^{(s)}$ to their appropriate positions in B_Δ,

$$\mathsf{B}_\Delta = \begin{bmatrix} \mathsf{R}_\mathsf{B}^{(1)} \mathsf{B}_\Delta^{(1)} & \dots & \mathsf{R}_\mathsf{B}^{(N_\mathrm{s})} \mathsf{B}_\Delta^{(N_\mathrm{s})} \end{bmatrix}. \tag{2.11}$$

To complement the partitioning introduced in (2.7-2.8), we define

$$\mathsf{B}_\mathrm{R}^{(s)} = \begin{bmatrix} \mathsf{0}_\mathrm{I}^{(s)} & \mathsf{B}_\Delta^{(s)} \end{bmatrix} \qquad \mathsf{B}_\Gamma^{(s)} = \begin{bmatrix} \mathsf{B}_\Delta^{(s)} & \mathsf{0}_\Pi^{(s)} \end{bmatrix} \qquad \mathsf{B}^{(s)} = \begin{bmatrix} \mathsf{B}_\mathrm{R}^{(s)} & \mathsf{0}_\Pi^{(s)} \end{bmatrix}.$$

Note that despite the nodal, vertex-based set Π being usually chosen such that all of the subdomains are sufficiently supported in case of mechanics, or connected in general, to avoid local stiffness matrices being indefinite, this choice is not strictly necessary for all subdomains

when the inter-domain conditions are held in a different manner, e.g. by enforcing local edge averages or some different auxiliary conditions through projector preconditioning, deflation, a transformation of basis, or local assembly. The only permanent condition is that the domain as a whole must fulfill the sufficient Dirichlet boundary conditions. These different, non-nodal conditions can often speed up the convergence dramatically. However, it is difficult to predict the optimal form of such coarse continuity conditions in a generically applicable manner.

The FETI-DP master system is given by

$$\begin{bmatrix} \mathsf{K}_{\mathrm{RR}} & \tilde{\mathsf{K}}_{\Pi\mathrm{R}}^{\mathsf{T}} & \mathsf{B}_{\mathrm{R}}^{\mathsf{T}} \\ \tilde{\mathsf{K}}_{\Pi\mathrm{R}} & \tilde{\mathsf{K}}_{\Pi\Pi} & 0 \\ \mathsf{B}_{\mathrm{R}} & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathsf{u}_{\mathrm{R}} \\ \tilde{\mathsf{u}}_{\Pi} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathsf{f}_{\mathrm{R}} \\ \tilde{\mathsf{f}}_{\Pi} \\ 0 \end{bmatrix} \tag{2.12}$$

Giving rise to an operator R providing assembly in primal constraints

$$\mathsf{R}^{\mathsf{T}} = \begin{bmatrix} \mathsf{R}^{(1),\mathsf{T}} & \mathsf{R}^{(2),\mathsf{T}} & \dots & \mathsf{R}^{(N_{\mathrm{s}}),\mathsf{T}} \end{bmatrix}$$

we could alternatively write

$$\begin{bmatrix} \widetilde{\mathsf{K}} & \mathsf{B}^{\mathsf{T}} \\ \mathsf{B} & 0 \end{bmatrix} \begin{bmatrix} \widetilde{\mathsf{u}} \\ \lambda \end{bmatrix} = \begin{bmatrix} \widetilde{\mathsf{f}} \\ 0 \end{bmatrix} \tag{2.13}$$

with

$$\widetilde{\mathsf{K}} = \mathsf{R}^{\mathsf{T}}\mathsf{K}\mathsf{R} \qquad \mathsf{K} = \mathrm{diag}\left(\mathsf{K}^{(1)}, \mathsf{K}^{(2)}, \dots, \mathsf{K}^{(N_{\mathrm{s}})}\right)$$

$$\widetilde{\mathsf{u}} = \mathsf{R}^{\mathsf{T}}\mathsf{u} \qquad \mathsf{u}^{\mathsf{T}} = \begin{bmatrix} \mathsf{u}^{(1),\mathsf{T}} & \mathsf{u}^{(2),\mathsf{T}} & \dots & \mathsf{u}^{(N_{\mathrm{s}}),\mathsf{T}} \end{bmatrix}$$

$$\widetilde{\mathsf{f}} = \mathsf{R}^{\mathsf{T}}\mathsf{f} \qquad \mathsf{f}^{\mathsf{T}} = \begin{bmatrix} \mathsf{f}^{(1),\mathsf{T}} & \mathsf{f}^{(2),\mathsf{T}} & \dots & \mathsf{f}^{(N_{\mathrm{s}}),\mathsf{T}} \end{bmatrix}$$

The specific equivalence between $\widetilde{\mathsf{K}}$ and the leading 2×2 block in Eq. (2.12) holds for an interpolation operator [55]

$$\mathsf{R} := \begin{bmatrix} \mathsf{I}_{\mathrm{R}} & 0 \\ 0 & \mathsf{R}_{\Pi} \end{bmatrix}$$

with local contributions

$$\mathsf{R}^{(s),\mathsf{T}} := \begin{bmatrix} \widehat{\mathsf{I}_{\mathrm{R}}^{(s)}} & 0 \\ 0 & \mathsf{R}_{\Pi}^{(s),\mathsf{T}} \end{bmatrix} \quad : W_s \to \widetilde{W}$$

where $\widehat{\mathsf{I}_{\mathrm{R}}^{(s)}} \in \mathbb{R}^{n_{\mathrm{R}} \times n_{\mathrm{R}}^{(s)}}$ is a matrix containing a block of possibly permuted $n_{\mathrm{R}}^{(s)} \times n_{\mathrm{R}}^{(s)}$ identity on appropriate positions and zeros elsewhere. In this case, $n_{\mathrm{R}}^{(s)}$ denotes the number of DOFs in the set R on a subdomain $\Omega^{(s)}$, and n_{R} is a total number of DOFs in this set. The space $\widetilde{W} \subset W := W_i \times \cdots \times W_{N_{\mathrm{s}}}$ is the space of functions continuous in primal variables defined by [29]

$$\widetilde{W} := \left\{ \mathsf{u} : \exists \mathsf{u}^{(s)} \in W^{(s)}, s = 1, \dots, N_{\mathrm{s}} \quad \text{s.t. } \mathsf{u} = \sum_{i=1}^{N_{\mathrm{s}}} \mathsf{R}^{(i),\mathsf{T}} \mathsf{u}^{(i)} \right\}. \tag{2.14}$$

Following the block Schur complement procedure and eliminating the system on the dual variables only, we come to a condensed system of the form

$$\mathsf{F}\lambda = \mathsf{d} \tag{2.15}$$

with left hand side operator $\mathsf{F} \succeq 0$ defined as

$$\mathsf{F} = \mathsf{B}_{\mathrm{R}}\mathsf{K}_{\mathrm{RR}}^{-1}\mathsf{B}_{\mathrm{R}}^{\mathsf{T}} + \mathsf{B}_{\mathrm{R}}\mathsf{K}_{\mathrm{RR}}^{-1}\tilde{\mathsf{K}}_{\Pi\mathrm{R}}^{\mathsf{T}}\tilde{\mathsf{S}}_{\Pi\Pi}^{-1}\tilde{\mathsf{K}}_{\Pi\mathrm{R}}\mathsf{K}_{\mathrm{RR}}^{-1}\mathsf{B}_{\mathrm{R}}^{\mathsf{T}} \tag{2.16}$$

or, for the purpose of keeping the notation high level, we alternatively present

$$\mathsf{F} = \mathsf{B}\widetilde{\mathsf{K}}^{-1}\mathsf{B}^{\mathsf{T}} \tag{2.17}$$

or

$$\mathsf{F} = \mathsf{B}_{\Gamma}\widetilde{\mathsf{S}}_{\Gamma}^{-1}\mathsf{B}_{\Gamma}^{\mathsf{T}}. \tag{2.18}$$

with $\widetilde{\mathsf{S}}_{\Gamma}$ being the Schur complement of the block of degrees of freedom in Γ of assembled matrix $\widetilde{\mathsf{K}}$. For later purposes, we also denote by S_{Γ} the non-assembled version of $\widetilde{\mathsf{S}}_{\Gamma}$.

The right-hand side reads

$$\mathsf{d} = \mathsf{B}_\mathrm{R}\mathsf{K}_\mathrm{RR}^{-1}\mathsf{f}_\mathrm{R} - \mathsf{B}_\mathrm{R}\mathsf{K}_\mathrm{RR}^{-1}\tilde{\mathsf{K}}_\mathrm{\Pi R}^\mathsf{T}\tilde{\mathsf{S}}_\mathrm{\Pi\Pi}^{-1}(\tilde{\mathsf{f}}_\Pi - \mathsf{K}_\mathrm{\Pi R}\mathsf{K}_\mathrm{RR}^{-1}\mathsf{f}_\mathrm{R})\,. \tag{2.19}$$

The application of F to a vector is practically performed by a solution of two systems in the sense of Eq. (2.16), where the beneficial structure is exploited with a wise use of a matrix inversion lemma. Clearly, the first term in the expression of F comprises a block-diagonal matrix, thereby is trivially parallelizable. This distribution into smaller parts, necessitating no inter-communication during computation, represents the essential factor in achieving numerical scalability of the method. The second term in F constitutes a slightly more intricate part. The new entity there, appearing in the Eq. (2.19) as well, is the coupled Schur complement assembled in a few selected primal variables, also known as (*prior*) coarse problem:

$$\tilde{\mathsf{S}}_\mathrm{\Pi\Pi} := \tilde{\mathsf{K}}_\mathrm{\Pi\Pi} - \tilde{\mathsf{K}}_\mathrm{\Pi R}\mathsf{K}_\mathrm{RR}^{-1}\tilde{\mathsf{K}}_\mathrm{R\Pi}. \tag{2.20}$$

The coarse problem, in its simplest variant, substitutes the role of a natural coarse space in original FETI, a term first used in [13].

Conversely to the first term of F, the coarse problem is critical to the robustness of the method. Recall that $\tilde{\mathsf{S}}_\mathrm{\Pi\Pi}$ operates at the global level and that its application directly mediates the global exchange of information. Since then, we can formulate two fundamental requirements for $\tilde{\mathsf{K}}_\mathrm{\Pi\Pi}$, or $\tilde{\mathsf{S}}_\mathrm{\Pi\Pi}$, respectively. Firstly, it is preferable to keep it small in dimension to reduce the computational complexity of this hard-to-parallelize operation. Second, we wish to take as many primal constraints as necessary to achieve a satisfactory convergence rate; this means that if incorporation of some extra primal constraints could substantially improve robustness, this enrichment is beneficial. Setting up such constraints with no or limited prior knowledge of coefficient distribution, decomposition into subdomains, their mutual effect on one another, as well as different factors such as scaling is yet difficult. Therefore, it is essential to thoughtfully consider the number of constraints enforced.

The formulation (2.15) is convenient because the solution can now be obtained as the energy minimiser of a quadratic form with positive (semi)definite matrix F; and therefore the Krylov subspace methods with a short recurrence property, such as the conjugate gradient (CG) method, can be adopted. Please note that the zero eigenvalues of F emerges from redundant constraints in B. However, the eigenmodes corresponding to zero eigenvalues are invisible to the gradient-based iteration method, and the solution in terms of displacement variables remains unique.

Also, despite the fact that the expression is probably the most widely used, it is not the only possible way to solve the master system iteratively. When the primal problem becomes too high in dimension to significantly harm the desired parallel scalability, the saddle point formulation can be taken as the stepping stone; see, e.g. [26], where inexact iterative solvers (such as generalized minimal residual method) were adopted.

2.3.1 Preconditioning

As is common for all iterative solvers, their efficiency is contingent upon the use of the right preconditioner. In this context, the word "right" is a bit vague. The appropriate preconditioner should be fairly cheap to apply, yet should store the needed information such that the preconditioned system has a (rather significantly) lower condition number and the eigenvalue distribution is better clustered. For the family of one-level FETI methods and for many domain decomposition methods, the preconditioner is set as a weighted sum of local contributions [9, 44, 55]. Taking this into account, the quality of the one-level preconditioner depends on two factors only: (i) how accurately the approximation of the inverse of the system matrix is computed on the subdomain level, and (ii) the provided weights of these contributions. In this thesis, we restrict ourselves only to the use of the most accurate (and, unavoidably, the most computationally demanding) choice of setting the localized inverse

approximation; cf. our work [36, 37]. Thus, the optimal Dirichlet preconditioner is given by

$$\mathsf{M}_{\mathrm{D}}^{-1} = \sum_{i=1}^{N_{\mathrm{s}}} \mathsf{R}_{\mathrm{B}}^{(i)} \mathsf{B}_{\mathrm{D}}^{(i)} \mathsf{R}_{\Gamma}^{(i)} \mathsf{S}_{\Gamma\Gamma}^{(i)} \mathsf{R}_{\Gamma}^{(i),\mathsf{T}} \mathsf{B}_{\mathrm{D}}^{(i),\mathsf{T}} \mathsf{R}_{\mathrm{B}}^{(i),\mathsf{T}} \tag{2.21}$$

with restriction $\mathsf{R}_{\Gamma}^{(s),\mathsf{T}}$ from DOFs on subdomain $\Omega^{(s)}$ in set Γ to those in Δ, or equivalently, with use of $\mathsf{S}_{\Delta}^{(i)} = \mathsf{K}_{\Delta\Delta}^{(i)} - \mathsf{K}_{\Delta\mathrm{I}}^{(i)} {\mathsf{K}_{\mathrm{II}}^{(i)}}^{-1} \mathsf{K}_{\mathrm{I}\Delta}^{(i)}$

$$\mathsf{M}_{\mathrm{D}}^{-1} = \sum_{s=1}^{N_{\mathrm{s}}} \mathsf{R}_{\mathrm{B}}^{(i)} \mathsf{B}_{\Delta,\mathrm{D}}^{(i)} \mathsf{S}_{\Delta}^{(i)} \mathsf{B}_{\Delta,\mathrm{D}}^{(i),\mathsf{T}} \mathsf{R}_{\mathrm{B}}^{(i),\mathsf{T}} \tag{2.22}$$

The lowest eigenvalue of the algorithm is bounded by one; this is explicitly stated in [6]. For additional references on the minimum and maximum eigenvalue bound estimates, we refer the reader to [28, 28, 32, 39, 55]

2.4 Scaling possibilities

Scaling of continuity constraints is an indispensable strategy to accelerate convergence of a domain decomposition scheme. Motivated by either physical or geometric reasoning, various types of scalings have been established over time, from the purely geometrical [44], applicable predomimantly to homogeneous problems, to the more sophisticated, and thus more computationally demanding ones [8, 41]. Note that we refer to this strategy as *scaling*, even though its application can be in general represented by non-diagonal matrices, which is the fact that contradicts the general perception of scaling.

For a two dimensional application, we will consider a general form of scaling matrices

$$\mathsf{D}_{\Gamma}^{(s)} = \sum_{\mathcal{E}_{sj} \subset \Gamma^{(s)}} \mathsf{R}_{E_{sj}}^{(s)} \mathsf{D}_{E_{sj}}^{(j)} \mathsf{R}_{E_{sj}}^{(s)\,\mathsf{T}}, \tag{2.23}$$

where $\mathsf{R}_{E_{sj}}^{(s)}$ represents the assembly operator that maps contributions from individual boundaries with neighboring subdomains into the entire $\Gamma^{(s)}$. Since there is no risk of confusion, we leave out the subscript Γ in the scaling matrices $\mathsf{D}_{\Gamma}^{(s)}$. The scaling matrices $\mathsf{D}^{(s)}$, $s \in \{1 \dots N_{\mathrm{s}}\}$ are needed to form a scaled jump operator

$$\mathsf{B}_{\mathrm{D}} = \begin{bmatrix} \mathsf{D}^{(1),\mathsf{T}} \mathsf{B}^{(1)} & \mathsf{D}^{(2),\mathsf{T}} \mathsf{B}^{(2)} & \dots & \mathsf{D}^{(N_s),\mathsf{T}} \mathsf{B}^{(N_s)} \end{bmatrix}. \tag{2.24}$$

Individual domain-wise scaling matrices $\mathsf{D}^{(s)}$ are constructed such that the resulting scaled jump operator preserves partition of unity property in a sense [16]

$$\sum_{s=1}^{N_s} \mathsf{B}^{(s)} \mathsf{B}_{\mathrm{D}}^{(s)\mathsf{T}} = \mathsf{I}. \tag{2.25}$$

Thus, the scaled jump operator $\mathsf{B}_{\mathrm{D}}^{(s)}$ can be viewed as a suitable generalized inverse of $\mathsf{B}^{(s)}$. Such construction allows for a consistent splitting of the gap represented by the residual.

The four frequently utilized forms of $\mathsf{D}_{E_{sj}}^{(s)}$ known from the literature are concisely described below. For all of them, we assume that the jump operator is of a signed Boolean type. Subsequently, although we do not investigate this particular case in the presented thesis, it is assumed that there are rows in B defined between each pair of neighboring subdomains. For instance, we would adopt all six Lagrange multipliers for a dual corner at a point of contact of four subdomains, despite any combination of three linearly independent Lagrange multipliers would be sufficient. This case is often referred to as a fully redundant set of Lagrange multipliers. Implied redundancy does not have any harmful influence on the solution which remains unique; only the acting gluing forces represented by Lagrange multipliers are not uniquely determined. On the contrary, it is favourable from an implementation perspective as it allows for a simple application of intended scaling. In practice, diagonal scaling matrices $\mathsf{D}^{(s)}$ are never built explicitly [44].

2.4.1 Multiplicity scaling

Starting with the most simple one, multiplicity scaling [44], originally proposed by Rixen and Farhat in 1999, serves to satisfy relationship (2.25) by splitting the jump proportionally to the number of attached subdomains. Let $\mathcal{N}_x$ be the set of subdomain indices with a boundary node located at $\boldsymbol{x}$, recall Eq. (2.3). With a certain abuse of notation, we use $\boldsymbol{x}$ for the coordinates while x denotes the node in disretization of Ω. In multiplicity scaling, we use local inverse multiplicities

$$m_l^{-1}(x) = \frac{1}{\#\mathcal{N}_x} \quad \forall\, l \in \mathcal{N}_x\ . \tag{2.26}$$

In this scaling, $\mathsf{D}_{E_{sj}}$ is given diagonal matrix with individual components $(\mathsf{D}^{(s)}_{E_{sj}})_{dd} = m_j^{-1}(x_{P(d)})$. Here, $P(d)$ denotes an index of the node pertinent to the given degree of freedom d.

2.4.2 ρ scaling

A general diagonal scaling for a fully redundant set of Lagrange multipliers can be established with the use of local counting functions [55]

$$\delta_l^{-1}(x) = \frac{\hat{\alpha}_l(x)}{\sum_{k\in\mathcal{N}_x}\hat{\alpha}_k(x)} \quad \forall x \in \Gamma_h^{(1)} \times \cdots \times \Gamma_h^{(N_s)}, \tag{2.27}$$

The values $\hat{\alpha} > 0$ are in the case of ρ scaling given by the maximum coefficient in the finite element support of x

$$\hat{\alpha}_s(x) = \max_{\boldsymbol{y}\in\omega(x)\cap\Omega^{(s)}} \alpha(y) \tag{2.28}$$

with α being either Young modulus E or ρ, e.g., thermal conductivity, and $\omega(x)$ denoting the support of a finite element functions pertinent to x.

This scaling pair-wise divides the gap between adjacent subdomains proportionally to the reciprocal maximum values of coefficients on elements within each affected subdomain. In contrast to stiffness scaling introduced next, ρ scaling delivers better the geometrically intended meaning of this scaling, especially in the case of non-uniform meshes. This scaling is fairly cheap and therefore widely used when the solver has access to the coefficients $\hat{\alpha}_s(x)$ on individual finite element nodes. For construction of $\mathsf{D}^{(s)}_{E_{sj}}$, we proceed analogously to Subsection 2.4.1, replacing $m_j^{-1}(x)$ with $\delta_j^{-1}\ (x)$.

2.4.3 Stiffness scaling

The second scaling introduced in [44] specifically targets heterogeneous problems with jumps in coefficients aligned along individual boundaries. Splitting of the gap still operates on pairs of degrees of freedom, this time in a manner that one might intuitively anticipate. For a mechanical problem, for instance, it is natural to expect that an appropriate splitting, aimed at optimally minimizing the energy after weighting, will tend to follow the stiffer than the significantly softer parts. In stiffness, or so called k-scaling, weighting is given by diagonal entries of local stiffness matrices as

$$\mathsf{D}^{(j)}_{E_{ij},dd} = \frac{\mathsf{K}^{(j)}_{E_{ij},dd}}{\sum_{k=i,j}\mathsf{K}^{(k)}_{E_{ij},dd}} \tag{2.29}$$

From a certain viewpoint, stiffness scaling can be viewed as an algebraic variant of the ρ scaling; compare with the previous section.

2.4.4 Deluxe scaling

Deluxe scaling was first introduced by Dohrmann and Widlund (2013) in [8]. Using the notation introduced by Rheinbach et al., the recipe for deluxe weighting matrices is as

follows [43]. In 2D, for each edge $\mathcal{E}_{ij}$ shared by subdomains i and j, we first seek for the restriction of $\mathsf{S}_\Gamma^{(l)}$, $l \in \{i, j\}$ to edge E_{ij}. With a minor deviation from the nomenclature used in the thesis, let us denote this term consistently with [43] by $\mathsf{S}_{E_{ij,0}}^{(l)}$. The subscript 0 emphasizes that this term represents the minimum energy extension from the edge $\mathcal{E}_{ij}$ into the interior of the subdomain with homogeneous Dirichlet conditions on $\Gamma^{(l)} \setminus E_{ij}$. Apart from the requirements on B discussed at the end of Subsection 2.4, we further make an assumption that the orientation of binding constraints is consistent across all edges. The edge-related part of the scaling matrix pertinent to subdomain $\Omega^{(l)}$, $l \in \{i, j\}$ has the following form

$$\mathsf{D}_{E_{ij}}^{(l)} = \left(\mathsf{S}_{E_{ij,0}}^{(i)} + \mathsf{S}_{E_{ij,0}}^{(j)}\right)^{-1} \mathsf{S}_{E_{ij,0}}^{(l)} \tag{2.30}$$

which inserted in Eq. (2.23) yields the final matrix $\mathsf{D}^{(i)}$. Note that the use of $\mathsf{D}_{E_{ij}}^{(j)}$ in the definition of $\mathsf{D}^{(i)}$ is due to the fact that FETI-DP operates on dual quantities. For a primal counterpart of FETI-DP, the BDDC method, $\mathsf{D}_{E_{ij}}^{(i)}$ would be used — then the formulation would match that given in [8].

As of now, everything is set correctly to define the scaled version of a jump operator. The product of the two symmetric terms on the right side of Eq. (2.30) is not symmetric, therefore, the transpose in each local term in expression (2.24) is needed.
Albeit delivering performance superior to the three previously introduced scaling strategies, deluxe scaling is indisputably computationally costly compared to diagonal scalings, which somehow limits its wider adoption in practice. However, we find it a powerful tool when dealing with varying coefficients in the interiors of subdomains. With information about (i) boundary conditions on $\partial\Omega$ and (ii) coefficient distribution inside subdomains, it provides a decent splitting of the gap especially in cases where, e.g., the high-coefficient aggregates vanish within the subdomain, that is, those that do not touch a complementary part of Γ w.r.t. given edge E_{sj}.

2.5 Enforcement of additional constraints

In the previous sections, we mentioned that an a priori coarse space is central to ensuring fast convergence as well as the quality of the initial estimate entering the iterative solver — both of these factors strongly affect the overall convergence and the resulting accuracy of the solution. When a system with defined primal conditions is ill-conditioned, and there is a risk of convergence rate related issues, it is desirable to enrich this set of continuity conditions by constructing a primally handled second level of the method, in which additional constraints are enforced. Possible forms of these additional conditions will be discussed in Chapter 3. For now, it suffices to say that the new conditions are general modes in their nature. These conditions can be understood as suitably chosen combinations of finite element basis functions at the dual interface of the subdomains, essential for ensuring rapid convergence. After a certain number of iterations the solver reaches a state when the approximated solution becomes predominantly discontinuous across the interface Γ particularly in modes that are poorly captured by the preconditioner. At this stage, the iterative solver itself ceases to be capable of generating these suitable Krylov search directions. This is caused by the gap in the quality of the preconditioner. By a constraint we mean, in accordance with the literature [31, 57], the expression

$$\mathsf{c}^\mathsf{T}\mathsf{B}\mathsf{u} = 0, \tag{2.31}$$

while by a constraint vector we understand vector c from the abovementioned equation. Using the constraint vector, we therefore require that a suitably chosen combination of jumps in the solution variables vanishes. Typically, we want to enforce more than one constraint. Thus, by inserting the constraint vectors into a rectangular matrix U, we get

$$\mathsf{U}^\mathsf{T}\mathsf{B}\mathsf{u} = 0\,. \tag{2.32}$$

For the sake of brevity, we introduce a matrix Q that does not operate on displacement jumps, but directly specifies individual modes, i.e. $\mathsf{Q}^\mathsf{T} = \mathsf{B}^\mathsf{T}\mathsf{U}$ and

$$\mathsf{Q}^\mathsf{T}\mathsf{u} = 0\,. \tag{2.33}$$

Note that we viewed the terms *mode* and *constraint* equivalently above, but in the following text we more often use the more general term *mode*, which does not directly imply the necessity of being enforced in the coarse space.

In this thesis, we exclusively consider constraints that operate on a single interface; that means, in the two-dimensional case, that each constraint can be defined only on one of the edges. Then, because the interaction is limited to a pair of substructures, the general expression from (2.31) simplifies to

$$\mathsf{c}^\mathsf{T}\mathsf{B}^{(i)}\mathsf{u}^{(i)} = -\mathsf{c}^\mathsf{T}\mathsf{B}^{(j)}\mathsf{u}^{(j)}\,. \tag{2.34}$$

Without loss of generality, we assume that our *constraint vector* c specifying an edge-related constraint operates on the part of the interface between substructures $\Omega^{(i)}$ and $\Omega^{(j)}$. From this point onwards, the joint occurrence of indices i and j typically signifies some relation to the edge E_{ij} between the subdomains $\Omega^{(i)}$ and $\Omega^{(j)}$. Denoting $\mathsf{B}^{(l)}_{E_{ij}}$, $l \in \{i, j\}$ a submatrix of $\mathsf{B}^{(l)}$ specifically restricted to edge E_{ij}, with non-contributing rows and trivial columns excluded, we can write

$$\mathsf{c}^\mathsf{T}_{E_{ij}}\mathsf{B}^{(i)}_{E_{ij}}\mathsf{u}^{(i)}_{|E_{ij}} = -\mathsf{c}^\mathsf{T}_{E_{ij}}\mathsf{B}^{(j)}_{E_{ij}}\mathsf{u}^{(j)}_{|E_{ij}}\,. \tag{2.35}$$

The vector $\mathsf{u}^{(i)}_{|E_{ij}}$ denotes the part of $\mathsf{u}^{(i)}$ corresponding to DOFs on an edge E_{ij}. With consistently established pairwise continuity constraints, $\mathsf{B}^{(l)}_{E_{ij}}$ is simply a permutation of the (negative) identity matrix. Recalling that $\mathsf{B}^{(i)}_{E_{ij}} = -\mathsf{B}^{(j)}_{E_{ij}}$ and $\mathsf{B}^{(l)}_{E_{ij}}$, $l \in \{i, j\}$ consists only of values from $\{-1, 0, 1\}$, we arrive at an expression for solution variables

$$\widehat{\mathsf{q}}^\mathsf{T}_{E_{ij}}\mathsf{u}^{(i)}_{|E_{ij}} = \widehat{\mathsf{q}}^\mathsf{T}_{E_{ij}}\mathsf{u}^{(j)}_{|E_{ij}}\,, \tag{2.36}$$

in which we can arbitrarily pick $\widehat{\mathsf{q}}^\mathsf{T}_{E_{ij}}$ to be either $\mathsf{c}^\mathsf{T}_{E_{ij}}\mathsf{B}^{(i)}_{E_{ij}}$ or $\mathsf{c}^\mathsf{T}_{E_{ij}}\mathsf{B}^{(j)}_{E_{ij}}$.

In the context of the relationship (2.36), the quantity $\widehat{\mathsf{q}}_{E_{ij}}$ is consistently referred to throughout this thesis as a *constraint mode*.

From our perspective, the exceptionally beneficial constraint modes are those reflecting the poorly preconditioned part of the system. However, the mere fact that preconditioning is not perfect does not necessarily pose a critical problem in practice. For instance, systems with dense clustered spectra and only a few outliers can be denoted as ill-conditioned. In such cases, as is demonstrated on numerous illustrative examples discussed later in this work, each enforced constraint mode typically results in saving one iteration. This is also theoretically justified, assuming that calculations are performed in exact arithmetic. Yet, for larger problems with more complex coefficient distributions, this is no longer a general rule. If the preconditioned system has too ill-conditioned spectrum, such as a flat one, a skeleton of the problem may be absent. Then an application of gradient-based solver such as the conjugate gradient (CG) method can prove tricky, as it might require more iterations than would be expected. Additionally, the convergence rate may be completely spoilt as iterates cease to converge towards the precise solution due to inaccuracies in numerical computations and the loss of orthogonality of search directions. Then, it is not uncommon for the solver to reach a plateau at a relatively high error norm, a case when the best solution obtained remains inaccurate even after a substantial number of iterations.

In spite of that, the *constraint vectors* $\mathsf{c}(\mathsf{M}^{-1})$ fundamentally represent functions of admissible harmful configurations in the solution variables field, roughly written as $\mathsf{c}(\mathsf{u}_{\text{bad}}(\mathsf{M}^{-1}))$. The vectors denoted here as $\mathsf{u}_{\text{bad}}(\mathsf{M}^{-1})$ are closely associated with the eigenmodes of (localized) eigenproblem(s) as defined later. These (approximated) eigenmodes will be frequently illustrated in this work so that the reader can develop a sense for the form of beneficial constraints.

The constraints presented above, targeting a trouble-making part of the system, can be

enforced through a range of different approaches. The enumeration of the most widely used ones is recalled in the text below to keep the thesis self-contained. Some of the approaches are described with a deeper level of explanation; consequently, the descriptions provided do not always align with the actual author's in-house implementation.

2.5.1 Projector preconditioning

Projector preconditioning, or so-called *deflation*, represents a straightforward way to enrich solver's coarse space with arbitrarily chosen set of new constraints. Contrary to approaches based on transformation of basis (ToB), compare with Subsections 2.5.2-2.5.4, projector preconditioning is relatively convenient from an implementation perspective, because it does not require any adjustments of the system products. Within deflation, the system itself remains untouched. We only construct a second coarse problem, independent of the original one, with the use of orthogonal projections. Thus, from this point of view, it might sound appealing to solve two smaller coarse problems than a larger single one. Here, we briefly recall the projector preconditioning and balancing approach based on the work of Klawonn and Rheinbach [28].

First, we introduce an orthogonal decomposition of the solution space, a procedure well known from the framework of FETI-1 or T-FETI [10, 30]. We let

$$\mathsf{P} = \mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{F} \tag{2.37}$$

be an F-orthogonal projection onto $\mathrm{Range}(\mathsf{U})$ and

$$\mathsf{I} - \mathsf{P} = \mathsf{I} - \mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{F} \tag{2.38}$$

the F-orthogonal complementary projection to P. The $\dagger$ symbol denotes the pseudoinverse, which is needed only if the matrix F is symmetric positive semidefinite, i.e., there is a redundancy of continuity constraints in the jump operator.

Then, we have the additive splitting of the exact solution of (2.15) in F-orthogonal subspaces

$$\lambda_{\mathrm{exact}} = \lambda' + \lambda^* \qquad \lambda' \in \mathrm{Range}(\mathsf{I} - \mathsf{P}), \quad \lambda^* \in \mathrm{Range}(\mathsf{P})\,. \tag{2.39}$$

The projection of the exact solution $\mathsf{P}\lambda_{\mathrm{exact}}$ onto $\mathrm{Range}(\mathsf{P}) = \mathrm{Range}\,(\mathsf{U})$ is computed as

$$\lambda^* = \mathsf{P}\lambda_{\mathrm{exact}} = \mathsf{P}\mathsf{F}^\dagger\mathsf{d} = \mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{d} \tag{2.40}$$

and it remains to compute λ'.

For $\vartheta \in \mathrm{Range}(\mathsf{U})$, we can write $\vartheta = \mathsf{U}\xi$ and

$$\begin{aligned}(\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{F}\vartheta &= (\mathsf{F} - \mathsf{F}\mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{F})\mathsf{U}\xi \\ &= \mathsf{F}\mathsf{U}\xi - \mathsf{F}\mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U}\xi \\ &= \mathsf{F}\mathsf{U}\xi - \mathsf{F}\mathsf{U}\xi \\ &= \mathsf{0},\end{aligned} \tag{2.41}$$

where we made use of symmetry of F. Thus, $\mathrm{Range}(\mathsf{U}) \subseteq \mathrm{Kernel}((\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{F})$. Let $\vartheta_\perp$ be F-orthogonal to $\mathrm{Range}(\mathsf{U})$, that is, $\mathsf{U}^\mathsf{T}\mathsf{F}\vartheta_\perp = \mathsf{0}$. Now

$$\begin{aligned}(\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{F}\vartheta_\perp &= (\mathsf{F} - \mathsf{F}\mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{F})\vartheta_\perp \\ &= \mathsf{F}\vartheta_\perp - \mathsf{F}\mathsf{U}(\mathsf{U}^\mathsf{T}\mathsf{F}\mathsf{U})^\dagger\mathsf{U}^\mathsf{T}\mathsf{F}\vartheta_\perp \\ &= \mathsf{F}\vartheta_\perp\end{aligned} \tag{2.42}$$

and no nontrivial $\vartheta_\perp$ lies in $\mathrm{Kernel}((\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{F})$. Consequently, we see that

$$\mathrm{Kernel}((\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{F}) = \mathrm{Range}(\mathsf{U})\,. \tag{2.43}$$

Qe our now approaching our goal. Our aim is to seek for λ' iteratively by solving a problem

$$(\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{F}\lambda = (\mathsf{I} - \mathsf{P})^\mathsf{T}\mathsf{d} \tag{2.44}$$

with imposed auxiliary conditions stored in U, resulting in iterates that by construction satisfy the condition $\mathsf{U}^\mathsf{T}\mathsf{B}\mathsf{u}$. To do so, we make use of a preconditioned conjugate gradient method. We only need to keep directions of search in $\mathrm{Range}(\mathsf{I} - \mathsf{P})$, which is done by projecting the

correction in each iteration. For symmetry reasons, we construct a deflated preconditioner as

$$\mathsf{M}_{\mathrm{PP}}^{-1} = (\mathsf{I} - \mathsf{P})\mathsf{M}_{\mathrm{D}}^{-1}(\mathsf{I} - \mathsf{P})^{\mathsf{T}} \tag{2.45}$$

and solve Eq. (2.44) iteratively with a symmetric preconditioner with an initial guess λ^* in a following way

$$\mathsf{M}_{\mathrm{PP}}^{-1}\mathsf{F}\lambda' = \mathsf{M}_{\mathrm{PP}}^{-1}\mathsf{d}. \tag{2.46}$$

The ultimate solution is then composed of two contributions: initial guess from (2.40) and iteratively found λ' from Eq. (2.46),

$$\lambda = \lambda' + \lambda^* \,. \tag{2.47}$$

Note that the system (2.46) is only positive semidefinite, because the eigenvalues corresponding to deflated constraints are mapped to zero.

An alternative approach, which prevents introducing a rank-deficiency, is to incorporate the particular solution λ^* directly into the preconditioner. This gives rise to the balancing preconditioner

$$\mathsf{M}_{\mathrm{bal}}^{-1} = \mathsf{M}_{\mathrm{PP}}^{-1} + \mathsf{U}(\mathsf{U}^{\mathsf{T}}\mathsf{F}\mathsf{U})^{\dagger}\mathsf{U}^{\mathsf{T}} \,. \tag{2.48}$$

The part of the spectrum mapped to zero by application of $\mathsf{M}_{\mathrm{PP}}^{-1}$ is in the case of $\mathsf{M}_{\mathrm{bal}}^{-1}$ shifted to one [28], i.e., the conjugate gradients algorithm is no longer insensitive to contributions in search directions lying in Range(U). With balancing, we only have to solve

$$\mathsf{M}_{\mathrm{bal}}^{-1}\mathsf{F}\lambda = \mathsf{M}_{\mathrm{bal}}^{-1}\mathsf{d} \,. \tag{2.49}$$

For more details, we refer the reader to [20, 28].

The deflation and balancing approaches are proficient tools for a coarse space augmentation. However, these techniques exhibit two closely related drawbacks. First, their performance quickly deteriorates if projections are computed inexactly [28]. In particular, this extremely detrimental behaviour becomes evident once CGs recognize small non-zero eigenvalues. In such a case, convergence issues usually emerge and projector preconditioning becomes unstable. Consequently, for our testing purposes, we rather opted for the balancing approach, as it is generally more stable. On the other hand, balancing can often be even more deceptive, because it might compute an incorrect solution without any prior indication. For very ill-conditioned problems, it is challenging to ensure that the term $(\mathsf{U}^{\mathsf{T}}\mathsf{F}\mathsf{U})^{\dagger}$ is computed precisely enough, making the convergence behavior dependent also on the particular implementation. For example, Kühn [31] states that the efficacy of deflation-based techniques strongly depends on the way the (generalized) inverse $(\mathsf{U}^{\mathsf{T}}\mathsf{F}\mathsf{U})^{\dagger}$ is handled; it is crucial to exploit the sparsity structure of U and F to reduce the computational cost to a minimum. He further mentions that the (pseudo)inverse of the given Galerkin projection can then be computed at a cost of the sparse Cholesky factorization [31]. We have not adopted this in our implementation yet. However, we do not see any reason why this should affect the precision achieved.

This brings us to the second drawback of the approaches that rely on projections: for large U, the application of projections becomes costly. Unfortunately, deflation-based approaches are not amenable to inexact solvers. For example, approaches utilizing transformation of basis with a partial subassembly are generally considered more robust and suitable for employing inexact solvers. These methods will be discussed in the subsequent subchapters. In fact, we were forced to implement a more complex generalized transformation of basis due to problems with projections: we frequently encountered convergence issues when solving highly heterogeneous problems, especially those with flat spectra.

2.5.2 Transformation of basis

Assuming that we want to augment the coarse space with constraints of non-nodal character, typically the adaptive constraints or, e.g. (weighted) averages, we introduce a transformation in which the modes we want to enforce are represented directly as components of solution vector $\bar{\mathsf{u}}$ in the new basis, such that

$$\mathsf{u}^{(s)} = \mathsf{T}^{(s)}\bar{\mathsf{u}}^{(s)}. \tag{2.50}$$

Here, we follow a convention established in the literature that T provides a transformation from new, generally non-nodal basis to the original one. Note that although we refer to matrix T as a transformation, it is not necessarily required to be orthonormal. The only requirement for the transformation is that n_c^s basis vectors $\mathsf{q}_i^{(s)}$ in T are orthogonal to the remaining basis vectors collected as columns in $\mathsf{Q}^{\perp,(s)}$, i.e.

$$\mathsf{T}^{(s)} = \begin{bmatrix} \mathsf{q}_1^{(s)} & \dots & \mathsf{q}_{n_c^s}^{(s)} & \mathsf{Q}^{\perp,(s)} \end{bmatrix} \qquad \mathsf{q}_i^{(s),\mathsf{T}} \mathsf{Q}^{\perp,(s)} = \mathsf{0}^{\mathsf{T}} \quad \forall\, 1 \leq i \leq n_c^s \,.$$

However, since we employ the transformation explicitly, we always construct T such that it has orthonormal columns. We start with an orthonormalized set of modes and compute an orthogonal complement $\mathsf{Q}^{\perp,(s)}$ to Range$(\begin{bmatrix}\mathsf{q}_i^{(s)},\dots,\mathsf{q}_{n_c^s}^{(s)}\end{bmatrix})$. It is clear that, for a given set of constraints (orthonormalized or not), the transformation is not uniquely determined. Thus, to obtain an effective transformation, one should carefully consider the balance between (i) the preservation of sparsity in the remaining dual variables and (ii) having a general formulation for the assembly of the block of T that only concerns the remaining DOFs. Note that the transformation is always nontrivial only on the degrees of freedom corresponding to a dual set of variables. Hence, in the interior and primal DOFs, the transformation is formally an identity. Assuming that dual DOFs are ordered in groups pertinent to individual edges, relation between the original and transformed variables is given by

$$\begin{bmatrix} \mathsf{u}_{\mathrm{I}}^{(s)} \\ \mathsf{u}_{\Delta'}^{(s)} \\ \mathsf{u}_{\Pi'}^{(s)} \end{bmatrix} = \mathsf{T}^{(s)} \begin{bmatrix} \overline{\mathsf{u}}_{\mathrm{I}}^{(s)} \\ \overline{\mathsf{u}}_{\Delta'}^{(s)} \\ \overline{\mathsf{u}}_{\Pi'}^{(s)} \end{bmatrix} = \begin{bmatrix} \mathsf{I} & & \\ & \operatorname{diag}_{\{\mathcal{E}_{sj} \cap \Gamma^{(s)} \neq \emptyset\}} (\mathsf{T}_{E_{sj}}^{(s)}) & \\ & & \mathsf{I} \end{bmatrix} \begin{bmatrix} \mathsf{u}_{\mathrm{I}}^{(s)} \\ \overline{\mathsf{u}}_{\Delta'}^{(s)} \\ \mathsf{u}_{\Pi'}^{(s)} \end{bmatrix}, \tag{2.51}$$

where constraint modes $\widehat{\mathsf{q}}_{E_{sj}}$ defined by (2.36) find their application

$$\mathsf{T}_{E_{sj}}^{(s)} = \begin{bmatrix} \widehat{\mathsf{q}}_{E_{sj},1} & \dots & \widehat{\mathsf{q}}_{E_{sj},n_c^{(s)}} & \widehat{\mathsf{Q}}_{E_{sj}}^{\perp,(s)} \end{bmatrix} \qquad \widehat{\mathsf{q}}_{E_{sj},i}^{(s),\mathsf{T}} \widehat{\mathsf{Q}}_{E_{sj}}^{\perp,(s)} = \mathsf{0}^{\mathsf{T}} \quad \forall\, 1 \leq i \leq n_c^{E_{sj}} \,.$$

In the expressions above, we have used a superscript $\bullet'$ to distinguish the a priori dual Δ' and primal Π' set. Here, we have directly introduced a transformation by means of block-diagonal contributions corresponding to individual edges. This implies that modes in the new basis are restricted to individual edges, i.e., no mode in the new basis shares node values with another edge. This is consistent with the architecture of the method, which benefits greatly from the localisation of contributions to individual interfaces. From now on, we omit the denotation for the transformed interior and primal nodes as they are not affected by the transformation, and we write $\overline{\mathsf{u}}^{(s)} = \begin{bmatrix} \mathsf{u}_{\mathrm{I}}^{(s)\mathsf{T}} & \overline{\mathsf{u}}_{\Delta'}^{(s)\mathsf{T}} & \mathsf{u}_{\Pi'}^{(s)\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$.

Being the new constraint modes explicitly expressed in the new basis, we can treat them as other primal constraints. Hence, a partial assembly in the new degrees of freedom is to be applied. The specific manner in which this assembly process is enforced, together with the treatment of the scaling, depends on the approach called. In fact, one can opt between two possibilities. Both are recalled in the following.

2.5.3 Standard Transformation of basis

In the standard transformation of basis (sToB), we introduce new constraints by a direct reclassification of solution variables from the dual set Δ to the primal set Π. From now on, we search for a solution in the transformed basis denoted by $\overline{\bullet}$ symbol, i.e., for each domain we have

$$\overline{\mathsf{K}}^{(s)} = \mathsf{T}^{(s),\mathsf{T}} \mathsf{K}^{(s)} \mathsf{T}^{(s)} \tag{2.52}$$

$$\overline{\mathsf{f}}^{(s)} = \mathsf{T}^{(s),\mathsf{T}} \mathsf{f}^{(s)} \,. \tag{2.53}$$

We will drop the overline in the terms not affected by a transformation. Note again that the transformation matrices $\mathsf{T}^{(s)}$ do not necessarily have to be orthogonal. The simplest and most straightforward way to incorporate the non-nodal constraints that are known in advance into the coarse problem is to insert the relations with actualized assembly and jump operators

into the scheme presented in Section 2.3. Then, we arrive at

$$\begin{bmatrix} \overline{\mathsf{K}}_{\mathrm{RR}} & \widetilde{\overline{\mathsf{K}}}_{\Pi\mathrm{R}}^{\mathsf{T}} & \mathsf{B}_{\mathrm{R}}^{\mathsf{T}} \\ \widetilde{\overline{\mathsf{K}}}_{\Pi\mathrm{R}} & \widetilde{\overline{\mathsf{K}}}_{\Pi\Pi} & 0 \\ \mathsf{B}_{\mathrm{R}} & 0 & 0 \end{bmatrix} \begin{bmatrix} \overline{\mathsf{u}}_{\mathrm{R}} \\ \widetilde{\overline{\mathsf{u}}}_{\Pi} \\ \overline{\lambda} \end{bmatrix} = \begin{bmatrix} \overline{\mathsf{f}}_{\mathrm{R}} \\ \widetilde{\overline{\mathsf{f}}}_{\Pi} \\ 0 \end{bmatrix} \tag{2.54}$$

and ultimately get a system similar to (2.15) operating in the new basis of the form

$$\overline{\mathsf{F}}\,\overline{\lambda} = \overline{\mathsf{d}}\,. \tag{2.55}$$

Now, all the primal constraints are enforced by the corresponding assembly. The jump operators are still defined such that there is one Boolean continuity constraint for each pair of remaining dual DOFs. The formulation of the FETI-DP algorithm in the new basis is essentially the same as in the original basis. To obtain a solution in the original basis, we only have to perform a back-transformation

$$\mathsf{u}^{(s)} = \mathsf{T}^{(s)}\overline{\mathsf{u}}^{(s)} \quad s = 1, \ldots, N_{\mathrm{s}}\,.$$

However, such approach faces two primary challenges. The first challenge is that constraints are often not predetermined, for instance, when they depend on factors such as scaling or coefficient distribution. Consequently, it is either impossible or disadvantageous to communicate between subdomains during the preprocessing phase. The second challenge is related to the definition of scaling itself, unless it is constant. This is due to a potential interaction between the new non-nodal primal and the remaining dual variables imposed by transformation matrices $\mathsf{T}^{(s)}$. To mitigate these obstacles in our implementation, we consistently define a new row in B containing a continuity condition for every edge-related pair of DOFs. Thus, form of the matrix B corresponds to a jump operator used in a standard FETI-DP algorithm with a nodal basis.

Henceforth, we assume that there are no non-nodal primal constraints in the prior coarse space. These non-nodal constraints will only be defined on the fly via a formal reclassification of variables, even if their form is known in advance. Note that they also can constitute the whole coarse space if there are none nodal a priori primal constraints, as is the case with averages over edges. Hence, if we denote by Π' and Δ' the (possibly empty) prior primal and prior dual set of variables and let $\mathrm{R}' = \mathrm{I} \cap \Delta'$, we can rewrite Eq. (2.12) as

$$\begin{bmatrix} \mathsf{K}_{\mathrm{R}'\mathrm{R}'} & \widetilde{\mathsf{K}}_{\Pi'\mathrm{R}'}^{\mathsf{T}} & \mathsf{B}_{\mathrm{R}'}^{\mathsf{T}} \\ \widetilde{\mathsf{K}}_{\Pi'\mathrm{R}'} & \widetilde{\mathsf{K}}_{\Pi'\Pi'} & 0 \\ \mathsf{B}_{\mathrm{R}'} & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathsf{u}_{\mathrm{R}'} \\ \tilde{\mathsf{u}}_{\Pi'} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathsf{f}_{\mathrm{R}'} \\ \tilde{\mathsf{f}}_{\Pi'} \\ 0 \end{bmatrix}. \tag{2.56}$$

Then, we introduce a posterior (non-nodal) primal constraint set Π^* and accordingly update the dual set Δ^*, and we expand (2.56) to

$$\begin{bmatrix} \mathsf{K}_{\mathrm{I}\,\mathrm{I}} & \overline{\mathsf{K}}_{\mathrm{I}\,\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\mathrm{I}\,\Pi^*} & \widetilde{\mathsf{K}}_{\mathrm{I}\,\Pi'} & 0 \\ \overline{\mathsf{K}}_{\Delta^*\mathrm{I}} & \overline{\mathsf{K}}_{\Delta^*\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\Delta^*\Pi^*} & \widetilde{\overline{\mathsf{K}}}_{\Delta^*\Pi'} & \mathsf{B}_{\Delta^*}^{\mathsf{T}} \\ \widetilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{I}} & \widetilde{\overline{\mathsf{K}}}_{\Pi^*\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\Pi^*\Pi^*} & \widetilde{\overline{\mathsf{K}}}_{\Pi^*\Pi'} & 0 \\ \widetilde{\mathsf{K}}_{\Pi'\mathrm{I}} & \widetilde{\overline{\mathsf{K}}}_{\Pi'\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\Pi'\Delta^*} & \widetilde{\mathsf{K}}_{\Pi'\Pi'} & 0 \\ 0 & \mathsf{B}_{\Delta^*} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathsf{u}_{\mathrm{I}} \\ \overline{\mathsf{u}}_{\Delta^*} \\ \widetilde{\overline{\mathsf{u}}}_{\Pi^*} \\ \tilde{\mathsf{u}}_{\Pi'} \\ \overline{\lambda} \end{bmatrix} = \begin{bmatrix} \mathsf{f}_{\mathrm{I}} \\ \overline{\mathsf{f}}_{\Delta^*} \\ \widetilde{\overline{\mathsf{f}}}_{\Pi^*} \\ \tilde{\mathsf{f}}_{\Pi'} \\ 0 \end{bmatrix} \tag{2.57}$$

with explicitly written out a posteriori assembled quantities

$$\widetilde{\overline{\mathsf{K}}}_{\Pi^*\Pi^*} = \sum_{i=1}^{N_{\mathrm{s}}} \mathsf{R}_{\Pi^*}^{(i)\,\mathsf{T}} \overline{\mathsf{K}}_{\Pi^*\Pi^*}^{(i)} \mathsf{R}_{\Pi^*}^{(i)} = \mathsf{R}_{\Pi^*}^{\mathsf{T}} \overline{\mathsf{K}}_{\Pi^*\Pi^*} \mathsf{R}_{\Pi^*} \qquad \widetilde{\overline{\mathsf{f}}}_{\Pi^*} = \sum_{i=1}^{N_{\mathrm{s}}} \mathsf{R}_{\Pi^*}^{(i)\,\mathsf{T}} \mathsf{f}_{\Pi^*}^{(i)},$$

$$\widetilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{R}^*} = \begin{bmatrix} \mathsf{R}_{\Pi^*}^{(1)\,\mathsf{T}} \overline{\mathsf{K}}_{\Pi^*\mathrm{R}^*}^{(1)} & \ldots & \mathsf{R}_{\Pi^*}^{(N_{\mathrm{s}})\,\mathsf{T}} \overline{\mathsf{K}}_{\Pi^*\mathrm{R}^*}^{(N_{\mathrm{s}})} \end{bmatrix} = \widetilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}^{\mathsf{T}} \quad \text{where} \quad \mathsf{R}_{\Pi^*}^{\mathsf{T}} = \begin{bmatrix} \mathsf{R}_{\Pi^*}^{(1),\mathsf{T}} & \mathsf{R}_{\Pi^*}^{(2),\mathsf{T}} & \ldots & \mathsf{R}_{\Pi^*}^{(N_{\mathrm{s}}),\mathsf{T}} \end{bmatrix}.$$

In contrast to Eqs. (2.54-2.55), here in the system we keep the Lagrange multipliers that are designed to enforce conditions that are now integrated into the coarse problem as components of Π^*. That is because the system assembled in Π^* is statically condensed on Lagrange multipliers. Written out, we expect the relation

$$\mathsf{B}_{\Delta'|\Pi^*}\,\overline{\mathsf{u}}_{\Delta'|\Pi^*} = 0 \tag{2.58}$$

which, given that $\mathsf{B}_{\Delta'|\Pi^*} = \mathsf{B}_{\Pi^*}$ and $\bar{\mathsf{u}}_{\Pi^*} = \mathsf{R}_{\Pi^*}\widetilde{\bar{\mathsf{u}}}_{\Pi^*}$, is directly implied by

$$\mathsf{B}_{\Pi^*}\,\mathsf{R}_{\Pi^*}^{\mathsf{T}} = 0 \tag{2.59}$$

to hold. For the sake of clarity, we note that by $\mathsf{B}_{\Delta'|\Pi^*}$ we understand the submatrix of $\mathsf{B}_{\Delta'|\Pi^*}$ with columns pertaining to DOFs in Π^*, and the zero blocks at positions three-five and five-three in Eq. (2.57) are a consequence of (2.59). Although there is no apparent motivation to keep two separate sets Π' and Π^*, we kept them to emphasize the twofold character of posterior constraints, which are now treated as primary while still being maintained in a dual arrangement. After the elimination of all solution variables, the system reads

$$\bar{\mathsf{F}}'\,\bar{\lambda}' = \bar{\mathsf{d}}'\,. \tag{2.60}$$

The problem with scaling is as follows. In standard FETI-DP the preconditioning step determining a new search direction in the kth iteration involves three steps: subsequent multiplication of the three matrices $\mathsf{B}_{\Delta,\mathrm{D}}^{\mathsf{T}}$, S_Δ and $\mathsf{B}_{\Delta,\mathrm{D}}$ with the kth residual $\mathsf{r}^k = \mathsf{B}\mathsf{u}^k$ vector,

$$\begin{aligned}\mathsf{z}^{k+1} &= \mathsf{M}_{\mathrm{D}}\,\mathsf{r}^k \\ &= \mathsf{B}_{\Delta,\mathrm{D}}\,\mathsf{S}_\Delta\,\mathsf{B}_{\Delta,\mathrm{D}}^{\mathsf{T}}\,\mathsf{B}\,\mathsf{u}^k \\ &= \sum_{s=1}^{N_{\mathrm{s}}}\left(\mathsf{R}_{\mathsf{B}_\Delta}^{(s)}\,\mathsf{D}^{(s),\mathsf{T}}\,\mathsf{B}_\Delta^{(s)}\,\mathsf{S}_\Delta^{(s)}\,\mathsf{B}_\Delta^{(s),\mathsf{T}}\,\mathsf{D}^{(s)}\,\mathsf{R}_{\mathsf{B}_\Delta}^{(s),\mathsf{T}}\right)\mathsf{B}\,\mathsf{u}^k\,,\end{aligned} \tag{2.61}$$

where $\mathsf{D}^{(s)}$ are appropriately chosen local scaling matrices that, once assembled, provide a partition of unity such that

$$\sum_{s=1}^{N_{\mathrm{s}}}\mathsf{R}_{\mathsf{B}_\Delta}^{(s)}\mathsf{D}^{(s)}\mathsf{R}_{\mathsf{B}_\Delta}^{(s),\mathsf{T}} = \mathsf{I}\,. \tag{2.62}$$

However, in the new basis, we operate with directly transformed entities $\bar{\mathsf{F}}'$ and $\bar{\mathsf{d}}'$, and the transformation of the preconditioner M_{D} is needed as well. This must be conducted with great caution because the transformation of the scaling is not straightforward.

We demonstrate this challenge of the scaling transformation with a simple two-dimensional problem decomposed into only two substructures $\Omega^{(1)}$ and $\Omega^{(2)}$, which meet at a single edge E_{12}. We assume that there is at least one non-nodal primal constraint on the interface E_{12}, and this constraint is given by a column in $\mathsf{T}_{E_{12}}$. The constraint is then enforced by a partial subassembly. With a slight abuse of notation, we temporarily let $\mathsf{T}_{E_{12}} =: \mathsf{T}_{\Delta'}^{(1)} = \mathsf{T}_{\Delta'}^{(2)}$ of the form

$$\mathsf{T}_{E_{12}} = \begin{bmatrix}\mathsf{T}_{\Delta^*} & \mathsf{T}_{\Pi^*}\end{bmatrix}\,. \tag{2.63}$$

Now, the equivalent of Eq. (2.61) is

$$\bar{\mathsf{z}}^{k+1} = \sum_{s=1}^{2}\left(\mathsf{R}_{\mathsf{B}_{\Delta^*}}^{(s)}\,\bar{\mathsf{D}}_{\Delta^*}^{(s),\mathsf{T}}\mathsf{B}_{\Delta^*}^{(s)}\bar{\mathsf{S}}_{\Delta^*}^{(s)}\mathsf{B}_{\Delta^*}^{(s),\mathsf{T}}\bar{\mathsf{D}}_{\Delta^*}^{(s)}\mathsf{R}_{\mathsf{B}_{\Delta^*}}^{(s),\mathsf{T}}\right)\mathsf{B}\bar{\mathsf{u}}^k\,. \tag{2.64}$$

The properly transformed scaling, which typically operates on the original Δ'; recall Subsection 2.5.2, would be

$$\bar{\mathsf{D}}^{(l)} = \mathsf{T}_{E_{ij}}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{E_{ij}} = \begin{bmatrix}\mathsf{T}_{\Delta^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Delta^*} & \mathsf{T}_{\Delta^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Pi^*} \\ \mathsf{T}_{\Pi^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Delta^*} & \mathsf{T}_{\Pi^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Pi^*}\end{bmatrix} \qquad l \in \{1,2\}. \tag{2.65}$$

Assuming the nodal-based scaling that satisfies $\mathsf{D}^{(1)} + \mathsf{D}^{(2)} = \mathsf{I}$, the sum $\bar{\mathsf{D}}^{(1)} + \bar{\mathsf{D}}^{(2)}$ clearly satisfies the partition of unity property as well for orthogonal transformation. In our computations, we are constrained to utilize only a restricted part $\bar{\mathsf{D}}_{\Delta^*}^{(l)}$ of $\bar{\mathsf{D}}^{(l)}$. However, this comes at the cost of omitting particular components of the scaling which have influence on the remaining dual set. For simplicity, we first consider the diagonal scaling. The residual $\bar{\mathsf{r}} := \mathsf{B}\bar{\mathsf{u}}$ can generally be nonzero only on the Lagrange multipliers defining Δ^*. With restriction $\mathsf{R}_{\mathsf{B}_{\Delta'}}^{(s),\mathsf{T}}$ in Eq. (2.64), we obtain $\begin{bmatrix}\Delta\bar{\mathsf{u}}_{\Delta^*}^{\mathsf{T}} & 0^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$ and we observe a splitting of the jump affecting the set Π^*, which due to the subsequent application of $\bar{\mathsf{S}}_{\Delta^*\Delta^*}^{(s)}$ cannot be incorporated into calculations. Additionally, the construction of a correction field in the third step encounters a similar

problem. This interaction can occur as long as the off-diagonal blocks of $\overline{\mathsf{D}}$ are nonempty. Hence, all but constant scaling, such as multiplicity scaling, are problematic. With constant scaling, we have by construction $\mathsf{T}_{\Pi^*}^{\mathsf{T}}\mathsf{T}_{\Delta^*} = 0$ and thus $\overline{\mathsf{D}}^{(l)} = \begin{bmatrix} \mathsf{T}_{\Delta^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Delta^*} & 0 \\ 0 & \mathsf{T}_{\Pi^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Pi^*} \end{bmatrix}$ or $\overline{\mathsf{D}}^{(l)} = \mathsf{D}^{(l)}$, depending on whether the transformation employed is orthonormal or not.

To the best of author's knowledge, there is no elegant solution that effectively resolves the issue of interaction in the scaling when using explicitly transformed variables. In an effort to approach the correct transformed scaling, we apply a restricted transformed scaling on DOFs in Δ^* together with use of orthogonal transformation matrices, i.e., we set

$$\begin{aligned} \overline{\mathsf{D}}_{\Delta^*}^{(l)} &= \mathsf{T}_{\Delta^*}^{\mathsf{T}}\mathsf{D}^{(l)}\mathsf{T}_{\Delta^*} \\ &= \mathsf{T}_{\Delta^*\Delta^*}^{\mathsf{T}}\mathsf{D}_{\Delta^*}^{(l)}\mathsf{T}_{\Delta^*\Delta^*} + \mathsf{T}_{\Pi^*\Delta^*}^{\mathsf{T}}\mathsf{D}_{\Pi^*}^{(l)}\mathsf{T}_{\Pi^*\Delta^*} \qquad l \in \{i, j\}\,, \end{aligned} \tag{2.66}$$

where $\mathsf{D}_{\Delta^*}^{(l)}$ is a submatrix of $\mathsf{D}^{(l)}$ pertaining to DOFs in Δ^*. Now we elucidate the necessity of having the original scaling accessible. Without this, it would not only be impossible to incorporate contributions from Π^*, but it would also require the normalization of dense matrices. Moreover, this approach allows us to reuse the same segments of code employed in standard FETI-DP and merely perform an additional transformation from the old basis to the new one. By maintaining the partition of unity, expression (2.66) is a suitable choice for scaling. Furthermore, we expect that this option will mirror the intention of the original scaling well in many cases.

The numerical results of the two highly heterogeneous tasks, presented in Fig. 2.1, support this claim. The first involves a stationary diffusion problem with a binary coefficient distribu-

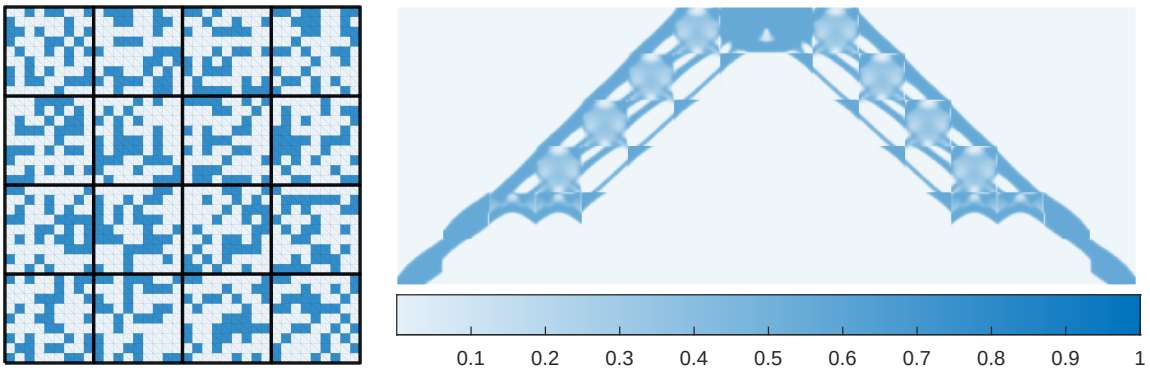


Figure 2.1: Left: Random binary voxel-based coefficient distribution within a stationary diffusion problem: $\frac{\rho_{\max}}{\rho_{\min}} = 10^6$. Dirichlet BCs are imposed on the whole $\partial\Omega$. Decomposition into 4×4 subdomains, $H/h = 9$. Vertex-based prior primal coarse space. **Right:** Messerschmitt-Bölkow-Blohm benchmark problem. $\frac{E_{\max}}{E_{\min}} = 10^6$. Decomposition into 6×16 subdomains, $H/h = 40$.

tion with a high heterogeneity ratio. Tables 2.1-2.2 store numerical results for a standard transformation of basis approach and a theoretically correct one, obtained using, e.g., deflation techniques. Although the choice among projector preconditioning, balancing and generalized transformation of basis depends on a personal preference, here we provide a comparison against a generalized transformation of basis because we observed the superior accuracy and robustness of generalized transformation of basis compared to the alternative methods relying on projections. Moreover, to guarantee a fair comparison, we compare the two methods using a relative L_2-error norm with a solution generated in the kth step compared to the reference solution obtained with a direct solver. In practice, we have to settle for an approximate error indicator, typically based on the norm of (preconditioned) residual. According to Tab. 2.1, the transformed and restricted scaling performs relatively well for both ρ and deluxe scaling. It is confirmed that the results for multiplicity scaling are identical, regardless of the method used. With ρ scaling, the condition number estimates as well as the initial and resulting L_2-error norm are relatively comparable. In the case of deluxe scaling, a significant difference

Stationary diffusion problem

	scaling	transformation	its.	ϵ_{L_2} (it. 1)	ϵ_{L_2} (it. 50)	κ_{est}
sToB	mult.	Eq. (2.66)	50	$2.82 \cdot 10^{-1}$	$6.23 \cdot 10^{-3}$	$1.84 \cdot 10^{5}$
gTob	mult.	correct	50	$2.82 \cdot 10^{-1}$	$6.23 \cdot 10^{-3}$	$1.84 \cdot 10^{5}$
sToB	ρ	Eq. (2.66)	50	$9.61 \cdot 10^{-2}$	$3.82 \cdot 10^{-5}$	$7.56 \cdot 10^{4}$
gTob	ρ	correct	50	$2.92 \cdot 10^{-1}$	$1.16 \cdot 10^{-5}$	$7.38 \cdot 10^{4}$
sToB	deluxe	Eq. (2.66)	50	$1.75 \cdot 10^{-1}$	$5.49 \cdot 10^{-5}$	$7.51 \cdot 10^{4}$
gTob	deluxe	correct	50	$2.71 \cdot 10^{-1}$	$4.90 \cdot 10^{-10}$	$1.60 \cdot 10^{4}$
sToB	ρ	none	50	$4.71 \cdot 10^{-1}$	$3.34 \cdot 10^{-2}$	$8.17 \cdot 10^{5}$
sToB	deluxe	none	50	$5.18 \cdot 10^{-1}$	$2.90 \cdot 10^{-2}$	$2.32 \cdot 10^{6}$

Table 2.1: Comparison of established transformation of scaling in standard transformation of basis for a scalar problem depicted in Fig. 2.1. Results for a vertex-based prior set Π' and enforced weighted averages after 50 iterations are shown. **Annotations: sToB/gToB** - standard/generalized transformation of basis, **transformation** - transformation of scaling, **its.** - iteration count (fixed), ϵ_{L_2}(it. k) - relative L_2-norm of difference between solution obtained in iteration k and a directly obtained reference solution.

in the convergence rate becomes apparent, as the reduction in relative norm is five orders of magnitude lower with a properly handled scaling. In any case, the transformation of scaling is undoubtedly essential. The non-transformed restricted scaling is less effective than constant scaling, which is inherently problematic in the context of heterogeneous problems with material discontinuities not aligned with interfaces.

The surprisingly subtle distinction between standard and generalized transformation of basis manifests itself in a linear elasticity problem with uniformly varying coefficients; see Tab. 2.2. It is evident that the solver performs similarly in both approaches. The most noticeable

Linear elasticity problem

	scaling	transformation	its.	ϵ_{L_2} (it. 1)	ϵ_{L_2} (it. 50)	κ
sToB	mult.	Eq. (2.66)	50	1.34	$6.45 \cdot 10^{-5}$	$8.96 \cdot 10^{1}$
gTob	mult.	correct	50	1.34	$6.45 \cdot 10^{-5}$	$8.96 \cdot 10^{1}$
sToB	ρ	Eq. (2.66)	50	$1.31 \cdot 10^{1}$	$7.75 \cdot 10^{-5}$	$2.96 \cdot 10^{2}$
gTob	ρ	correct	50	$9.25 \cdot 10^{-1}$	$6.91 \cdot 10^{-7}$	$4.76 \cdot 10^{1}$
sToB	deluxe	Eq. (2.66)	50	$2.12 \cdot 10^{-1}$	$9.98 \cdot 10^{-9}$	$2.00 \cdot 10^{1}$
gTob	deluxe	correct	50	$6.30 \cdot 10^{-1}$	$8.64 \cdot 10^{-9}$	$1.97 \cdot 10^{1}$
sToB	ρ	none	50	2.20	2.35	$3.92 \cdot 10^{5}$
sToB	deluxe	none	50	2.06	2.10	$3.33 \cdot 10^{5}$

Table 2.2: Comparison of established transformation of scaling in standard transformation of basis for elasticity problem depicted in Fig. 2.1. Results for a vertex-based prior set Π' and enforced weighted averages after 50 iterations are shown. **Annotations: sToB/gToB** - standard/generalized transformation of basis, **transformation** - transformation of scaling, **its.** - iteration count (fixed), ϵ_{L_2}(it. k) - relative L_2-norm of difference between solution obtained in iteration k and a directly obtained reference solution.

difference is now observed with the ρ scaling in terms of condition number, where the preconditioner in standard transformation of basis is slightly less effective. In summary, our observations generally do not indicate that the restricted scaling is particularly sensitive to any specific choice of scaling. However, it is not always the case that the difference in performance is negligible. Especially with adaptive techniques, some detrimental modes are often detected in the preconditioned system, and the condition number remains high.

A note to defend the existence of sToB is necessary; it has been first introduced in the context of imposing auxiliary constraints in the form of arithmetic averages, see [25, 29]. Being independent of the scaling used, the relations for constructing transformations for arithmetic averages are general, and once restricted, they also preserve the nodal character among the remaining dual unknowns. Specifically, the remaining dual unknowns represent nodal fluctuations from the imposed averages, i.e. the first- or second-order differences. In particular, we can set T_{Π^*} to store (non-normalized) edge moments of zeroth and possibly

first order and then directly construct $\mathsf{T}_{\overline{\Pi}^*}^{\perp} = \begin{bmatrix} \mathsf{I}_{\Delta^*\Delta^*} & -\mathsf{T}_{\Delta^*\Pi^*}\mathsf{T}_{\Pi^*\Pi^*}^{-\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$, which leads us to the following form of transformation

$$\begin{aligned} \mathsf{T}_{E_{ij}} &= \begin{bmatrix} \mathsf{T}_{\overline{\Pi}^*}^{\perp} & \mathsf{T}_{\Pi^*} \end{bmatrix} \\ &= \begin{bmatrix} \mathsf{I} & \mathsf{T}_{\Delta^*\Pi^*} \\ -\mathsf{T}_{\Pi^*\Pi^*}^{-1}\mathsf{T}_{\Delta^*\Pi^*}^{\mathsf{T}} & \mathsf{T}_{\Pi^*\Pi^*} \end{bmatrix} \qquad \text{s.t.} \quad \mathsf{T}_{\Pi^*}^{\mathsf{T}}\mathsf{T}_{\overline{\Pi}^*}^{\perp} = \mathsf{0}\,. \end{aligned} \tag{2.67}$$

Submatrix $\mathsf{T}_{\Pi^*\Pi^*}$ is certainly invertible unless the edge is aligned with one of the coordinate axes, then only a simple permutation is needed to ensure invertibility. Evidently, $\mathsf{T}_{E_{ij}}^{\mathsf{T}}\mathsf{T}_{E_{ij}} \neq \mathsf{I}$, but we can set a scaling only from the first term in the second line of Eq. (2.66), bypassing the need for a scaling transformation. Again, this is a feasible option to employ due to the preserved partition of unity. Nevertheless, a further deviation from the intended meaning of the scaling can be expected.

To summarize our description, standard transformation of basis (sToB) approach allows for enforcing non-nodal constraints. However, adaptively selected non-nodal modes, where the constraints are specifically dependent on prior scaling, the standard transformation of basis is not adequate anymore unless a constant scaling is used. The same holds for non-adaptive constraints as any other scaling than a constant one after transformation cannot carry all the information of the original scaling; a part of the scaling is irretrievably lost. The loss is due to the interaction between the chosen posterior modes and the remaining dual variables, because the application of the $\overline{\mathsf{D}}\mathsf{B}$ in $\mathsf{P}_{\overline{\mathsf{D}}} := \mathsf{B}^{\mathsf{T}}\overline{\mathsf{D}}\mathsf{B}$ operator does not preserve the continuity in the assembled posterior primal variables if transformed scaling is used. This continuity is enforced afterwards through a multiplication with B^{T}, yet a part of the scaling is neglected; we refer the reader for more details to [20, 42], where an illustrative counterexample is given. This violates the assumptions built in the theoretical background, where many condition number bounds relying upon the $\mathsf{P}_{\mathsf{D}} = \mathsf{B}^{\mathsf{T}}\mathsf{D}\mathsf{B}$ operator were successfully proven for the different adaptive approaches. Exactly the same problem would remain if, instead of transforming a scaling to the new basis, a reverse transformation from the new basis to the original one was provided by an appropriate modification of the continuity conditions in jump operator. Then, we could directly transform the jump operator as $\overline{\mathsf{B}} := \mathsf{B}_{\Delta'}\mathsf{T}_{\Delta'}$, to obtain jumps in the original basis. Generally, not a single continuity condition in $\overline{\mathsf{B}}_{\Delta^*}$ is now satisfied. Only transformed back to the new basis these conditions would be satisfied once again. Now, being the appropriate scaling constant, the issue would not arise. Unfortunately, this is not the case for ρ, stiffness or deluxe scaling, and, consequently, values after application of $\overline{\mathsf{P}}_{\mathsf{D}}$ are no longer zeroed in a posteriori primal variables — meaning that the continuity in Π^* is disrupted when transforming back to the new basis — unless the new basis is fully nodal (a specific case corresponding to a partial assembly). It is clear that the meaning of the scaling would be affected, resulting in inconsistency in the scaling. The key problem again lies in the restriction of the assigned part of the gap followed by the application of the local Schur complement; compare with Eq. (2.64).

2.5.4 Generalized Transformation of basis

We have now made a conceptual step towards a remedy: transformation of basis correctly handling scaling, known as a generalized transformation of basis (gToB). As mentioned in the previous sections, some components in the scaling can be neglected during a conversion of constraint(s) from a dual to primal set in a standard transformation of basis (sToB), which violates the intended character of the scaling. What has the most significant impact on the theory is the fact, that non-zero values can occur in in the posterior primal variables after the application of the localized P_{D} operator, unless some restrictive assumptions on the scaling and transformations hold.

Thus, the remedy is to enforce a posteriori primal variables in the space of Lagrange multipliers [42], as is the case in approaches using deflation. To this end, we make use of the

restriction operators R_{Π^*} introduced in Subsection 2.5.3 and unlike in sToB, we apply them not only to the vector of solution variables but to jump operators as well.

Again, we tacitly expect that the priori coarse space is limited to nodal constraints. As discussed in Subsection 2.5.3, this has been established as indispensable for preserving the desired interpretation of the scaling. However, when employing adaptive techniques, it is not necessary to adhere to this assumption; the primary requirement is to preserve the scaling once auxiliary constraints are imposed.

Starting from Eq. (2.56) we further express domain-wise quantities in the sense of Eqs.(2.52-2.53) and distinguish between $\overline{\mathsf{u}}_{\Delta^*}$ and $\overline{\mathsf{u}}_{\Pi^*} = \begin{bmatrix}\overline{\mathsf{u}}_{\Pi^*}^{(1),\mathsf{T}} & \dots & \overline{\mathsf{u}}_{\Pi^*}^{(N_\mathrm{s}),\mathsf{T}}\end{bmatrix}^\mathsf{T}$. For enhanced clarity, we henceforth proceed with unknowns segmented into sets $\mathrm{I}, \Delta^*, \Pi^*$, and Π', where the a posteriori chosen constraint modes in Π^* build on the coarse space and scaling. In generalized transformation of basis, a partial subassembly in $\overline{\mathsf{u}}_{\Pi^*}$ is not handled in a classical sense, i.e., by a direct elimination of a corresponding product $\widetilde{\overline{\mathsf{K}}}_{\Pi^*\Pi^*} = \mathsf{R}_{\Pi^*}^\mathsf{T}\overline{\mathsf{K}}_{\Pi^*\Pi^*}\mathsf{R}_{\Pi^*}$. Instead, the required continuity is achieved through a repeated process of (i) disassembly, (ii) subsequent averaging to ensure the continuity is preserved, and (iii) reassembly, which takes place in the jump operator. This, together with a special structure of the jump operator, allows us to work with quantities and a scaling both operating in the original basis.

Following the augmentation of unknowns, we can split transformation

$$\mathsf{T}_{\Delta'} =: \begin{bmatrix}\mathsf{T}_{\Delta^*} & \mathsf{T}_{\Pi^*}\end{bmatrix} \tag{2.68}$$

into two blocks accordingly. We remind that a priori sets are denoted by a prime $\bullet'$: primal constraints stored in Π' remain unchanged and for prior dual set we can write $\Delta' = \Delta^* \cap \Pi^*$, where asterisk marks the a posteriori (transformed) variables. Then, for every subdomain affected by at least one a posteriori constraint, we utilize jump operators $\overline{\mathsf{B}}_{\Delta^*}$ and $\widetilde{\overline{\mathsf{B}}}_{\Pi^*}$, which are composed of domain-wise parts

$$\overline{\mathsf{B}}_{\Delta^*}^{(s)} = \mathsf{B}_{\Delta'}^{(s)}\mathsf{T}_{\Delta^*}^{(s)} \tag{2.69}$$

$$\widetilde{\overline{\mathsf{B}}}_{\Pi^*}^{(s)} = \mathsf{B}_{\Delta'}^{(s)}\mathsf{T}_{\Pi^*}^{(s)}\mathsf{R}_{\Pi^*}^{(s)}\,. \tag{2.70}$$

Note that for subdomains for which no additional constraints are applied, $\overline{\mathsf{B}}_{\Delta^*}^{(s)}$ is given by properly transformed $\mathsf{B}_{\Delta'}^{(s)}$ and with $\widetilde{\overline{\mathsf{B}}}_{\Pi^*}^{(s)}$ being an empty matrix. Finally, we can rewrite Eq. (2.57) as

$$\begin{bmatrix}
\mathsf{K}_{\mathrm{I}\,\mathrm{I}} & \overline{\mathsf{K}}_{\mathrm{I}\,\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\mathrm{I}\,\Pi^*} & \widetilde{\mathsf{K}}_{\mathrm{I}\,\Pi'} & 0 \\
\overline{\mathsf{K}}_{\Delta^*\mathrm{I}} & \overline{\mathsf{K}}_{\Delta^*\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\Delta^*\Pi^*} & \widetilde{\overline{\mathsf{K}}}_{\Delta^*\Pi'} & \overline{\mathsf{B}}_{\Delta^*}^\mathsf{T} \\
\widetilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{I}} & \widetilde{\overline{\mathsf{K}}}_{\Pi^*\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\Pi^*\Pi^*} & \widetilde{\overline{\mathsf{K}}}_{\Pi^*\Pi'} & \widetilde{\overline{\mathsf{B}}}_{\Pi^*}^\mathsf{T} \\
\widetilde{\mathsf{K}}_{\Pi'\mathrm{I}} & \widetilde{\overline{\mathsf{K}}}_{\Pi'\Delta^*} & \widetilde{\overline{\mathsf{K}}}_{\Pi'\Pi^*} & \widetilde{\mathsf{K}}_{\Pi'\Pi'} & 0 \\
0 & \overline{\mathsf{B}}_{\Delta^*} & \widetilde{\overline{\mathsf{B}}}_{\Pi^*} & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\mathsf{u}_\mathrm{I} \\ \overline{\mathsf{u}}_{\Delta^*} \\ \widetilde{\overline{\mathsf{u}}}_{\Pi^*} \\ \widetilde{\mathsf{u}}_{\Pi'} \\ \lambda
\end{bmatrix}
=
\begin{bmatrix}
\mathsf{f}_\mathrm{I} \\ \overline{\mathsf{f}}_{\Delta'} \\ \widetilde{\overline{\mathsf{f}}}_{\Pi^*} \\ \widetilde{\mathsf{f}}_{\Pi'} \\ 0
\end{bmatrix}, \tag{2.71}$$

In the equation above we have denoted by $\widetilde{\bullet}$ quantities assembled in Π^* while the wide tilde accent $\widetilde{\bullet}$ those assembled Π' in order to distinguish between the two sets. Again, this system is usually condensed to Lagrange multipliers and the solution is sought for in an iterative manner, using

$$\mathsf{F}_{\mathrm{gToB}}\lambda = \mathsf{d}_{\mathrm{gToB}}, \tag{2.72}$$

with the modified left-hand side matrix $\mathsf{F}_{\mathrm{gToB}} \succeq 0$. The positive definiteness of $\mathsf{F}_{\mathrm{gToB}}$ can only occur in the case of an empty set Π^*, which means that it practically never happens. The rank deficiency arises in the condensation step of the averaging procedure in the block of a matrix corresponding to constraints given by Π^*. Clearly, the partial product resulting from the multiplication of the assembled Schur complement on Π^* in the new basis from the left by R_{Π^*} and from the right by $\mathsf{R}_{\Pi^*}^\mathsf{T}$ corresponds to a matrix with rows and columns identical in positions given by continuity conditions stored in Π^*. Transformed back to the original basis, this pair-wise structure is lost, but the rank deficiency naturally remains.

Let us note that, due to the jump operators being transformed themselves, the conditions within these operators retain its original meaning. Hence, the exact solution vector for Lagrange multipliers λ is the same for gToB as well as the standard FETI-DP algorithm in Eq. (2.56).

To give the reader a clearer understanding of the operations performed, we explicitly write down the individual terms arising from the recursive application of the block Gaussian elimination. As we believe the operations are then more apparent than in a formal, concise setting with one global Schur complement matrix. To avoid lengthy expressions, we use R^* collectively for I and Δ^*, but we still distinguish between set Π' and set Π^*.

Applying the block Gaussian elimination three times leads to a system operator

$$\begin{aligned}
\mathsf{F}_{\mathrm{gToB}} = \quad & \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\overline{\mathsf{B}}_{\mathrm{R}^*}^{\mathsf{T}} \\
& + \left(\tilde{\overline{\mathsf{B}}}_{\Pi^*} - \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\left(\tilde{\overline{\mathsf{B}}}_{\Pi^*} - \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)^{\mathsf{T}} \\
& + \left[\overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi'} + \left(\tilde{\overline{\mathsf{B}}}_{\Pi^*} - \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\left(\tilde{\tilde{\overline{\mathsf{K}}}}_{\Pi^*\Pi'} - \tilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi'}\right)\right] \\
& \left(\tilde{\tilde{\overline{\mathsf{S}}}}_{\mathrm{R}^*,\Pi'} - \tilde{\tilde{\overline{\mathsf{O}}}}_{\mathrm{R}^*,\Pi'\Pi^*}\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\tilde{\overline{\mathsf{O}}}_{\mathrm{R}^*,\Pi^*\Pi'}\right)^{-1} \\
& \left[\overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi'} + \left(\tilde{\overline{\mathsf{B}}}_{\Pi^*} - \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\left(\tilde{\tilde{\overline{\mathsf{K}}}}_{\Pi^*\Pi'} - \tilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi'}\right)\right]^{\mathsf{T}},
\end{aligned} \tag{2.73}$$

in which now only the first of the three terms can be executed fully in parallel because $\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}$ is block diagonal. In the second and third terms, subdomain-wise contributions have to be collected before the execution of a coarse solve, and distributed to individual subdomains afterwards . The corresponding right hand side follows as

$$\begin{aligned}
\mathsf{d}_{\mathrm{gToB}} = \quad & \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\overline{\mathsf{f}}_{\mathrm{R}^*} \\
& + \left(\tilde{\overline{\mathsf{B}}}_{\Pi^*} - \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\left(\tilde{\overline{\mathsf{f}}}_{\Pi^*} - \tilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\overline{\mathsf{f}}_{\mathrm{R}^*}\right) \\
& - \left[\overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi'} + \left(\tilde{\overline{\mathsf{B}}}_{\Pi^*} - \overline{\mathsf{B}}_{\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\left(\tilde{\tilde{\overline{\mathsf{K}}}}_{\Pi^*\Pi'} - \tilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi'}\right)\right] \\
& \left(\tilde{\tilde{\overline{\mathsf{S}}}}_{\mathrm{R}^*,\Pi'} - \tilde{\tilde{\overline{\mathsf{O}}}}_{\mathrm{R}^*,\Pi'\Pi^*}\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\tilde{\overline{\mathsf{O}}}_{\mathrm{R}^*,\Pi^*\Pi'}\right)^{-1} \\
& \left[\left(\tilde{\overline{\mathsf{f}}}_{\Pi'} - \tilde{\overline{\mathsf{K}}}_{\Pi'\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{f}}}_{\mathrm{R}^*}\right) - \left(\tilde{\tilde{\overline{\mathsf{K}}}}_{\Pi'\Pi^*} - \tilde{\overline{\mathsf{K}}}_{\Pi'\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\tilde{\overline{\mathsf{K}}}_{\mathrm{R}^*\Pi^*}\right)\tilde{\overline{\mathsf{S}}}_{\mathrm{R}^*,\Pi^*}^{-1}\left(\tilde{\overline{\mathsf{f}}}_{\Pi^*} - \tilde{\overline{\mathsf{K}}}_{\Pi^*\mathrm{R}^*}\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}^{-1}\overline{\mathsf{f}}_{\mathrm{R}^*}\right)\right].
\end{aligned} \tag{2.74}$$

In both expressions $\overline{\mathsf{S}}$ and $\overline{\mathsf{O}}$ are auxiliary notations as follows. $\overline{\mathsf{S}}$ denotes the Schur complement. Specifically, the two subscripts separated by a comma symbolize that the set of indices within the first subscript is condensed to the submatrix corresponding to DOFs attained in the second subscript, i.e.,

$$\overline{\mathsf{S}}_{X,Y} = \overline{\mathsf{K}}_{YY} - \overline{\mathsf{K}}_{YX}\mathsf{K}_{XX}^{-1}\mathsf{K}_{XY}.$$

The letter O represents a generally non-square product on off-diagonal sub-blocks of stiffness matrices, resulting in a rectangular matrix. It is defined in a similar fashion: the first subscript indicates the set of DOFs to be condensed out and the remaining two subscripts define the sets of DOFs along the first and second dimensions, respectively, of the matrix. That means

$$\overline{\mathsf{O}}_{A,BC} = \overline{\mathsf{K}}_{BC} - \overline{\mathsf{K}}_{BA}\overline{\mathsf{K}}_{AA}^{-1}\overline{\mathsf{K}}_{AC}.$$

Finally, the tilde accents whether the resulting product is assembled in any of the set Π^* or Π^*.

Now the reason behind keeping the sets Π' and Π^* separated becomes apparent. The construction of the preconditioner currently involves a partial assembly, which is handled via modification in the scaled jump operators. In the disassembly as well as in the assembly step, we have to cope with retaining a consistency of the preconditioner while accounting for the temporarily decoupled nature of the a posteriori primal variables. To do so, the scaled jump

operators are adjusted in a sense

$$\overline{\mathsf{B}}^{(s)}_{\mathrm{D},\Delta^*} = \mathsf{D}^{(s),\mathsf{T}}\overline{\mathsf{B}}^{(s)}_{\Delta^*} \qquad = \mathsf{B}^{(s)}_{\mathrm{D}}\mathsf{T}^{(s)}_{\Delta^*} \tag{2.75}$$

$$\tilde{\overline{\mathsf{B}}}^{(s)}_{\mathrm{D},\Pi^*} = \mathsf{D}^{(s),\mathsf{T}}\tilde{\overline{\mathsf{B}}}^{(s)}_{\Pi^*}(\mathsf{R}^{\mathsf{T}}_{\Pi^*}\mathsf{R}_{\Pi^*})^{-1} = \mathsf{B}^{(s)}_{\mathrm{D}}\mathsf{T}^{(s)}_{\Pi^*}\mathsf{R}^{(s)}_{\mu,\Pi^*} \tag{2.76}$$

with a domain-wise defined multiplicity-scaled prolongation operators $\mathsf{R}^{(s)}_{\mu,\Pi^*} := \mathsf{R}^{(s),\mathsf{T}}_{\Pi^*}(\mathsf{R}^{\mathsf{T}}_{\Pi^*}\mathsf{R}_{\Pi^*})^{-1}$ needed for construction of

$$\mathsf{R}^{\mathsf{T}}_{\mu,\Pi^*} := \begin{bmatrix}\mathsf{R}^{(1),\mathsf{T}}_{\mu,\Pi^*} & \dots & \mathsf{R}^{(N_s),\mathsf{T}}_{\mu,\Pi^*}\end{bmatrix}^{\mathsf{T}}. \tag{2.77}$$

As a consequence of the partial assembly involved, even the preconditioning step is no longer perfectly parallelizable, i.e., we do not have an expression that involves solely independent local solves as, for instance, is the case of in Eq. (2.22). Preconditioner takes the form

$$\mathsf{M}^{-1}_{\mathrm{D,gToB}} = \begin{bmatrix}\overline{\mathsf{B}}_{\mathrm{D},\Delta^*} & \tilde{\overline{\mathsf{B}}}_{\mathrm{D},\mu,\Pi^*}\end{bmatrix}\begin{bmatrix}\overline{\mathsf{S}}_{\mathrm{I},\Delta^*} & \tilde{\overline{\mathsf{O}}}_{\mathrm{I},\Delta^*\Pi^*}\\ \tilde{\overline{\mathsf{O}}}_{\mathrm{I},\Delta^*\Pi^*} & \tilde{\overline{\mathsf{S}}}_{\mathrm{I},\Pi^*}\end{bmatrix}\begin{bmatrix}\overline{\mathsf{B}}^{\mathsf{T}}_{\mathrm{D},\Delta^*}\\ \tilde{\overline{\mathsf{B}}}^{\mathsf{T}}_{\mathrm{D},\mu,\Pi^*}\end{bmatrix} \tag{2.78}$$

which, in expanded form, reads

$$\begin{aligned}\mathsf{M}^{-1}_{\mathrm{D,gToB}} = &\begin{bmatrix}\mathsf{B}_{\mathrm{D}}\mathsf{T}_{\Delta^*} & \mathsf{B}_{\mathrm{D}}\mathsf{T}_{\Pi^*}\mathsf{R}_{\mu,\Pi^*}\end{bmatrix}\\ &\begin{bmatrix}\overline{\mathsf{K}}_{\Delta^*\Delta^*} - \overline{\mathsf{K}}_{\Delta^*\mathrm{I}}\mathsf{K}^{-1}_{\mathrm{II}}\overline{\mathsf{K}}_{\mathrm{I}\,\Delta^*} & (\overline{\mathsf{K}}_{\Delta^*\Pi^*} - \overline{\mathsf{K}}_{\Delta^*\mathrm{I}}\mathsf{K}^{-1}_{\mathrm{II}}\overline{\mathsf{K}}_{\mathrm{I}\,\Pi^*})\mathsf{R}_{\Pi^*}\\ \mathsf{R}^{\mathsf{T}}_{\Pi^*}(\overline{\mathsf{K}}_{\Pi^*\Delta^*} - \overline{\mathsf{K}}_{\Pi^*\mathrm{I}}\mathsf{K}^{-1}_{\mathrm{II}}\overline{\mathsf{K}}_{\mathrm{I}\,\Delta^*}) & \mathsf{R}^{\mathsf{T}}_{\Pi^*}(\overline{\mathsf{K}}_{\Pi^*\Pi^*} - \overline{\mathsf{K}}_{\Pi^*\mathrm{I}}\mathsf{K}^{-1}_{\mathrm{II}}\overline{\mathsf{K}}_{\mathrm{I}\,\Pi^*})\mathsf{R}_{\Pi^*}\end{bmatrix}\\ &\begin{bmatrix}\mathsf{T}^{\mathsf{T}}_{\Delta^*}\mathsf{B}^{\mathsf{T}}_{\mathrm{D}}\\ \mathsf{R}^{\mathsf{T}}_{\mu,\Pi^*}\mathsf{T}^{\mathsf{T}}_{\Pi^*}\mathsf{B}^{\mathsf{T}}_{\mathrm{D}}\end{bmatrix}.\end{aligned} \tag{2.79}$$

The transformed matrices $\overline{\mathsf{K}}$ are commonly not assembled. In practice, we rather take advantage of the orthonormality of T and perform a forward or backward substitution on a factorized $\mathsf{S}_{\Delta'}$. We avoid explicitly assembling the transformed matrices in our implementation to reduce the risk of accumulation of possibly unidirectional propagation of errors eventually posed by the inaccurate computation of transformed ill-conditioned matrices. The assembly of transformed matrices can be time-consuming when the transformations $\mathsf{T}^{(s)}$ lose their sparsity, for instance, when face-related constraints on meshes with fine resolution are enforced in three dimensions. For completeness, we also mention an alternative: to solve a set of local auxiliary saddle point problems introduced with the use of new Lagrange multipliers [29] to obtain the desired vectors transformed to the original basis. Then, only a matrix-vector multiplication with (transposed) $\mathsf{T}^{(s)}$ and a resulting product is needed to obtain the outcome in the desired basis. Nevertheless, we only use the standard approach.

It is important to emphasise that the statement of the generalized transformation of basis presented above deviates from what is suitable from an implementation perspective. In the code, it does not make much sense to separate the coarse problem into a priori and a posteriori parts. As can be seen in Eq. (2.73), a three-level formulation which does not exploit the repetitive occurrence of some of the products requires nine local solves with $\overline{\mathsf{K}}_{\mathrm{R}^*\mathrm{R}^*}$ per iteration. In fact, unifying Π^* and Π' would lead to only three. Hence, we only wanted to highlight the relevance of the products involved. Because the literature on the gToB is limited, and the available sources understandably focus mostly on highly efficient parallel implementations or on theoretical aspects of this approach, which often determines the notation used, we wanted to provide the exposition of gToB with distinct Π' and Π^* in intentionally detailed perspective on the matter.

Chapter 3

Coarse Space Enhancements

First, we provide a brief overview of the most widely used approaches to enhancing the coarse space: from simple and generic approaches that are considered to be useful in specific (yet relatively common) cases to very sophisticated approaches that involve solving local generalized eigenvalue problems on the interfaces. In fact, there are two main directions we can trace.

In this work, we refer to these two main branches as non-adaptive and adaptive. Non-adaptive, also known as heuristic, work with a very limited amount of data and are usually cost-effective. These approaches are represented by arithmetic and weighted averages and a Frugal approach. The adaptive approaches, on the other hand, construct auxiliary constraints that are problem-specific, i.e., they depend on the distribution of coefficients of the underlying PDE, boundary conditions, and provided scaling. Such constraints are usually sought through solving a Generalized EigenValue Problem (GEVP) on the interface between a pair of adjacent subdomains. This is evidently linked with a significant computational overhead, as it represents a computationally expensive operation. However, the constraints acquired in this manner are highly beneficial because reliable estimators or even proven relations for the condition number bound are provided for many such GEVP formulations. By setting the approaches' threshold, adaptive strategies allow the user to indirectly control the resulting condition number; in contrast, non-adaptive approaches usually provide control over the number of constraints, but the effectiveness of these heuristic constraints remains uncertain. If we consider all the eigenmodes corresponding to eigenvalues exceeding a certain tolerance, pleasant benefits can be achieved when dealing with very challenging problems.

The existing strategies for the construction of auxiliary constraints then serve as a stepping stone for the two approaches developed by the author, which are presented at the end of this section. The first approach involves a heuristic construction of a reduced basis for the solution of GEVPs, an approach that is in our understanding adaptive. The second approach involves enriching the set of primal variables with a few selected primal nodes. Although this approach reflects the specifics of the problem, for simplicity reasons we kept it independent of the scaling. Hence, this is a typical representative of a heuristic strategy. We do not merely introduce these strategies known from the literature; we also discuss their suitability, identifying cases where they are effective and where they tend to fail, and propose slight modifications aimed at enhancing the robustness of the method provided by these heuristics.

3.1 Weighted averages

The concept of enforcing conditions in the form of arithmetic averages over selected boundary entities dates back to the early phase of the method [12]. At that time, the averages primarily targeted three-dimensional tasks, for which the primal constraints binding only selected nodes were not sufficiently robust, and more general edge- and face-related constraints based solely on geometrical information restored the robustness, albeit only for problems that are either homogeneous in each subdomain or where heterogeneity appears only within the subdomain

(no discontinuities occur along subdomain interfaces). The first case can be effectively resolved through scaling, and the second case generally does not significantly affect the conditioning of the system. However, decompositions following material distribution are often not feasible, for example, for continuously varying coefficients or when the decomposition may lead to subdomains with bad aspect ratios.

The robustness achieved by using averages, or zeroth- and potentially first-order moments, is due to their ability to handle rigid body modes. In fact, these averages, or zeroth- and potentially first-order moments in general, are closely associated with the null space components of a physical body. A standard illustration involves two adjacent subdomains made of a very stiff material, which are surrounded from all sides by a medium with a significantly lower material coefficient. If the two subdomains were not assembled at any of the common nodes, there would be at least three or six constraints, depending on the dimensionality, needed to capture the relative rigid body modes of the two substructures. However, when heterogeneity comes into play, the benefit of these averages deteriorates very quickly, and such constraints are no longer sufficient. As a remedy, Klawonn and Rheinbach in [27] suggested the use of weighted averages

$$\frac{\sum_{x_k \in E_{ij}} \hat{\alpha}(x_k) u_m(x_k)}{\sum_{x_k \in E_{ij}} \hat{\alpha}^2(x_k)} , \quad m = 1, \ldots, n_{\text{wa}} . \tag{3.1}$$

Originally, for mechanical problems, only translational weighted averages were proposed, for which n_{wa} corresponds to the dimension of the problem. Weighted rotations were later introduced in [18] such that the averages on edge E_{ij} read

$$\frac{\hat{\mathsf{r}}_{E_{ij},m}^{\mathsf{T}} \mathsf{u}_{|E_{ij}}}{\hat{\mathsf{r}}_{E_{ij},m}^{\mathsf{T}} \hat{\mathsf{r}}_{E_{ij},m}} , \quad m = 1, \ldots, n_{\text{wa}} , \tag{3.2}$$

where $\hat{\mathsf{r}}_{E_{ij}}$ denotes the weighted rigid body modes of a floating edge. In particular, for two-dimensional problem of elasticity we have three modes

$$\mathsf{r}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \qquad \mathsf{r}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \qquad \mathsf{r}_3 = \begin{bmatrix} -\boldsymbol{x}_2^1 + \hat{\boldsymbol{x}}_2 \\ \boldsymbol{x}_1^1 - \hat{\boldsymbol{x}}_1 \\ -\boldsymbol{x}_2^2 + \hat{\boldsymbol{x}}_2 \\ \boldsymbol{x}_1^2 - \hat{\boldsymbol{x}}_1 \\ \vdots \\ -\boldsymbol{x}_2^n + \hat{\boldsymbol{x}}_2 \\ \boldsymbol{x}_1^n - \hat{\boldsymbol{x}}_1 \end{bmatrix} ; \tag{3.3}$$

while in the case of a diffusion problem, only one mode suffices

$$\mathsf{r}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} . \tag{3.4}$$

These rigid body modes $\mathsf{r}_{E_{ij}}$ restricted to an edge E_{ij} are pointwise scaled with corresponding value of $\hat{\alpha}(x)$ to obtain the weighted average constraint mode $\hat{\mathsf{r}}_{E_{ij}}$, needed to enforce constraints of the form

$$\hat{\mathsf{r}}_{E_{ij}}^{\mathsf{T}} \mathsf{u}_{|E_{ij}}^{(i)} = \hat{\mathsf{r}}_{E_{ij}}^{\mathsf{T}} \mathsf{u}_{|E_{ij}}^{(j)} . \tag{3.5}$$

For instance, values of $\hat{\mathsf{r}}_{3,E_{ij}}$ at a nodal point $x_k \in E_{ij}$ for a rotational mode are set as $\begin{bmatrix} \hat{\alpha}(x_k)(-\boldsymbol{x}_2^k + \hat{\boldsymbol{x}}_2) \\ \hat{\alpha}(x_k)(\ \boldsymbol{x}_1^k - \hat{\boldsymbol{x}}_1)) \end{bmatrix}$. In the expressions above, the centre of rotation $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2)$ coincides with the geometric centre of the edge. The weighting coefficients $\hat{\alpha}(x)$ are set as a maximum material coefficient at support $\omega(x)$ of a nodal point x

$$\hat{\alpha}(x) = \max_{\boldsymbol{y} \in \omega(x)} \alpha(y) \tag{3.6}$$

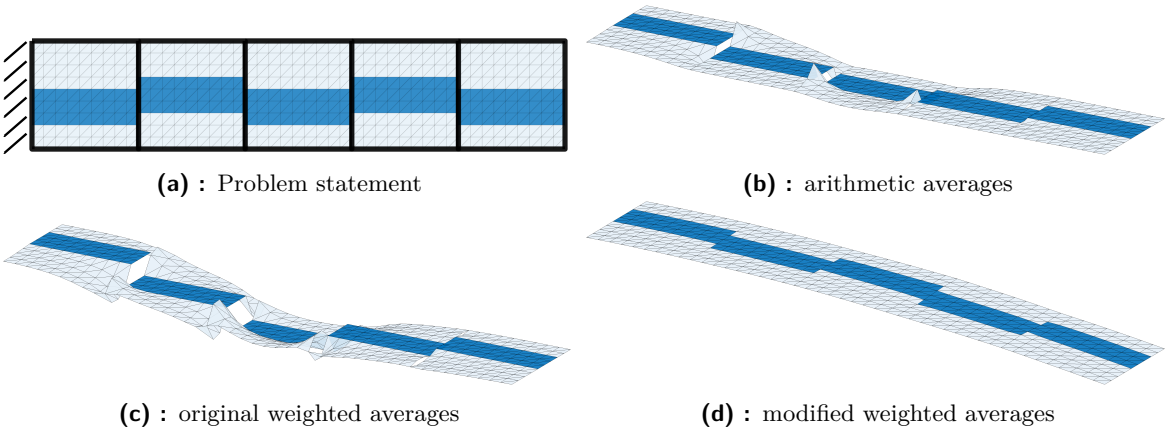
This enhancement for heterogeneous problems is beneficial for handling a single high-coefficient segment intersecting a given geometrical entity (either edge of face). For the conventional construction of weighting as stated in Eq. (3.6), it is assumed that a substantially stiffer part, such as a rigid channel, traverses soft regions in two subdomains with aligned material discontinuities along the interface. In such a case, weighted averages do a remarkable job. Since the weighted averages do not introduce any additional cost compared to the arithmetic ones, it is convenient to use them in a default setting.

3.1.1 Proposed modified weighted averages

We believe that a generally more suitable choice for the scaling weight could be as follows

$$\hat{\alpha}(x) = \min_{l\in\{i,j\}} \hat{\alpha}^{(l)}(x) := \min_{l\in\{i,j\}} \left(\max_{y\in\omega(x)\cap\Omega_l} \alpha(y) \right). \tag{3.7}$$

With such a choice, the weighted averages should be able to deal with discontinuities that are not perfectly aligned across the interface.



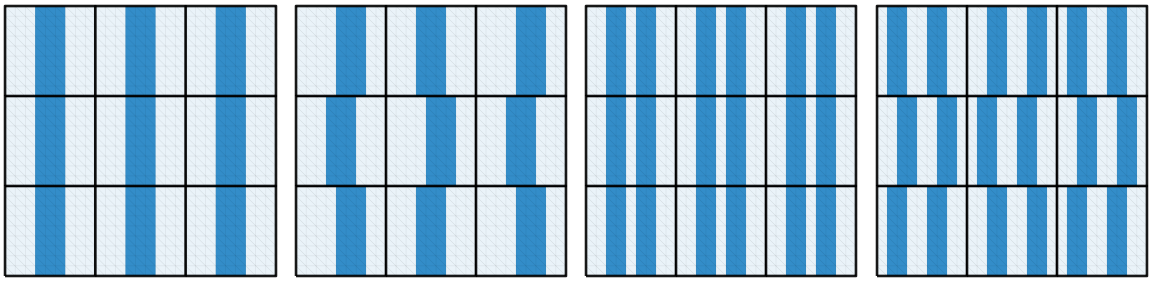
(a) : Problem statement **(b) :** arithmetic averages

(c) : original weighted averages **(d) :** modified weighted averages

Figure 3.1: Illustrative example showcasing the impact of averages. A problem setup with high-coefficient channels shifted at the interfaces is shown in **(a)**. Dirichlet BCs are imposed on the left edge of $\partial\Omega$. The next three images show intermediate solutions in the second iterations of CG: **(b)** arithmetic averages, **(b)** classic weighted averages, **(b)** modified weighted averages.

Let us provide an example of a steady-state heat equation problem consisting of five subdomains with straight conductive channels that are shifted by one element at each interface as shown in Fig. 3.1a. The temperature at the end of one of the channels is kept fixed, and the channels are subjected to a constant source term. With the classical vertex-based coarse space and ρ scaling, the initial coarse space lacks one additional constraint at each of the edges.

Figs 3.1b-d show the second iterations of the CG method. Clearly, the solution is severely discontinuous at all interfaces for both arithmetic and original weighted averages. The intended connectivity of the dark blue channels is clearly not achieved with the standard weighted averages. It is violated by the jumps that arise at conductive elements; this is due to a very low energy these gradients have. In contrast, the second iteration with a slightly adjusted version of the weighted averages seems to approach the accurate solution. Please note that the temperature values in Fig. 3.1d are scaled up 100 times compared to the previous two cases, to emphasise the obtained accuracy of the approximated solution. We further investigate the performance in a case of problems shown in Fig. 3.2. In the second case, we can observe behavior similar to the previous case. Here, the significance of the proposed modification is additionally supported by quantitative numerical results, provided in Tab. 3.1. In the first problem of the case, the weighted averages are able to reduce the condition number to 1.10. In this case the proposed modification performs identically to the classic weighted averages, since there is no difference in coefficient profiles across the interface. In test problem (2),



(a) : Test problem 1 **(b) :** Test problem 2 **(c) :** Test problem 3 **(d) :** Test problem 4

Figure 3.2: The second test case computing: four different coefficient distributions for a stationary diffusion problem. In the figure, deep blue represents coefficient $\rho_{max} = 10^6$, light blue ( )$\rho_{min} = 1$. Dirichlet BCs are imposed on the whole $\partial\Omega$. Vertex-based primal coarse space and ρ scaling.

Condition numbers for different coefficient distributions

Test problem		**1** (Fig. 3.2a)	**2** (Fig. 3.2b)	**3** (Fig. 3.2c)	**4** (Fig. 3.2d)
arithmetic	κ	$2.51 \cdot 10^4$	$8.71 \cdot 10^4$	$9.65 \cdot 10^3$	$1.14 \cdot 10^5$
	it.	9	12	16	19
weighted	κ	**1.10**	$3.05 \cdot 10^4$	$5.98 \cdot 10^3$	$1.14 \cdot 10^5$
	it.	2	10	10	19
Eq. 3.7	κ	**1.10**	**1.25**	$5.98 \cdot 10^3$	$1.13 \cdot 10^5$
	it.	2	3	10	14

Table 3.1: Numerical results for four test problems. Stopping criteria: $\epsilon_{L_2} \leq 1 \cdot 10^{-6}$

however, the classic weighted averages fail to reduce the condition number for the same reason that arithmetic averages fail in case (1). The proposed adjusted version is able to handle this scenario effectively, leading to the condition number reduced by four orders of magnitude. Eventually, for problems involving multiple channels, as is the case of test problems (3) and (4), a single constraint is not sufficient. Nevertheless, it can be observed from the iteration count that the proposed weighting converges more rapidly towards the reference solution in test problem (4), in which the two channels are mutually shifted. Hence, weighted averages serve as an efficient tool for coarse-space correction when dealing with heterogeneities. Being independent of the scaling or the construction of the jump operators, they could be a suitable option for the prior coarse space as well.

The substantial benefit of the weighted averages is their very inexpensive setup. No information about the coefficient distribution inside the subdomains is needed for their construction, only the coefficient profiles at the shared interface of adjacent subdomains. Unlike some of generally tough-to-picture strategies for construction of auxiliary constraints, averages (or, in general, low-order moments) comprise one's intuitive expectations and their behaviour is very predictable.

3.2 Adaptive coarse spaces

The adaptive coarse spaces have recently attracted significant attention [21, 23], especially when addressing highly heterogeneous problems [31]. These adaptive approaches most commonly involve a solution of generalized eigenvalue problems at the interfaces between neighbouring domains, thus minimising the required communication among otherwise well-parallelizable processes. One advantage of the adaptive techniques is their sensitivity to specific problem- and scaling-dependent poorly conditioned parts of the original coarse space. As a result, the adaptive techniques facilitate enriching the coarse space with locally acquired modes that have a significant impact on limiting the resulting condition number. Despite being considered

relatively computationally expensive, they have become increasingly popular over the last few years due to their robustness. Numerous adaptive approaches have undergone theoretical analysis [43], often resulting in an upper bound on condition numbers regardless of mesh size and material heterogeneity [19, 24].

Several adaptive approaches have been introduced over the years, for more details see, e.g., [35, 43]. A potential approach based on the localized P_D operator and parallel sum of matrices exists [43]. This approach leads to the solution of a small eigenproblem, restricted only to the dual DOFs on the part of the interface shared by the two subdomains. As demonstrated by comparing local spectra of different adaptive DD approaches, this alternative formulation could potentially result in a smaller number of acquired adaptive constraints [17]. However, in this thesis, we exclusively focus on the most commonly used adaptive approach, which is briefly described in the subsequent subsection.

Also note that all the strategies for constructing admissible constraints in order to set up a robust coarse space are scaling-dependent.

3.2.1 Eigenvalue problem by Mandel and Sousedík

A pioneering work in the field of adaptive approaches in the context of BBDC and FETI-DP methods dates back to 2007, when Mandel and Sousedík [35] stated that the condition number bound of the preconditioned system $\mathsf{M}^{-1}\mathsf{F}$ based on the P_D operator, satisfying

$$\kappa(\mathsf{M}^{-1}\mathsf{F}) \leq \sup_{\mathsf{w}\in\widetilde{W}} \frac{\|\mathsf{P}_\mathrm{D}\mathsf{w}\|^2_{\mathsf{S}_\Gamma}}{\|\mathsf{w}\|^2_{\mathsf{S}_\Gamma}}\,, \tag{3.8}$$

is limited by the maximal eigenvalue of the system operating on W, yet projected onto the space of continuous primal constraints $\widetilde{W}$

$$\Pi\mathsf{B}_\Gamma^\mathsf{T}\mathsf{B}_{\Gamma,\mathrm{D}}\mathsf{S}_\Gamma\mathsf{B}_{\Gamma,\mathrm{D}}^\mathsf{T}\mathsf{B}_\Gamma\Pi\mathsf{w} = \lambda\,\Pi\mathsf{S}_\Gamma\Pi\,\mathsf{w} \tag{3.9}$$

through projections $\Pi : W \to \widetilde{W}$.

Driven by Eq. (3.9), which is not suitable for practical implementations due to its global nature, the authors of [35] presented a localized estimate. Specifically, the condition number indicator is determined as the maximum eigenvalue of a set of localized problems corresponding to Eq. (3.9), each defined on an interface between a pair of adjacent subdomains. Each localized generalized eigenvalue problem (GEVP) reads:

$$\text{Find } \mathsf{w}_{ij} \in \widetilde{W}_{ij} \qquad \mathsf{B}_{ij}^\mathsf{T}\mathsf{B}_{\mathrm{D},ij}\mathsf{S}_{ij}\mathsf{B}_{\mathrm{D},ij}^\mathsf{T}\mathsf{B}_{ij}\mathsf{w}_{ij} = \lambda_{ij}\mathsf{S}_{ij}\mathsf{w}_{ij} \tag{3.10}$$

where the solution space $\widetilde{W}_{ij}$ is a space of functions continuous in the primal variables shared by the two subdomains $\Omega^{(i)}$ and $\Omega^{(j)}$. In the equation above, S_{ij} stands for a completely decoupled Schur complement

$$\begin{bmatrix} \mathsf{S}_{\Gamma\Gamma}^{(i)} & 0 \\ 0 & \mathsf{S}_{\Gamma\Gamma}^{(j)} \end{bmatrix}$$

and B_{ij} stores local continuity constraints on $\Gamma^{(i)}\cap\Gamma^{(j)}$, obtained as a submatrix of $\begin{bmatrix}\mathsf{B}_\Gamma^{(i)} & \mathsf{B}_\Gamma^{(j)}\end{bmatrix}$ preserving only rows with one $+1$ and one -1 value. The local version of the scaled jump operator $\mathsf{B}_{\mathrm{D},ij}$ is then obtained in a completely analogous fashion. Afterwards, we can write $\mathsf{P}_{\mathrm{D},ij} = \mathsf{B}_{\mathrm{D},ij}^\mathsf{T}\mathsf{B}_{ij}$. Moreover, we let $\mathsf{R}_{\Gamma,ij}$ be a submatrix of $\begin{bmatrix}\mathsf{R}^{(i)} \\ \mathsf{R}^{(j)}\end{bmatrix}$ restricted to values on $\Gamma^{(ij)}$

$$\mathsf{R}_{\Gamma,ij} = \begin{bmatrix} \mathsf{I}_\Delta^{(i)} & 0 & 0 \\ 0 & 0 & \mathsf{R}_\Pi^{(i)} \\ 0 & \mathsf{I}_\Delta^{(j)} & 0 \\ 0 & 0 & \mathsf{R}_\Pi^{(j)} \end{bmatrix}.$$

With the restricted assembly operator $\mathsf{R}_{\Gamma,ij}^\mathsf{T}$ at hand, the solution of Eq. (3.10) on $\widetilde{W}_{ij}$ is straightforward making use of the partial subassembly. In a parallel code, we aim to benefit from independent local solves, in which, ideally, no matrix products have to be assembled.

Thus, instead of a finite element assembly in the given variables, we adopt an additional Euclidean projection $\Pi_{ij} : W_{ij} := W_i \times W_j \to \widetilde{W}_{ij}$. This projection is simply obtained as

$$\Pi_{ij} = \mathsf{R}_{\Gamma,ij}(\mathsf{R}_{\Gamma,ij}^{\mathsf{T}}\mathsf{R}_{\Gamma,ij})^{-1}\mathsf{R}_{\Gamma,ij}^{\mathsf{T}} \tag{3.11}$$

Clearly, it holds $\mathrm{Range}(\Pi_{ij}) = \widetilde{W}_{ij}$ and $(\mathrm{Kernel}\,\mathsf{S}_{ij})^{\perp} \subseteq \widetilde{W}_{ij}$. The matrix Π_{ij} remains an identity on all but DOFs attached in $\Pi^{(i)} \cap \Pi^{(j)}$. Thus, in practice, this matrix is never explicitly computed by means of the expression above. Instead, multiplication with Π_{ij} can be handled with a few cheaply obtained vector-vector multiplications. We also note that the application of projection matrices does not violate the sparsity pattern of S_{ij} because a maximum of two values communicate at a time.

With the newly introduced quantities, Eq. (3.10) can be rewritten more suitably for an efficient numerical solution [35]:

$$\text{Find } \mathsf{w}_{ij} \in \mathrm{Range}\,\Pi_{ij} \;:\quad \Pi_{ij}\mathsf{P}_{\mathrm{D},ij}^{\mathsf{T}}\mathsf{S}_{ij}\mathsf{P}_{\mathrm{D},ij}\Pi_{ij}\mathsf{w}_{ij} = \lambda_{ij}\Pi_{ij}\mathsf{S}_{ij}\Pi_{ij}\mathsf{w}_{ij} \tag{3.12}$$

We will refer to the left-hand side of the equation above as the high-energy side, while the right-hand side will be interchangeably called the low-energy side. With Eq. (3.12) at hand, we wish to reduce the condition number of the preconditioned system by enforcing adaptive constraints

$$\overbrace{\mathsf{w}_{ij}^{m,\mathsf{T}}\mathsf{P}_{\mathrm{D},ij}^{\mathsf{T}}\mathsf{S}_{ij}\mathsf{B}_{\mathrm{D},ij}^{\mathsf{T}}}^{\mathsf{c}_{ij}^{m,\mathsf{T}}}\mathsf{B}_{ij}\mathsf{v}_{ij} = 0 \quad \forall m : \lambda_{ij}^{m} \geq \mathsf{tol}\,.$$

Unless the superstructure of two bodies $\Omega^{(i)}$ and $\Omega^{(j)}$ joined at primal vertices is not positioned on a part of Dirichlet boundary sufficient to prevent common or relative rigid body modes of this superstructure, the right hand side of Eq. (3.12) in only positive-semidefinite. For most of the eigensolvers a positive definite right-hand side matrix is needed. For this reason, a second l_2-orthogonal projection $\overline{\Pi}$ onto $\mathrm{Range}\,(\Pi_{ij}\mathsf{S}_{ij}\Pi_{ij} + t(\mathsf{I} - \Pi_{ij}))$ is used to ensure positive definiteness of the right-hand side of the generalized eigenvalue problem. So, numerically, we solve

$$(\Pi_{ij}\mathsf{P}_{\mathrm{D},ij}^{\mathsf{T}}\mathsf{S}_{ij}\mathsf{P}_{\mathrm{D},ij}\Pi_{ij})\mathsf{w}_{ij,k} = \lambda_{ij,k}(\overline{\Pi}_{ij}(\Pi_{ij}\mathsf{S}_{ij}\Pi_{ij} + t(\mathsf{I} - \Pi_{ij}))\overline{\Pi}_{ij} + \bar{t}(\mathsf{I} - \overline{\Pi}_{ij}))\mathsf{w}_{ij,k} \tag{3.13}$$

with parameters $t, \bar{t} > 0$. Once again, we face a problem of finding an orthonormal basis of a given subspace, $\mathrm{Kernel}\Big(\Pi_{ij}\mathsf{S}_{ij}\Pi_{ij} + t(\mathsf{I} - \Pi_{ij})\Big)$, because then the application of $\overline{\Pi}$ can be replaced by just a few vector-vector multiplications. Luckily, this can be achieved quite cheaply if both geometrical information about nodal coordinates and data containing Dirichlet conditions are accessible. For more technical details about effective parallel implementation of this eigenproblem, we refer the reader to [22].

In [43], Rheinbach et al. proved for elliptic PDEs in two dimensions that the preconditioned system with enforced all constraint vectors c_{ij}^{m} corresponding to $\mu_{ij}^{m} \geq \mathsf{tol}$ satisfies

$$\kappa(\mathsf{M}_{\mathrm{MS,bal}}^{-1}\mathsf{F}) \leq N_{\mathcal{E}}^{2}\mathsf{tol}\,,$$

where $N_{\mathcal{E}}$ stands for the maximum number of edges of a subdomain. Hence, the condition number is independent of the mesh resolution, coefficient distribution and its contrast. The same holds if generalized transformation of basis approach or projector preconditioning is adopted [20, 28]. In our experiments, we observed that the highest local eigenvalue remaining in the system generally serves as a reliable condition number estimator.

3.3 Frugal Approach

To keep the thesis self-contained, we recall a frugal approach here. The idea behind the Frugal approach of Heinlein et al. [18] is closely related to adaptive approaches, particularly that introduced by Mandel and Sousedík [35]. In fact, frugal constraints are trying to mimic the most harmful eigenmodes without the need to solve any generalized eigenvalue problem, providing a computationally economic alternative to the original adaptive coarse

space. We later build on some of its concepts, and we suggest minor adjustments and discuss the limitations and suitability of the frugal constraints.

According to [18], we search for interface modes $\mathsf{v}_{\mathsf{ij}}^{\Gamma}$ for which alternative formulation close to [35] holds:

$$\left\langle \mathcal{H}_{\mathrm{I}}\left(\mathsf{P}_{\mathrm{D}_{ij}}\mathsf{w}_{ij}^{\Gamma}\right), \mathsf{K}_{ij}\ \mathcal{H}_{\mathrm{I}}\left(\mathsf{P}_{\mathrm{D}_{ij}}\mathsf{v}_{ij}^{\Gamma}\right)\right\rangle = \mu_{ij}\left\langle \mathcal{H}_{\mathrm{I}}\left(\mathsf{w}_{ij}^{\Gamma}\right), \mathsf{K}_{ij}\ \mathcal{H}_{\mathrm{I}}\left(\mathsf{v}_{ij}^{\Gamma}\right)\right\rangle, \tag{3.14}$$

where $\mathsf{K}_{ij} = \mathrm{diag}(\mathsf{K}_i, \mathsf{K}_j)$ and $\mathcal{H}_{\mathrm{I}}$ denotes minimum energy extensions from $\Gamma^{(i)} \cup \Gamma^{(j)}$ to the interior of the subdomains (possibly including nodes on $\partial\Omega_N$). Here, slightly abusing the notation, we understand the minimum energy operators as expressions defined on discretized quantities. The piece-wise discrete harmonic extension $\mathsf{w}^{(l)} = \mathcal{H}_{\mathrm{I}}^{l}(\mathsf{w}_{\Gamma}^{(l)})$ with respect to the inner product defined by $\mathsf{K}^{(l)}$ satisfies [55]

$$\left\langle \mathcal{H}_{\mathrm{I}}^{l}(\mathsf{v}_{\Gamma}^{(l)}),\ \mathsf{K}^{(l)}\ \mathcal{H}_{\mathrm{I}}^{l}(\mathsf{v}_{\Gamma}^{(l)})\right\rangle = \min_{\mathsf{v}^{(l)}|_{\Gamma^{(l)}}=\mathsf{v}_{\Gamma}^{(l)}} \left\langle \mathsf{v}^{(l)},\ \mathsf{K}^{(l)}\mathsf{v}^{(l)}\right\rangle.$$

The application of a discrete harmonic extension operator $\mathcal{H}_{\mathrm{I}}^{l}$ from a boundary to the interior of a substructure is closely related to the Schur complement $S_{\Gamma}^{(l)}$. In a matrix representation, the interior values $\mathsf{w}_{\mathrm{I}}^{(l)}$ of $\mathsf{w}^{(l)}$ are completely defined by values of $\mathsf{w}_{\Gamma}^{(l)}$ by solving a system with a zero right hand side

$$\mathsf{K}_{\mathrm{II}}^{(l)}\mathsf{w}_{\mathrm{I}}^{(l)} + \mathsf{K}_{\mathrm{I}\Gamma}^{(l)}\mathsf{w}_{\Gamma}^{(l)} = 0$$

Now, we seek for eigenmodes of Eq. (3.14) for which $\mu_{ij} \geq \mathsf{tol}$. In the original formulation [35], however, the eigenmodes v_{ij}^{Γ} are by construction continuous in the primal variables shared by the subdomains $\Omega^{(i)}$ and $\Omega^{(j)}$. In [18] it was observed that for many real-world coefficient distributions a good approximation of $\mathsf{v}_{\mathsf{ij}}^{\Gamma}$ can be constructed heuristically without a solution of any generalized eigenproblem. Instead, only the distribution of material coefficient on $\Gamma^{(i)} \cup \Gamma^{(j)}$ in a sense similar to that used in, e.g. ρ scaling or weighted averages, is needed for constructing new constraints. Specifically, for each finite element node on $\Gamma^{(l)}$ we define values of coefficient maxima on the elements whose basis functions have non-empty support at the node x (denoted as $\omega(x)$)

$$\hat{\alpha}^{(l)}(x) = \max_{\boldsymbol{y}\in\omega(x)\cap\Omega_l} \alpha(y) \qquad l \in \{i, j\},$$

where α stands for either ρ or E depending on the type of PDE under consideration. Then, these values are point-wise scaled with appropriate values corresponding to degrees of freedom in rigid body modes $\mathsf{r}^{(l)}$ of individual subdomains restricted to $\Gamma_h^{(l)}$, leading to the definition of $\hat{\mathsf{r}}_{E_{ij}}^{(l)}$

$$\hat{\mathsf{r}}_{E_{ij},m}^{(l)}(x) = \hat{\alpha}^{(l)}(x)\,\mathsf{r}_{E_{ij},m}^{(l)}(x) \qquad \forall x \in \Gamma^{(l)} \quad l \in \{i,j\} \quad m = 1, \ldots, n_{\mathrm{RBM}}^{E_{ij},(l)}$$

for each of $n_{\mathrm{RBM}}^{E_{ij},(l)}$ considered. Since there is no need for orthogonalization of $\mathsf{r}_{E_{ij},m}$, we shift the centre of rotational mode in case of two-dimensional mechanics problem to the geometric centre of the edge E_{ij}. Thus, we use the subscript E_{ij} to highlight that values in $\hat{\mathsf{r}}_{E_{ij}}^{(l)}$ are related to the edge, and for the same reason, we revise this extra superscript in $n_{\mathrm{RBM}}^{E_{ij},(l)}$. Hence, the number of constructed rigid body modes $n_{\mathrm{RBM}}^{E_{ij},(l)}$ is determined only by the character of the equation, thus being the same for all possible edges (except for polutional one-node edges). We admit that the term "rigid body modes" might be slightly misleading in this context, since it remains irrespective of $\dim\left(\mathrm{Kernel}(\mathsf{K}^{(l)})\right)$, and we always assume the whole set of rigid body modes pertinent to PDE under consideration, i.e. $n_{\mathrm{RBM}} = 1$ for the diffusion problem and $n_{\mathrm{RBM}} = 3$ for elasticity problem.

Driven by careful observation, authors in [18] proposed constructing $\mathsf{v}_{\mathsf{ij}}^{\Gamma}$ to set the values in the following way

$$\mathsf{v}_{\Gamma,ij,m}^{(l)}(x) := \begin{cases} \hat{\mathsf{r}}_{E_{ij},m}^{(l)}(x) & \text{if} \quad x \in \Gamma_h^{(l)} \setminus \Pi^{(l)} \\ 0 & \text{if} \quad x \in \Pi^{(l)}, \end{cases} \tag{3.15}$$

for which a vector constructed as

$$\mathsf{v}^{\Gamma}_{ij,m} = \begin{bmatrix} \mathsf{v}^{(i)}_{\Gamma,ij,m} \\ -\mathsf{v}^{(j)}_{\Gamma,ij,m} \end{bmatrix} \tag{3.16}$$

then results in a frugal constraint $\mathsf{c}_{ij,m} := \mathsf{B}_{\mathrm{D},ij}\mathsf{S}_{ij}\mathsf{P}_{\mathrm{D},ij}\mathsf{v}^{\Gamma}_{ij}$, a potentially suitable choice for the augmentation of the coarse space. Moreover, it would be beneficial if $\mathsf{v}^{(j)}_{E_{ij}}$ approximates the highest eigenmodes of Eq. (3.12). Then it would be reasonable to expect the ratio

$$\mu_{ij} = \frac{|\mathcal{H}_{\mathrm{I}}(\mathsf{P}_{\mathrm{D}_{ij}}\mathsf{v}^{\Gamma}_{ij,m})|_{\mathsf{K}_{ij}}}{|\mathcal{H}_{\mathrm{I}}(\mathsf{v}^{\Gamma}_{ij,m})|_{\mathsf{K}_{ij}}} \tag{3.17}$$

to be a reliable estimator of the dominant eigenvalues λ of Eq. (3.12).

Let us first comment on the specific construction of $\mathsf{v}^{\Gamma}_{ij,m}$. As expected in Eq. (3.15), the space of functions continuous in $\Pi^{(i)} \cap \Pi^{(j)}$ is limited to vectors that vanish at primary nodes. Specifically, we seek for heuristically constructed interface modes from a specific subspace $\widetilde{W}_{ij}$, which we denote

$$\widetilde{W}_{ij,0} = \left\{ \mathsf{w}_{ij} \in W_{h,i} \times W_{h,j} : \mathsf{w}_{ij|(\Pi^{(i)} \cap \Pi^{(j)})} = 0 \right\}.$$

This limitation is completely valid and justifiable for heuristic approaches. The process of identifying vectors that results in high ratios μ_{ij} in $\widetilde{W}_{ij,0}$ rather than $\widetilde{W}_{ij}$ is favourable for a numerical solution because no application of projections Π_{ij} would be needed, thus preserving a completely local character of the multiplication with S_{ij}. While it may seem appealing to solve GEVP (3.13) on $\widetilde{W}_{ij,0}$, this formulation would completely fail to recognise some of the constraints for specific coefficient distributions as explained next.

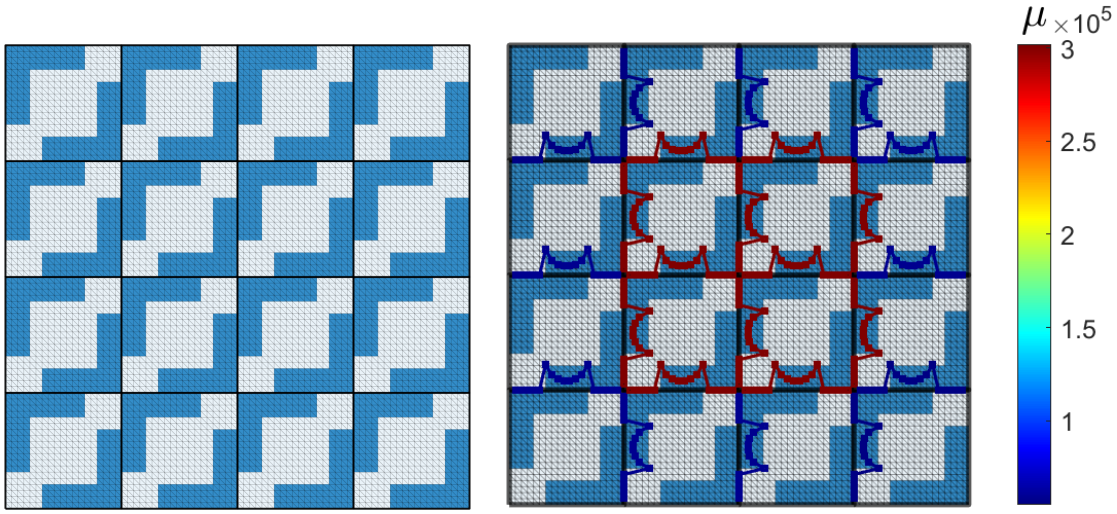


Figure 3.3: A synthetic example illustrating the unsuitability of searching for eigenmodes in $\widetilde{W}_{ij,0}$. With prescribed homogeneous boundary conditions at the locations of primal nodes, the GEVP does not recognize any of the harmful modes. **Left:** Coefficient distribution within a stationary diffusion problem: deep blue represents coefficient $\rho_{max} = 10^6$, light blue ( )$\rho_{min} = 1$. Dirichlet BCs are imposed on the whole $\partial\Omega$. We assume decomposition into 4×4 subdomains with a vertex-based primal coarse space; each subdomain consists of two distinct L-blocks. Each L-block has a connection to only one primal constraint. **Right:** Visualization of all adaptive constraints from solution of Eq. (3.13) with $\mathsf{tol} = 100$ on subspace $\widetilde{W}_{ij}$ (the weights of individual constraints in B are shown as profiles along individual edges, their color depicts the corresponding eigenvalue). One constraint is found on every edge. Deluxe scaling was used.

For better understanding, let us provide an illustrative example shown in Fig. 3.3. This synthetic, highly heterogeneous problem consists of regularly placed L-shaped blocks, each with a corner in one of the primal vertices on a domain decomposed into 16 square subdomains. Here we comment on the suitability of constraints obtained by a solution of the GEVP on

$\widetilde{W}_{ij,0}$ and $\widetilde{W}_{ij}$. It serves as a motivation for further slight modifications of frugal approach. Evidently, a vertex-based prior coarse space does not lead to a desirable condition number. The condition number $\kappa(\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{F}) \approx 2.22 \cdot 10^5$ is comparable to the coefficient contrast in this case, even for the deluxe scaling. To obtain a reasonably low condition number, the coarse space has to be augmented with constraints that prevent the L-blocks from floating. For a scalar problem, it is natural to anticipate that one constraint is sufficient to bind two L-blocks touching across boundary. According to the eigenvalue analysis, see Fig. 3.4, the preconditioned spectrum has nine distinct eigenvalues. Consequently, only nine globally

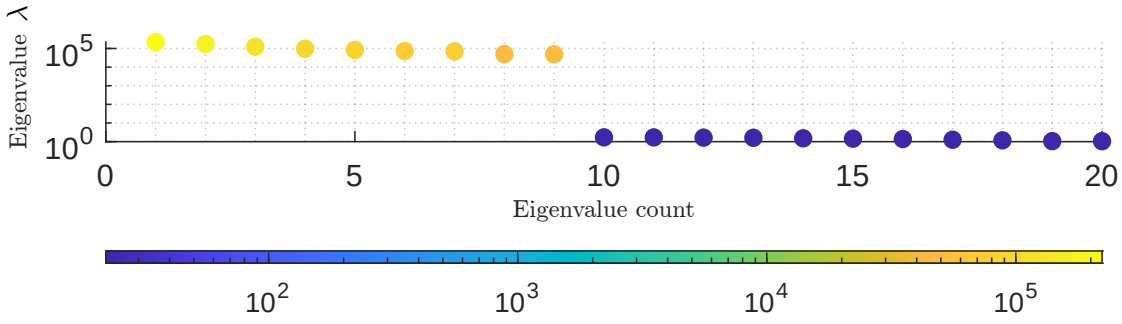


Figure 3.4: 20 highest eigenvalues λ of the $\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{F}$ of the problem depicted in Fig. 3.3

optimal constraints would suffice to drop the condition number by several orders of magnitude. However, one missing constraint in the prior coarse is found on each pair of the edges due to imposed locality of the GEVP, leading to unnecessary large coarse space. Even then, a careful selection of nine locally obtained constraints would suffice to prevent all the interior L-blocks from floating, leading to almost identical condition number. The reason behind this behavior is that certain combinations of the adaptively computed constraints are prone to yielding nearly redundant information, especially when dealing with binary distributions.

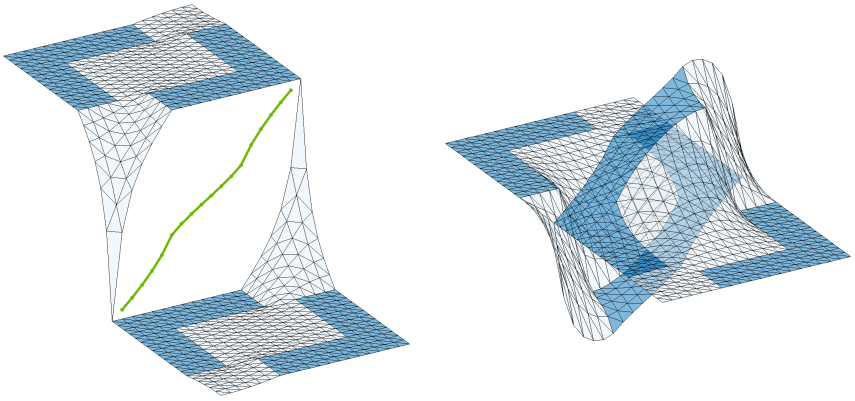


Figure 3.5: Visualization of the dominant eigenmode on two interior subdomains, indicated by the red colour in Fig. 3.3. **Left:** Illustration of the right-hand side of Eq. (3.13) with an applied minimum-energy extension into the interiors of the subdomains. The splitting of the gap based on deluxe scaling is depicted by a green line. This configuration results in low energy **3.51**. **Right:** The visiualization of the right hand side of Eq. (3.13) illustrates expected correction in temperature field based on the averaging of the gap and minimum energy extension into the interiors of the subdomains. Homogeneous BCs on the complement of edge w.r.t. $\Gamma_h^{(i)} \cup \Gamma_h^{(j)}$ follow from application of the localized P_{D} operator. This configuration corresponds to high energy $\mathbf{1.06 \cdot 10^6}$.

Note that in the case of binary distribution of coefficients the number of locally detected constraints is often predictable. The only feasible option to achieve the low-energy mode on the right-hand side of Eq. (3.10) is for the present problem to mutually shift the two L-blocks

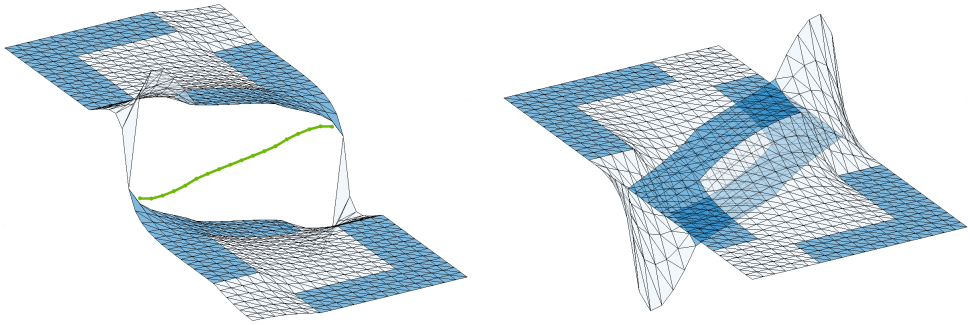


Figure 3.6: First mode on the same edge as in Fig. 3.5 found in space $\widetilde{W}_{ij,0}$, i.e. with values fixed at zero at shared primal vertices. **Left:** Visualization of splitting of the gap based on deluxe scaling on low-energy side of Eq. (3.12), resulting in energy $1.63 \cdot 10^5$. **Right:** Visualization of low-energy right-hand side of GEVP in $\widetilde{W}_{ij,0}$ leading to energy $2.84 \cdot 10^5$. This eigenmode corresponds to a low eigenvalue $\lambda_0 \approx 1.74$.

that touch, keeping each at a constant temperature level to avoid gradients on conductive elements; compare with Fig. 3.5. Our example was intentionally devised to pinpoint the limitations of seeking for modes in $\widetilde{W}_{ij,0}$. Clearly, it does not permit this form of relative shifts. Consequently, the best it can deliver is an eigenmode corresponding to eigenvalue only 1.74, depicted in Fig. 3.6.

As of now, we have identified a specific weak spot in finding constraints when we are restricted to construction in the sense of Eq. (3.15), which frugal constraints use. Recall that the frugal approach aims to identify modes at interfaces that are poorly captured by the preconditioner, i.e., the modes for which the chosen scaling leads to a poor correction of the solution. It comes as no surprise that the main objective of substructuring methods is to enforce continuity in elements with high coefficients, which are presumed to play a crucial role in mediating global information. Following the weighted averages, the focus is thus particularly on elements with high coefficients. The corresponding degrees of freedom belonging to these elements are heuristically prioritized through appropriate weighting, recall previously defined coefficients $\hat{\alpha}^{(l)}$. Unlike weighted averages, where the modes in solution variables are by construction identical among the two subdomains, frugal constraints use two individual weighted averages and construct the mode in such a way that the values on one subdomain are prescribed in one "direction" and on the second subdomain in the opposite "direction". These modes are designed to simulate the low-energy side of the GEVP with the greatest possible jump between high-coefficient elements across the interface. This is then used to let the scaling-dependent localized P_D operator construct the constraint mode. Furthermore, it enables us to evaluate and determine the relevance of this constraint, thus it gives us the choice to discard this mode.

Additionally, the limitation of frugal constraints has been illustrated on a very exotic example. To be fair, similar distributions are unlikely to be encountered in many realistic applications. Unfortunately, a similar limitation arises in mechanical problems in which the decomposition does not align with the distribution of the material. In such cases, a vertex-based coarse space is often prone to missing rotational modes, which are difficult to detect by the construction of modes with fixed centres of rotations, recall (3.3).

3.3.1 A modified construction of frugal constraints

We believe that a trivial improvement in the sense of ρ scaling could, in most cases, significantly improve the relevance of the estimator μ. We propose setting values in $\Pi^{(i)} \cap \Pi^{(j)}$ as weighted average between values stored in $\mathsf{v}^{(i)}_{\Gamma,ij,m}$ and $-\mathsf{v}^{(j)}_{\Gamma,ij,m}$, respecting the coefficient distribution

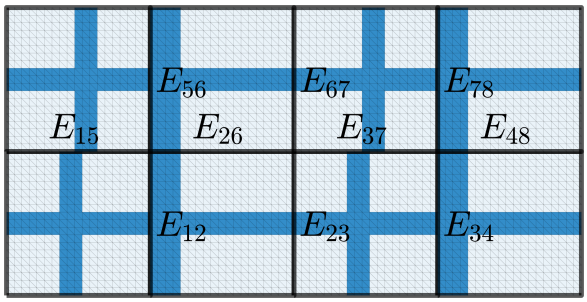


Figure 3.7: Left: Coefficient distribution : deep blue represents coefficient $\alpha_{max} = 10^6$, light blue ( ) $\alpha_{min} = 1$, where α stands for ρ in stationary diffusion or E in linear elasticity case. Decomposition of $\Omega := (0,4) \times (0,2)$ into 4×2 square subdomains. Vertex-based coarse space and deluxe scaling used. Dirichlet BCs are imposed on the left boundary at $x = 0$. The subdomains contain two channels each that touch variably at the marked edges.

on elements pertinent to nodes in Π_{ij} as well. The modified construction thus reads:

$$\mathsf{v}^{(l)}_{\mathrm{a},ij,m}(x) := \begin{cases} \hat{\mathsf{r}}^{(l)}_{E_{ij},m}(x) & \text{if} \quad x \in \Gamma^{(l)} \setminus \left(\Pi^{(i)} \cap \Pi^{(j)}\right) \\ (-1)^{\delta_{lj}} \cdot \left[\hat{\mathsf{r}}^{(i)}_{E_{ij},m}(x) - \dfrac{\hat{\alpha}^{(j)}}{\hat{\alpha}^{(i)} + \hat{\alpha}^{(j)}} \left(\hat{\mathsf{r}}^{(i)}_{E_{ij},m}(x) + \hat{\mathsf{r}}^{(j)}_{E_{ij},m}(x)\right)\right] & \text{if} \quad x \in \left(\Pi^{(i)} \cap \Pi^{(j)}\right), \end{cases} \tag{3.18}$$

which can be further simplified to

$$\mathsf{v}^{(l)}_{\mathrm{b},ij,m}(x) := \begin{cases} \hat{\mathsf{r}}^{(l)}_{E_{ij},m}(x) & \text{if} \quad x \in \Gamma^{(l)} \setminus \left(\Pi^{(i)} \cap \Pi^{(j)}\right) \\ (-1)^{\delta_{lj}} \left(\hat{\mathsf{r}}^{(i)}_{E_{ij},m}(x) - \hat{\mathsf{r}}^{(j)}_{E_{ij},m}(x)\right) & \text{if} \quad x \in \left(\Pi^{(i)} \cap \Pi^{(j)}\right). \end{cases} \tag{3.19}$$

We refer to this variant as "**frugal a**", while we denote the original variant according to Eq. (3.16) simply "**frugal orig**". Note that in the formulae above, we also distinguish between shared and remaining primal vertices. Indeed, we treat the primal nodes at $(\Gamma^{(i)} \cap \Gamma^{(j)}) \setminus \Gamma^{(ij)}$ in an identical manner as all the dual DOFs. In this way, we are again approaching the eigenproblem introduced in the previous section. Thus, we introduce a third variant as a compromise between the aforementioned two.

$$\mathsf{v}^{(l)}_{\mathrm{a},ij,m}(x) := \begin{cases} \hat{\mathsf{r}}^{(l)}_{E_{ij},m}(x) & \text{if} \quad x \in \Gamma^{(l)} \setminus \left(\Pi^{(i)} \cap \Pi^{(j)}\right) \\ 0 & \text{if} \quad x \in \left(\Pi^{(i)} \cap \Pi^{(j)}\right) \end{cases}$$

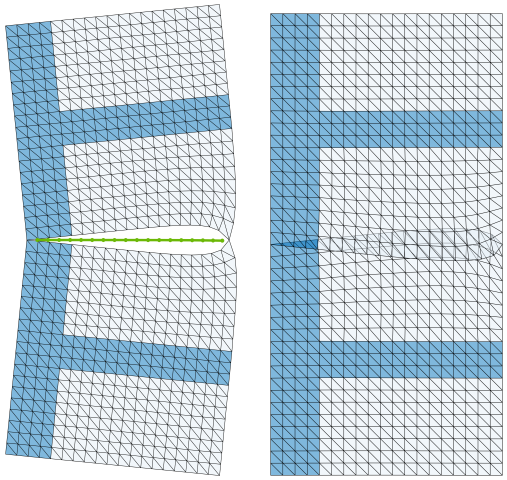
which we call "**frugal b**". Note that in the variant **frugal b**, $\mathsf{v}_{\mathrm{b},ij}$ is set to 0 only at all primal nodes shared by two subdomains $\Pi^{(i)} \cap \Pi^{(j)}$, not at all $\Pi^{(i)} \cup \Pi^{(j)}$ primal nodes. Let us motivate this second modification with a second example, shown in Fig. 3.7. This problem consists of eight subdomains, each having two conductive channels crossing the interface. These channels are either aligned or shifted by a few elements. Dirichlet BCs are imposed on the left side of and the problem is subjected to a constant flux. In the diffusion problem case, a complete set of ten frugal constraints is necessary to decrease the condition number from the initial state with $\kappa(\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{F}) \approx 2.82 \cdot 10^5$ and four eigenvalues exceeding 10^5, to the constrained system with $\kappa(\mathsf{M}_{\mathrm{bal}}^{-1}\mathsf{F}) \approx 1.30$. Hence, this setting demonstrates when the application of frugal constraints is beneficial. Moreover, in this problem, it is difficult to exclude many of the constraints, because only channels at edges E_{26} and E_{48} are connected through primal vertices, and the same holds for an edge E_{15}, where both domains are strongly impacted by Dirichlet BCs.

In the linear elasticity case, the situation is similar. If all thirty, i.e. three frugal modes for each of the ten edges are taken into account, a condition number decreases from the initial value of $\kappa(\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{F}) \approx 5.27 \cdot 10^5$ with fourteen distinct eigenvalues in range $\langle 5.5 \cdot 10^3, 5.3 \cdot 10^5 \rangle$ down to the system with $\kappa(\mathsf{M}_{\mathrm{bal}}^{-1}\mathsf{F}) \approx 1.69$. Now, the original eigenproblem (3.13) identifies only 22 potentially suitable constraints that exceed tolerance $\mathsf{tol} = 100$. Assuming that the frugal approach can accurately estimate the same eigenmodes as GEVP, our interest now lies in omitting the eigenmodes associated with low eigenvalues λ. To this end, we would like

to evaluate the eigenvalue estimator μ for each constraint to have the option to reduce the number of additional constraints in future applications.

	E_{12}	E_{23}	E_{34}	E_{56}	E_{67}	E_{78}	E_{15}	E_{26}	E_{37}	E_{48}
Stationary diffusion										
GEVP (Eq. 3.13)	$\mathbf{2.03 \cdot 10^5}$	$\mathbf{2.49 \cdot 10^5}$	$\mathbf{1.96 \cdot 10^5}$	$\mathbf{2.14 \cdot 10^5}$	$\mathbf{2.41 \cdot 10^5}$	$\mathbf{2.07 \cdot 10^5}$	1.33	2.34	$\mathbf{2.61 \cdot 10^5}$	2.34
frugal orig	$3.25 \cdot 10^{-1}$	$2.26 \cdot 10^{-1}$	$3.56 \cdot 10^{-1}$	$3.59 \cdot 10^{-1}$	$2.15 \cdot 10^{-1}$	$3.88 \cdot 10^{-1}$	$9.58 \cdot 10^{-1}$	$9.63 \cdot 10^{-1}$	$\mathbf{1.80 \cdot 10^4}$	$9.63 \cdot 10^{-1}$
frugal a	2.77	$\mathbf{2.19 \cdot 10^4}$	$\mathbf{3.74 \cdot 10^4}$	3.40	$\mathbf{2.08 \cdot 10^4}$	$\mathbf{4.07 \cdot 10^4}$	$9.58 \cdot 10^{-1}$	1.75	$\mathbf{1.80 \cdot 10^4}$	1.75
frugal b	$3.25 \cdot 10^{-1}$	$\mathbf{2.19 \cdot 10^4}$	$3.56 \cdot 10^{-1}$	$3.59 \cdot 10^{-1}$	$\mathbf{2.08 \cdot 10^4}$	$3.88 \cdot 10^{-1}$	$9.58 \cdot 10^{-1}$	1.75	$\mathbf{1.80 \cdot 10^4}$	1.75
Linear elasticity										
	E_{12}	E_{23}	E_{34}	E_{56}	E_{67}	E_{78}	E_{15}	E_{26}	E_{37}	E_{48}
	$\mathbf{2.13 \cdot 10^5}$	$\mathbf{5.00 \cdot 10^5}$	$\mathbf{1.36 \cdot 10^5}$	$\mathbf{2.01 \cdot 10^5}$	$\mathbf{5.14 \cdot 10^5}$	$\mathbf{1.28 \cdot 10^5}$	8.28	$\mathbf{7.00 \cdot 10^4}$	$\mathbf{1.83 \cdot 10^5}$	$\mathbf{7.00 \cdot 10^4}$
GEVP (Eq. 3.13)	$\mathbf{3.14 \cdot 10^4}$	$\mathbf{1.58 \cdot 10^4}$	$\mathbf{3.08 \cdot 10^4}$	$\mathbf{5.01 \cdot 10^4}$	$\mathbf{1.19 \cdot 10^4}$	$\mathbf{4.86 \cdot 10^4}$	2.80	1.91	$\mathbf{2.47 \cdot 10^4}$	1.91
	$\mathbf{7.85 \cdot 10^3}$	$\mathbf{6.38 \cdot 10^3}$	$\mathbf{7.78 \cdot 10^3}$	$\mathbf{8.54 \cdot 10^3}$	$\mathbf{6.45 \cdot 10^3}$	$\mathbf{8.43 \cdot 10^3}$	1.55	1.61	$\mathbf{5.79 \cdot 10^3}$	1.44
	$3.33 \cdot 10^{-1}$	$3.33 \cdot 10^{-1}$	$2.49 \cdot 10^{-1}$	$3.25 \cdot 10^{-1}$	$3.13 \cdot 10^{-1}$	$2.36 \cdot 10^{-1}$	$6.59 \cdot 10^{-2}$	$9.06 \cdot 10^{-1}$	$\mathbf{1.35 \cdot 10^3}$	$9.06 \cdot 10^{-1}$
frugal orig	$4.51 \cdot 10^{-2}$	$6.93 \cdot 10^{-3}$	$4.53 \cdot 10^{-2}$	$7.06 \cdot 10^{-2}$	$5.36 \cdot 10^{-3}$	$6.98 \cdot 10^{-2}$	2.33	$7.29 \cdot 10^{-1}$	$\mathbf{8.31 \cdot 10^3}$	$7.29 \cdot 10^{-1}$
	$2.39 \cdot 10^{-2}$	$1.22 \cdot 10^{-3}$	$2.52 \cdot 10^{-2}$	$2.48 \cdot 10^{-2}$	$1.11 \cdot 10^{-3}$	$2.51 \cdot 10^{-2}$	$1.22 \cdot 10^{-2}$	$2.96 \cdot 10^{-1}$	$\mathbf{1.67 \cdot 10^2}$	$2.96 \cdot 10^{-1}$
	2.31	$\mathbf{2.78 \cdot 10^4}$	$\mathbf{2.25 \cdot 10^4}$	2.42	$\mathbf{2.74 \cdot 10^4}$	$\mathbf{2.24 \cdot 10^4}$	$6.59 \cdot 10^{-2}$	1.37	$\mathbf{1.35 \cdot 10^3}$	1.37
frugal a	1.73	$\mathbf{7.15 \cdot 10^2}$	$\mathbf{5.04 \cdot 10^3}$	4.56	$\mathbf{5.63 \cdot 10^2}$	$\mathbf{7.86 \cdot 10^3}$	2.33	1.39	$\mathbf{8.31 \cdot 10^3}$	1.39
	$3.06 \cdot 10^{-1}$	$\mathbf{3.30 \cdot 10^2}$	$\mathbf{1.22 \cdot 10^3}$	$6.12 \cdot 10^{-1}$	$\mathbf{2.62 \cdot 10^2}$	$\mathbf{1.33 \cdot 10^3}$	$1.22 \cdot 10^{-2}$	$9.80 \cdot 10^{-1}$	$\mathbf{1.67 \cdot 10^2}$	$9.80 \cdot 10^{-1}$
	$3.33 \cdot 10^{-1}$	$\mathbf{2.78 \cdot 10^4}$	$2.49 \cdot 10^{-1}$	$3.25 \cdot 10^{-1}$	$\mathbf{2.74 \cdot 10^4}$	$2.36 \cdot 10^{-1}$	$6.59 \cdot 10^{-2}$	1.37	$\mathbf{1.35 \cdot 10^3}$	1.37
frugal **b**	$4.51 \cdot 10^{-2}$	$\mathbf{7.15 \cdot 10^2}$	$4.53 \cdot 10^{-2}$	$7.06 \cdot 10^{-2}$	$\mathbf{5.63 \cdot 10^2}$	$6.98 \cdot 10^{-2}$	2.33	1.39	$\mathbf{8.31 \cdot 10^3}$	1.39
	$2.39 \cdot 10^{-2}$	$\mathbf{3.30 \cdot 10^2}$	$2.52 \cdot 10^{-2}$	$2.48 \cdot 10^{-2}$	$\mathbf{2.62 \cdot 10^2}$	$2.51 \cdot 10^{-2}$	$1.22 \cdot 10^{-2}$	$9.80 \cdot 10^{-1}$	$\mathbf{1.67 \cdot 10^2}$	$9.80 \cdot 10^{-1}$

Table 3.2: Highest (one in the case of two-dimensional stationary diffusion, three in the linear elasticity case) eigenvalues (or their estimates) corresponding to computed modes on each of the edges E_{ij} using one of four methods: GEVP proposed by Mandel and Sousedík, **frugal orig**, **frugal a** and **frugal b**. The values represent eigenvalues λ for adaptive modes obtained by solution of the eigenproblem 3.13 or their estimator μ in case of frugal modes, respectively. Values greater than 100 are highlighted in bold.

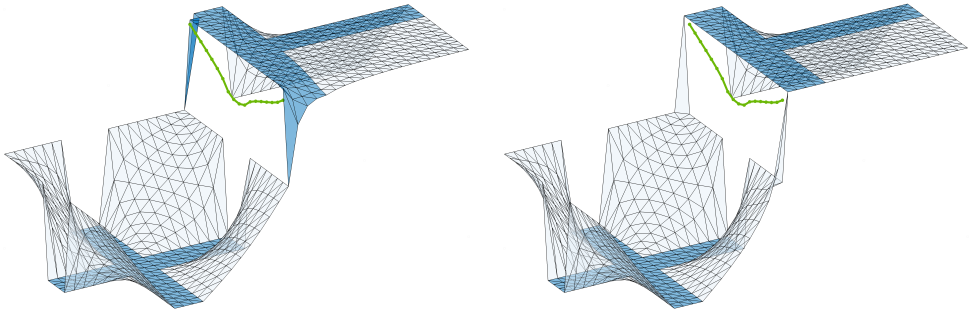


(a) (b)

Figure 3.8: Illustration of the rotational mode in elasticity problem on E_{26} which is hard to capture with frugal approach: **(a)** Side of low energy: $3.57 \cdot 10^{-1}$. **(b)** Side of high energy: $2.50 \cdot 10^4$

The comparison of the three variants of a heuristic frugal approach with reference locally optimal constraints obtained by the solution of the GEVP is provided in Tab. 3.2. First, it is evident that there is a notable resemblance between the scalar and vector-valued problems. The only edge that certainly does not contain any harmful mode is E_{15}, because relatively high coefficient segments that intersect edge E_{15} are directly connected to the Dirichlet boundary. Next edges E_{26} and E_{48} follow in their low harmfulness. In the scalar case, the high-coefficient blocks are completely handled via shared primal vertex. In mechanics, this connection in the primal vertices suffices to cope with relative translational modes, however the rotational mode centred on this vertex is hard to capture for a frugal approach. Typically, these rotational modes are not needed for a well-connected structure because of the (primal or translational) constraints arising from other edges. Edges E_{23} and E_{67} represent examples

in which the variant **frugal orig** results in an inaccurate energy estimate of all constraints, which both **frugal a** and **frugal b** correctly evaluate high value of μ in the corresponding columns in Tab. 3.2. It is also evident from Tab. 3.2 that similar behaviour does not extent to edges E_{34} and E_{78}, where both primal vertices are located on the jump between high and low coefficients; see Fig. 3.9. Here, the importance of the proposed **variant a** becomes apparent. According to Tab. 3.2, it is the only variant that properly estimates the λ from GEVP by its μ value. The difference between construction of constraints in sense of **frugal a** and the standard variant is also clearly visible in Fig. 3.9, where modes appearing in denominator of μ are shown. In **frugal a**, we avoid prescribing a gradient on a significantly more conductive elements by appropriate shifting of the temperature at shared primal vertices. This is not the case in **frugal orig**, where the temperature is kept fixed at zero irrespective of the material distribution. Edge E_{37} is an ideal representative of where no modifications have to be made;



(a) (b)

Figure 3.9: Stationary diffusion: Frugal mode on the edge E_{34}. Low energy sides and corresponding energies in parentheses are shown for: **(a) frugal orig**: $(1.81 \cdot 10^{18})$. **(b) frugal a**: $(1.72 \cdot 10^{13})$.

estimator μ correctly accepts all the constructed frugal constraints in diffusion as well as elasticity. It is necessary to note that the layout of the material distribution as given in substructures $\Omega^{(3)}$ and $\Omega^{(7)}$ is exactly what we aim to avoid in practice, as the primary nodes do not fulfill the role of a global skeleton. The most challenging and the only situations where the frugal approach completely fail to evaluate the adequacy of constructed constraints are the cases with one subdomain partly lying on the Γ_D, for our test problem from Fig. 3.7 see E_{12} and E_{56}.

To conclude, the original frugal formulation does not seem to be applicable for an adaptive reduction of the number of the enforced constraints. The simplified proposed version **frugal b** might be useful for well-behaved decomposition or when we accept the risk of overlooking potentially suitable constraints that could be captured if a more sensitive selection was adopted. Finally, the proposed variant **frugal a** is a recommended strategy for a smart, problem-specific selection, as it successfully recognized 20 out of 30 adaptive constraints in total.

3.4 Reduced-basis strategy for obtaining adaptive constraints

The adaptive techniques described above remain considered computationally expensive and time-consuming despite the robustness they provide and despite the associated (significant) reduction in the number of iterations needed to achieve a satisfactorily accurate solution. This is particularly due to data transfers between computational units are not negligible, and the set-up and computation of the eigenproblem represent a costly operation even when localized among pairs of substructures. Moreover, especially when it is not necessary to

compute the eigenproblems at all interfaces, a significant imbalance is introduced, disrupting the otherwise perfectly parallelizable architecture of the method itself. An important insight is that in real-world problems, we typically require only a few – if any at all – constraints at each interface; that is, in most instances, the number of eigenmodes required is markedly (or orders of magnitude) less than the dimension of the eigenproblem. Therefore, the adoption of iterative solvers for identifying (potentially approximate) dominant eigenmodes, such as those employing the generalized Rayleigh quotient, emerges as a seemingly appropriate choice [53]. However, in practice, direct Krylov-Schur-based sparse eigensolvers, such as those implemented in PETSc which enable parallel computations are still frequently used due to their robustness [3, 22]. A principal drawback with the formulation of the eigenproblem introduced in Subsection 3.2.1 is that, with a direct solver, it inherently produces a substantial number of eigenmodes. Not only is the majority of these eigenmodes likely to be discarded due to being assessed as unnecessary, but we can also confidently claim that a large number of eigenmodes is predetermined to carry no useful information. This is due to the subsequent application of the P_D operator to our selected eigenmodes: It is evident that if $\Gamma_h^{(ij)}$ contains more dual DOFs than our interface, edge E_{ij} for instance, many of the $\mathsf{B}_{\mathrm{D},ij}\mathsf{S}_{ij}\mathsf{P}_{\mathrm{D},ij}\mathsf{v}_{ij}$ products will either be completely zeroed out or be linearly dependent on the others. As we have demonstrated in the preceding sections, the acquisition of all the desired constraints necessitates the employment of a correct formulation of an adaptive approach. Various modifications, which otherwise hold considerable potential to enhance the efficiency of obtaining these adaptive conditions, generally fail to produce all the sought-after modes. The goal of this section is thus different. We aim to compute a properly formulated eigenproblem without reliance on any additional simplifying assumptions, but within an ideally very low-dimensional subspace. While this is a common strategy in numerous eigensolvers (hence labelled Krylov-), e.g. [3, 47], our effort here is to focus on an explicit construction of the reduced basis by relying exclusively on geometrical and material parameters. Direct approximation of the constraints proved challenging in cases of uniformly varying coefficients or when multiple constraints are required. However, drawing on principles of heuristic methodologies we can attempt to estimate the basis vectors from which the maximum number of necessary conditions could be constructed. Practically, we will thus adopt the concept of (weighted) averages, but we will employ an eigensolver for the assembly of the constraints because it is challenging to produce an optimal, linearly independent, yet complete set of necessary constraints. Consequently, we face an objective to construct a basis Ψ, ideally of smallest feasible dimension, which spans the space of the dominant eigenvectors we seek for.

As a first step in constructing such a basis, let us focus on binary-valued scalar problems. For scalar problems, visualizations become somewhat simpler, and additionally, these problems do not feature harder-to-handle rotational modes. If the material coefficient distribution is binary and the heterogeneity ratio is high, it is easy to determine whether each high-coefficient segment is already in the coarse space, simply by checking whether it is connected to the Dirichlet boundary through primal constraints or not. Since the eigenmodes arising from the solution of localized GEVPs are typically strongly associated with very low-energy functions, it can be anticipated that the solution variables within one high-coefficient segments will be of the same value.

Let us provide a minimalistic example comprising two interior subdomains of a comb-like shaped high-coefficient elements; see the scheme in Fig. 3.10 left. In this example, a single adaptive constraint is needed if ρ scaling is used. Clearly, we aim to introduce two basis functions for the lower subdomain and ideally a single basis function for the upper subdomain. To keep the heuristics cheap, however, we will not work with information inside the subdomains. Instead, we will limit ourselves to coefficient profiles similar to those used in ρ scaling. Unlike ρ scaling, where values are defined nodal-wise, we will adhere to values belonging to individual elements. Thus, we introduce coefficient profiles $\xi^{(s)}$, see their visualization on in Fig. 3.10 top-right. Assuming that $(x^1_{\Gamma(s)}, x^2_{\Gamma(s)}, \ldots, x^{n^{(s)}_\Gamma}_{\Gamma(s)})$ is a sequence of

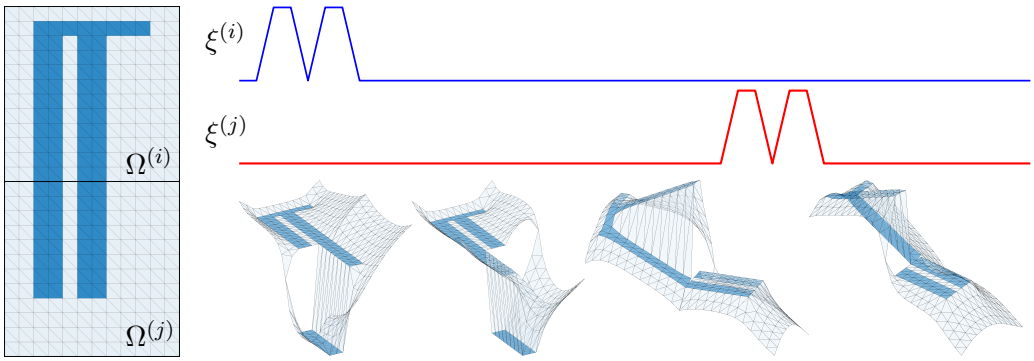


Figure 3.10: Illustration of the four constructed basis functions on an interface between two subdomains with a comb-like conductive segment. **Left:** Stationary diffusion problem and two interior subdomains. Similarly to previous examples, dark blue shows region with high coefficient and light blue region with 10^6 lower coefficient. **Top right:** Coefficient profiles ξ on subdomain interfaces $\Gamma^{(s)}$ in a counter clock-wise ordering starting from the local bottom-left node: blue for the upper subdomain, red for the bottom one. **Bottom right:** Four constructed basis functions with minimum energy extension to the interiors of subdomains.

nodes on $\Gamma^{(s)}$ in a consecutive order, we can define values of coefficient profile for each element edge between two unsupported degrees of freedom on $\Gamma_h^{(s)}$ (recall that DOFs belonging to the set I are not attached to $\Gamma_h^{(s)}$ as

$$\xi_o^{(s)} := \alpha_{\left\{x^o_{\Gamma^{(s)}}, x^{o+1 \bmod n_\Gamma^{(s)}}_{\Gamma^{(s)}}\right\}} \quad o = 1 \dots n_\Gamma^{(s)} - 1, \text{ and possibly } n_\Gamma^{(s)}, \tag{3.20}$$

where $\alpha_{\left\{x^o_{\Gamma^{(s)}}, x^{o+1 \bmod n_\Gamma^{(s)}}_{\Gamma^{(s)}}\right\}}$ denotes the coefficient α (being either ρ or E) of an element in $\Omega^{(s)}$ with an edge connecting nodes $x^o_{\Gamma^{(s)}}$ and $x^{o+1}_{\Gamma^{(s)}}$. The modulo operation addresses situations where the boundary $\Gamma^{(s)}$ is a closed curve, and thus there is an element with nodes $\{x^1_{\Gamma^{(s)}}, x^{n_\Gamma^{(s)}}_{\Gamma^{(s)}}\}$. If the coefficient distribution itself is not accessible in the solver, the corresponding off-diagonal entries may take the role of α in Eq. (3.20).

We now proceed to the next step in the construction of a heuristic basis: identifying clusters of elements with significantly high coefficients within the coefficient profiles $\xi_o^{(s)}$. In the binary case, distinguishing between elements with low and high coefficients is relatively straightforward. For this purpose, we utilize the indicator function, denoted by $\chi(\xi_o^{(s)})$, where $\xi_o^{(s)}$ corresponds to the coefficient value of an element o in local ordering. This function is defined as:

$$\chi(\xi_o^{(s)}) = \begin{cases} 1 & \text{if } x \geq \mathsf{tol}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

A cluster $C^{(s)}$ on $\Gamma^{(s)}$ is defined as a contiguous sequence of elements with nodes $x^o_{\Gamma^{(s)}}, x^{o+1}_{\Gamma^{(s)}}, \dots, x^p_{\Gamma^{(s)}}$ such that each $\chi(x^j_{\Gamma^{(s)}}) = 1$ for all $o \leq j \leq p$. The indices o and p mark the start and end of a cluster, respectively. To avoid unnecessary introduction of new basis functions, we account for a particular case when $\Gamma^{(s)}$ is a closed curve by setting $x^{n_\Gamma^s+1}_{\Gamma^{(s)}} \leftarrow x^1_{\Gamma^{(s)}}$, i.e. linking the end of the sequence back to the start. The total number of clusters $n_C^{(s)}$ on a subdomain is determined by counting the transitions from $\chi(x) = 0$ to $\chi(x) = 1$ shown if Fig. 3.20.

We now formalize the construction of the basis for our example. The basis consists of piecewise constant functions, specifically designed to represent the clusters identified in the coefficient profiles. Each basis vector, ψ_k, corresponds to a unique cluster $C_k^{(s)}$ on $\Gamma^{(s)}$, $s = i, j$

and is defined as follows:

$$\psi_k^{(s)}(x) = \begin{cases} 1 & \text{if } x \in C_k^{(s)} \quad \forall k = 1 \dots n_C^{(s)}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.22}$$

This formulation implies that $\psi_k^{(s)}$ acts as a characteristic function on a set of DOFs given by each cluster $C^{(s)}$, taking the value 1 within the spatial domain of its corresponding cluster $C_k^{(s)}$ and 0 elsewhere. As of now, we have not considered the fact that the solution must be from $\widetilde{W}_{ij}$. To incorporate this requirement, we could directly enforce values at nodes from Π_{ij} to be continuous across the interface. For that, we set

$$\overline{\psi}_k^{(i)}(x) = \begin{cases} 1 & \text{if } x \in C_k^{(j)} \text{ and } x \in \Pi_{ij}\,, \\ 0 & \text{otherwise.} \end{cases} \tag{3.23}$$

and the same definition applies vice versa for j.

Now we finally have everything ready to construct a reduced basis

$$\Psi = \begin{bmatrix} \Psi^{(i)} & \overline{\Psi}^{(i)} \\ \overline{\Psi}^{(j)} & \Psi^{(j)} \end{bmatrix} \tag{3.24}$$

with $\Psi^{(s)} \in \mathbb{R}^{n_\Gamma^{(s)} \times n_C^{(s)}}$.

We are now approaching a basis from which we can expect to obtain the adaptive constraints needed for enhancing coarse space. Practically, we could now apply a Galerkin projection onto S_{ij} and $\mathsf{P}_{\mathrm{D}ij}^{\mathsf{T}}\mathsf{S}_{ij}\mathsf{P}_{\mathrm{D}ij}$, i.e., instead of solving GEVP (3.13) we have to find the eigenmodes and corresponding eigenvalues of

$$\Psi^{\mathsf{T}}\mathsf{P}_{\mathrm{D}ij}^{\mathsf{T}}\mathsf{S}_{ij}\mathsf{P}_{\mathrm{D}ij}\Psi\mathsf{v}_\Psi = \lambda_\Psi \Psi^{\mathsf{T}}\mathsf{S}_{ij}\Psi\mathsf{v}_\Psi \tag{3.25}$$

and reconstruct the approximated original eigenmodes by setting $\mathsf{v} = \Psi\,\mathsf{v}_\Psi$. In this case, Ψ itself represents the role of both projections in eigenproblem (3.13).

The purpose of the construction presented so far is to provide a stepping stone for the next development. Once implemented, it turned out that such a simplified approach is suitable only for simplified binary problems similar to the one given, and even then only for scalar tasks. However, there is still a room for improvement, because the aforementioned basis still operated on the entire $\Gamma^{(i)} \times \Gamma^{(j)}$, which is not necessary. While the goal is to drastically reduce the dimensionality of localized GEVPs, it might generally be acceptable for the solution of the eigenproblem itself if we ended up with a slightly larger basis. This is because the computational time for solving small eigenproblems is negligible, even with a slightly increased number of constraints. However, since we will be using minimum-energy extensions that are needed anyway, we will appreciate reducing the number of basis vectors to an absolute possible minimum. The reason for this is that having just a few right-hand sides for the evaluation of minimum energy extensions allows us to potentially utilize inexact iterative solvers. For example, with regular decomposition into square subdomains, we would expect the cumulative sum of dimensions from solved eigenproblems to be approximately or up to (depending on whether we discarded some eigenproblems with no clusters on the shared edge or not) for times larger, which is unfortunate.

Let us proceed to establish the final form of our reduced basis. The first distinction from the previous form is that we scan for clusters of high-coefficient elements only on a shared entity, specifically an edge. The second deviation from the previously introduced procedure is that the clusters have global character right from the assembly phase. For this purpose, we redefine the coefficient profiles $\xi^{(s)}$, $s \in \{i, j\}$ to incorporate both domains simultaneously, introducing the profile $\xi^{(ij)}$ as

$$\xi^{(ij)} := \begin{bmatrix} \xi^{(i)} \\ \xi^{(j)} \end{bmatrix},$$

where the ordering must be modified such that every two consecutive values of $\xi^{(ij)}$ share a geometrically coinciding node at E_{ij}.

Newly, clusters within the shared coefficient profile $\xi^{(ij)}$ are identified by locating the index

k corresponding to the maximum coefficient value that exceeds a specified threshold $\mathsf{tol}_{\mathrm{RB}}$, relative to the local values within its neighborhood. This is expressed as:

$$C^{(ij)} = \left\{ k \;\middle|\; \overline{\xi}_k^{(ij)} = \max_{m \in [t_{\mathrm{start}}, t_{\mathrm{end}}]} \xi_m^{(ij)} \text{ and } \overline{\xi}_k^{(ij)} > \mathsf{tol}_{\mathrm{RB}} \cdot \max\left(\xi_{t_{\mathrm{start}}}^{(ij)}, \xi_{t_{\mathrm{end}}}^{(ij)} \right) \right\}, \tag{3.26}$$

where t_{start} and t_{end} are dynamically determined for each segment within the profile where the maximum value meets or exceeds the given user-defined threshold.

We start by scanning the coefficient profiles $\xi^{(ij)}$ from the first value, i.e., by setting $t_{\mathrm{start}} = 1$, and actualize the temporary minimum $\hat{\xi}_{\mathrm{min}}$ and its corresponding index t_{start} based on the values in coefficient profile, which determines the potential start of the interval. Once a value exceeding $\mathsf{tol}_{\mathrm{RB}} \cdot \hat{\xi}_{\mathrm{min}}$ is recognized, we start again to search for the end point of the interval by testing expression (3.26). Once this criterion is met, we mark the index of the node corresponding to the highest value within the coefficient profile restricted to this interval and use it for the construction of a basis function $\overline{\psi}_k^{(i)}$, i.e. our cluster. A cluster is characterized by the element with the highest coefficient, as it is expected that, after minimal energy extension, it will accurately approximate ill-posed modes. After successful identification of a cluster, the elements within the interval $(t_{\mathrm{start}}, t_{\mathrm{end}})$ are marked as visited, prohibiting their use in the construction of additional clusters. Subsequent clusters are determined by iteratively applying this evaluation process to the remaining unvisited segments of $\xi^{(ij)}$. This procedure continues until all potential peaks within the profile have been evaluated. For vector-valued problems, we rather split the clusters defined on intervals where t_{start} operates on the other domain than t_{end}. This is due to the fact that this splitting is beneficial for capturing rotational modes. Hence, if this occurs, we only force the node shared by the two subdomains to be a representative of t_{end} on the first subdomain and t_{start} on the second one: the basis functions are then constructed accordingly to the previous case.

Admittedly, the construction of the reduced basis is not unique as it depends on the chosen direction and starting point of the search process. At the same time, the described procedure is difficult to generate for two-dimensional entities arising three dimensional tasks. Current setup serves as a proof-of-concept that obtaining a suitable low-dimensional and yet well-performing basis using only a limited information about material distribution on the edge/face shared by a pair of adjacent substructures is possible. For the extension, methods such as generic clustering algorithms in machine learning techniques can be used for identification of high-coefficient aggregates.

The process described by Eq. (3.26) is complicated by the necessity of working with relative coefficient ratios, which introduces further challenges in identifying clusters. An illustrative problem depicted in Fig. 3.11 demonstrates the need for a relative threshold. This problem has been intentionally altered from that presented in Fig. 3.1a to include coefficient distribution consisting of four values of $\rho \in \{1, 10^2, 10^4, 10^6\}$. The transition between the two subdomains always occurs at the interface and it is consistently equal to 100. Adaptive methods in this case yield almost the same four constraints and eigenvalues as for the problem with binary-valued coefficients, where the material contrast equals 100. Hence, although the global heterogeneity ratio $\rho_{\mathrm{max}}/\rho_{\mathrm{min}} = 10^6$, the eigenvalues corresponding to the visualized constraints remain comparable to the maximum coefficient jump on each edge, which is, approximately 10^2. In order to effectively capture eigenvalues in the order of 10^2, which are anticipated in the system, it is practical to set the tolerance value tol also in the order of the coefficient jump, or slightly lower. The tolerance used in our heuristic has an analogous significance; therefore $\mathsf{tol}_{\mathrm{RB}}$ can be set slightly below the value of coefficient contrast. This is advantageous because it keeps the only parameter entering this heuristic procedure meaningful, hence eliminating the need for a complex parameter-tweaking. In our implementation, the same threshold for tolerance was consistently applied in both this criterion and for the eigenproblem when working with binary-coefficient problems.

For vector problems, we first identify clusters exactly as in the scalar case and construct an

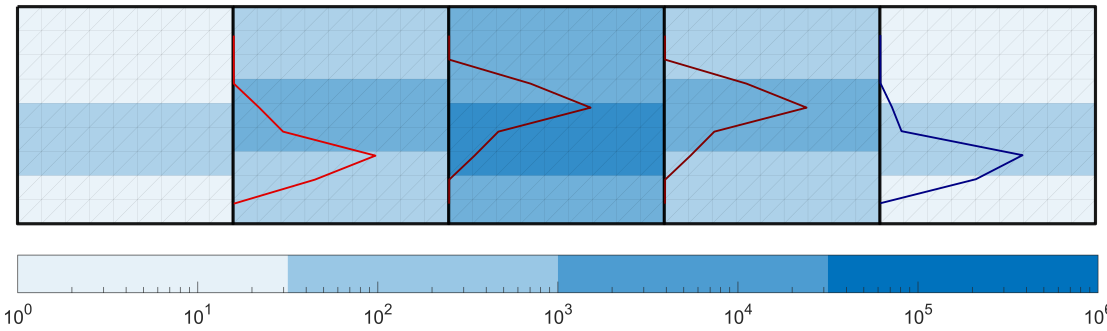


Figure 3.11: Alternated stationary diffusion problem. Colorbar denotes four different values of coefficient ρ. Red and blue lines visualize adaptively obtained constraint. Corresponding eigenvalues are ranging from 85.93 to 96.94

independent nodal basis Ψ_0 in accordance to expressions (3.22-3.24); then for each cluster we construct an independent basis consisting of rigid-body modes $\hat{\Psi}$.

First, we define a DOF-wise basis Ψ matrix from the nodal basis Ψ_0 such that each node represented by a one in Ψ_0 will have all associated DOFs marked by one in Ψ. This is formally accomplished by utilising the Kronecker product with a column vector $1 \in \mathbb{R}^{n_{\text{DOFs/elem}}}$ consisting of ones, where $n_{\text{DOFs/elem}}$ is the number of degrees of freedom per element, i.e. $n_{\text{DOFs/elem}} = 1$ for scalar problems and $n_{\text{DOFs/elem}} = d$, the dimension, for vector problems:

$$\Psi = \Psi_0 \otimes 1_{n_{\text{DOFs/elem}} \times 1}. \tag{3.27}$$

We now introduce $\hat{\Psi}$ such that it stores the pointwise-multiplied columns of Ψ with the corresponding rigid body modes of individual clusters. As the clusters are line segments of positive measure and we restrict ourselves to two-dimensional problems, a matrix $\mathsf{R}_{\text{rbm}}^{(ij)}$ storing the rigid body modes of the connected superstructure $\overline{\Omega}^{(i)} \cup \overline{\Omega}^{(j)}$ can be simply constructed and its columns point-wise multiplied with columns of $\hat{\Psi}$ instead. In general, for three-dimensional problems, clusters may possibly have fewer linearly independent rigid body modes.

The resulting usable part of the basis, $\hat{\Psi} \in \mathbb{R}^{n_C^{(ij)} \times (n_{\text{DOFs}}^{\Gamma(ij)} \cdot n_{\text{rbm}})}$, where $n_{\text{DOFs}}^{\Gamma(ij)}$ represents the number of decoupled interface degrees of freedom, and n_{rbm} is the problem-dependent count of rigid body modes. For the sake of clarity, the matrix is now given by

$$\hat{\Psi} = (\Psi \otimes 1_{1 \times n_{\text{rbm}}}) \odot (\mathsf{R}_{\text{rbm}}^{(ij)} \otimes \mathbf{1}_{1 \times n_C^{(ij)}}). \tag{3.28}$$

This basis indeed satisfies the continuity conditions in Π_{ij}, but in practice, it proves to be inapplicable. This is due to sharp transitions in basis functions, leading to steep gradient in neighboring lower-coefficient elements. This typically results in high energy modes in most cases. We have to control this.

However, by establishing this framework, we have laid the groundwork for the application of minimum energy extensions. If we denote $\mathcal{B}_{ij}$ as the set of degrees of freedom (DOFs) affected by the clusters and $\mathcal{B}_{ij}^c$ as its complement in $\Gamma_h^{(ij)}$, we can introduce another discrete harmonic extension operator $\mathcal{H}_{\mathcal{B}_{ij}^c}$ as follows:

$$\mathcal{H}_{\mathcal{B}_{ij}^c}(\mathsf{x}) := \underset{\tilde{\mathsf{y}} \in \widetilde{W}_{ij} \cap \text{Kernel}(\Pi_{ij}\mathsf{S}_{ij}\Pi_{ij} + t(\mathsf{I} - \Pi_{ij}))^{\perp}}{\arg\min} \left\{ \langle \tilde{\mathsf{y}}, \mathsf{S}_{ij}\tilde{\mathsf{y}} \rangle \mid \tilde{\mathsf{y}}_{|\mathcal{B}_{ij}} = \mathsf{x} \right\}, \tag{3.29}$$

which extends the values of $\psi_{m|\mathcal{B}_{ij}}$, defined on the set $\mathcal{B}_{ij}$, harmonically across the complement set $\mathcal{B}_{ij}^c$. The conditions applied here ensure the solvability of the original GEVP and were discussed in Subsection 3.2.1.

With operator $\mathcal{H}_{\mathcal{B}_{ij}^c}$ at hand, we now extend all basis vectors defined in $\hat{\Psi}$

$$\hat{\psi}_m \leftarrow \mathcal{H}_{\mathcal{B}_{ij}^c}(\hat{\psi}_{m|\mathcal{B}_{ij}}) \quad \forall m = 1, \ldots, (n_C^{(ij)} \cdot n_{\text{rbm}}). \tag{3.30}$$

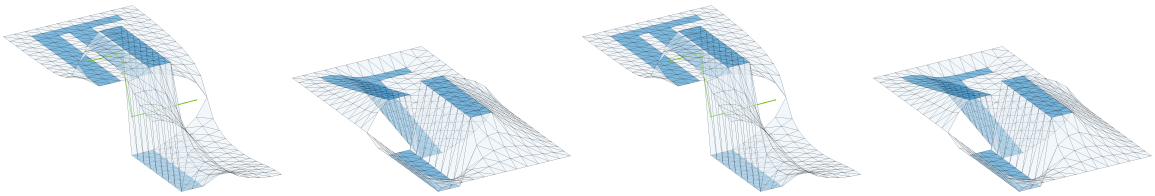
Of course, each application of this energy-minimizing operator is not for free. One possible way to apply $\mathcal{H}_{\mathcal{B}_{ij}^c}$ is to directly perform the second round of static condensation, i.e., construct a Schur complement of the (properly projected) Schur complement on the right-hand side

of Eq. (3.13). However, this would require us to explicitly assemble this product and then factorize a relatively large block belonging to the set $\mathcal{B}^c_{ij}$. Although a similar approach is applied in the preconditioning step of the Krylov-Schur algorithm, for our purposes, we prefer to avoid the direct assembly of this product and make maximum use of the sparsity of the matrices. Therefore, instead, we utilize a partially matrix-free deflation conjugate gradient algorithm. The principle of the deflated conjugate gradient method was introduced previously in Subsection 2.5.1 discussing projector preconditioning. There, we formally incorporated projections onto the admissible search space directly into the preconditioner and utilized standard preconditioned conjugate gradients (CG). However, we do not describe the deflated CG algorithm here, as it is considered a generally well-known. More details can be found in the relevant literature; cf. [48].

The advantage now is that we do not require any multiplication with projection matrices; instead, in each iteration the algorithm only requires zeroing out the contributions in search directions at DOFs with prescribed values, $\mathcal{B}_{ij}$. It is sufficient to iteratively solve the system of linear equations with a homogeneous right-hand side, where only the application of the right-hand side product of the Eq. (3.13) to the vector is necessary. In our preliminary results, preconditioning with incomplete Cholesky factorisation was needed to obtain satisfactorily accurate basis vectors that are applicable. Thus, it might not be a more efficient way for the solution of GEVP, but it is presented as an option.

For illustration, we can refer back to problem showed in Fig. 3.10, which we have previously passed over without commenting on the basis functions. The four basis functions displayed there were obtained using this methodology, although the previous version would have been equally adequate in this case.

Closer examination of the basis functions reveals distinct, seemingly nonphysical jumps between subdomains (most notably in the third and fourth basis functions). These arise as artefacts of each cluster being defined using only one element, an approach adopted for its simplicity. In this case, it would arguably have been preferable to construct clusters from two elements. However, at this point, this is merely a cosmetic detail.



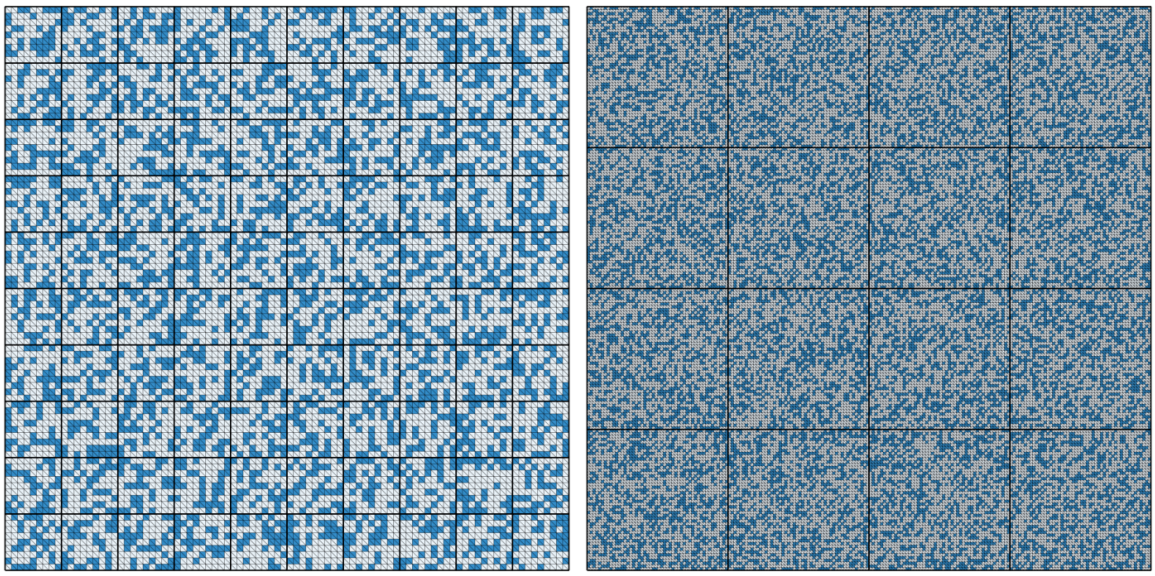
(a) : RHS (full basis) Energy: $\mathbf{1.00}$ **(b) :** LHS (full basis) Energy: $\mathbf{2.25 \cdot 10^3}$ **(c) :** RHS (red. basis) Energy: $\mathbf{35.6}$ **(d) :** LHS (red. basis) Energy: $\mathbf{8.02 \cdot 10^4}$

Figure 3.12: Obtained adaptive constraints with full and reduced basis on interface between subdomains depicted in Fig. 3.10. Resulting energies are provided. **Annotation:** LHS - left hand side, RHS - right hand side

Fig. 3.12 displays the low and high energy sides of one adaptive constraint obtained on a problem shown in Fig. 3.10. The first two subfigures depict both sides of the eigenproblem for eigenmode obtained on the full basis, whereas Figs. 3.12c-d show the same resulting from the reduced eigenproblem of dimension 4×4 using the reduced basis depicted in Fig. 3.10. It is noteworthy that the eigenvalues ($2.25 \cdot 10^3 \approx 8.02 \cdot 10^4/35.6$) are practically identical, and the modes also appear visually similar. Additionally, the strange jumps have disappeared. This suggests that this approach is moving in the right direction.

3.4.1 Applicability of the reduced-basis approach

Let us continue to focus on the binary problems for a while and conduct numerical tests to verify that this approach is correct and yields accurate results. Instead of synthetic problems with various boxes or channels crossing interfaces, we generate random binary voxel-based



(a) : Fine subdomain resolution **(b) :** Fine mesh resolution

Figure 3.13: Two stationary diffusion problems with random binary material distributions. **Left**: Decomposition with increased subdomain resolution, resulting in a larger number of smaller subdomains. **Right**: Test problem utilising a fine mesh, partitioned into 5000 triangular linear finite elements per subdomain and decomposition into 4×4 subdomains using the reduced basis depicted in Fig. 3.10.

material distributions. This helps us to avoid solving only eigenproblems where the nature of the conditions is highly predictable. Therefore, we generated two realizations with a ratio of conductive to non-conductive voxels fixed to $0.4 : 0.6$; see Fig. 3.13. This ratio was selected to promote connectivity among elements without resulting in overly large connected aggregates. The contrast in material coefficient is routinely set at a value of 10^6. In particular, we investigated two variants within this set-up.

One variant is designed to ascertain whether constraints on all edges are evaluated accurately; to this end, we chose a decomposition into 10×10 subdomains with a ratio of the subdomain size to the element size (H/h) of 9. Here, H represents a typical size of a subdomain, and h indicates a typical size of an element within the subdomain.

The second variant aims at determining whether more complex constraints on edges with a larger number of elements will be sufficiently accurate; therefore, it is decomposed into 4×4 subdomains, with $H/h = 50$ in this case. Note that the binary distribution combined with subdomains with fine mesh resolution goes against out motivation behind our reduced basis.

Figure 3.14 (and its subsequent counterparts) show dimensions of the original GEVP and the proposed reduced basis alternative for each edge, along with dominant eigenvalues λ (as dots) and their reduced basis estimates μ (as circles).

In the first task with a refined mesh, our approach successfully identified all modes corresponding to eigenvalues greater than the threshold tol set to 50, as illustrated in Fig. 3.14. The original formulation of the eigenproblem identified up to thirteen adaptively acquired modes at certain interfaces. Our reduced basis approach effectively approximated all modes, including those barely exceeding the given threshold (see the lowest marker at edge 29). Furthermore, despite the random binary distribution of material coefficients approaching a scenario that results in the maximum possible number of clusters in the given mesh, the dimensions of the reduced bases remain at least one order of magnitude lower than of the original setting. In the second variant, displayed in Fig. 3.13a, our approach performs similarly well; see Fig. 3.15. Both approaches yielded nearly indistinguishable eigenvalues. It is noteworthy that both GEVP formulations led to an identical number of 267 constraints. The practical equivalence of the eigenmodes is upheld by the numerical results: In both cases, the solver required 12

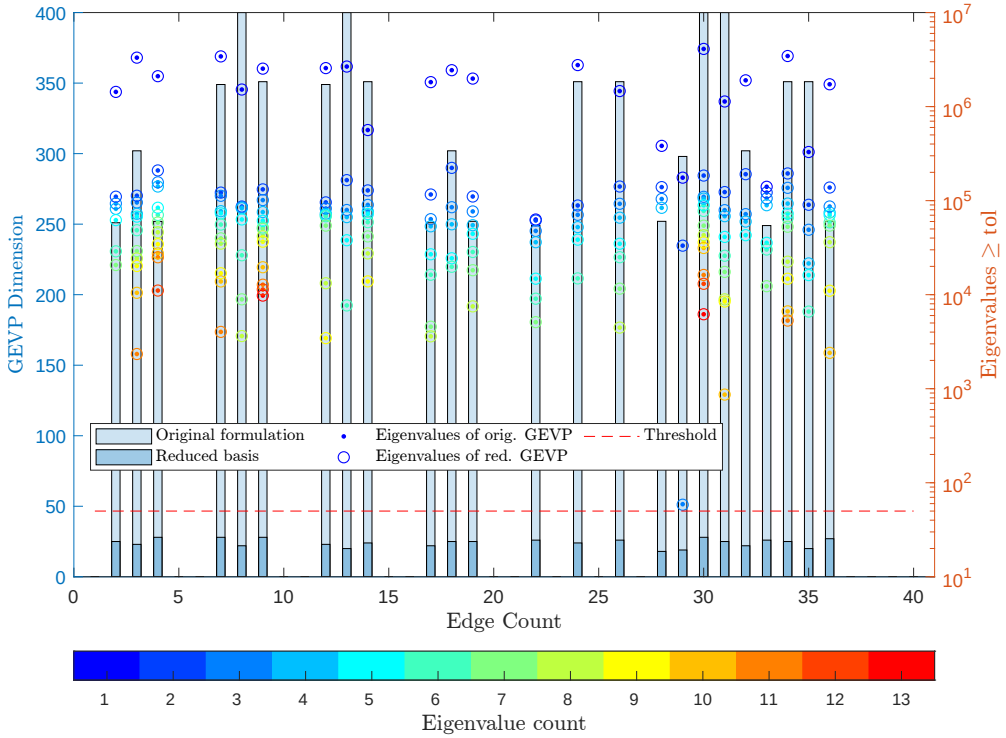


Figure 3.14: Analysis of eigenvalues of **scalar** problem 3.13b obtained by the original and reduced basis GEVPs. The histograms represent the dimensions of the GEVPs solved. All eigenvalues exceeding the specified threshold tol are visualized with markers, each marker color-coded according to their order on the interface. The close match of eigenvalues across both methods demonstrates that the adaptive constraints generated by the original and the novel reduced-basis approach are almost identical.

iterations to satisfy the criterion $\epsilon_{L_2} \leq 10^{-6}$ with a condition number equal to 14.7. We do not provide these results there because they do not contain any additional relevant information.

All is not so perfect when elasticity problems are considered; especially when problems involving spatially varying coefficients with high contrast are addressed. As a prominent representative of such a problem we take a final iteration of a topology optimization process, which will be presented in Section 4.1. For now, it suffices to say that it is a linear elasticity problem with highly varying material coefficients, with the contrast in material coefficients approaching 10^6. In this particular case, our heuristic does not exhibit the same level of effectiveness as observed in previous binary scalar-valued problems. The results of the analysis of eigenvalues obtained from both methods are shown in Fig. 3.16. Encouraging results are demonstrated by the reduced-basis approach, which successfully identifies the most extreme outlier eigenvalues. In few instances of moderate eigenvalues, the approach has successfully identified all (up to three) eigenvalues, even when they are marginally above the threshold. In a substantial number of cases, it responds quite well to the second and third eigenvalues, though the accuracy of approximation deteriorates. Nonetheless, there are cases where the reduced basis approach fails to detect relatively high second and third largest eigenvalues; see all the columns in which there are cyan and yellow dots without correspondingly colored circles.

On the other hand, the dimensionality reduction in this example is by almost an additional order of magnitude larger than in the previous scalar problem. Furthermore, in practice, even more substantial savings can be expected, as we are still using a relatively coarse mesh (30×30 elements per subdomain) solely to ensure that the values of the dimensions of the reduced bases are visible in the histograms. While the original formulation solved a large number of GEVPs of dimension 480, the maximum dimension in the new approach is 15. Another

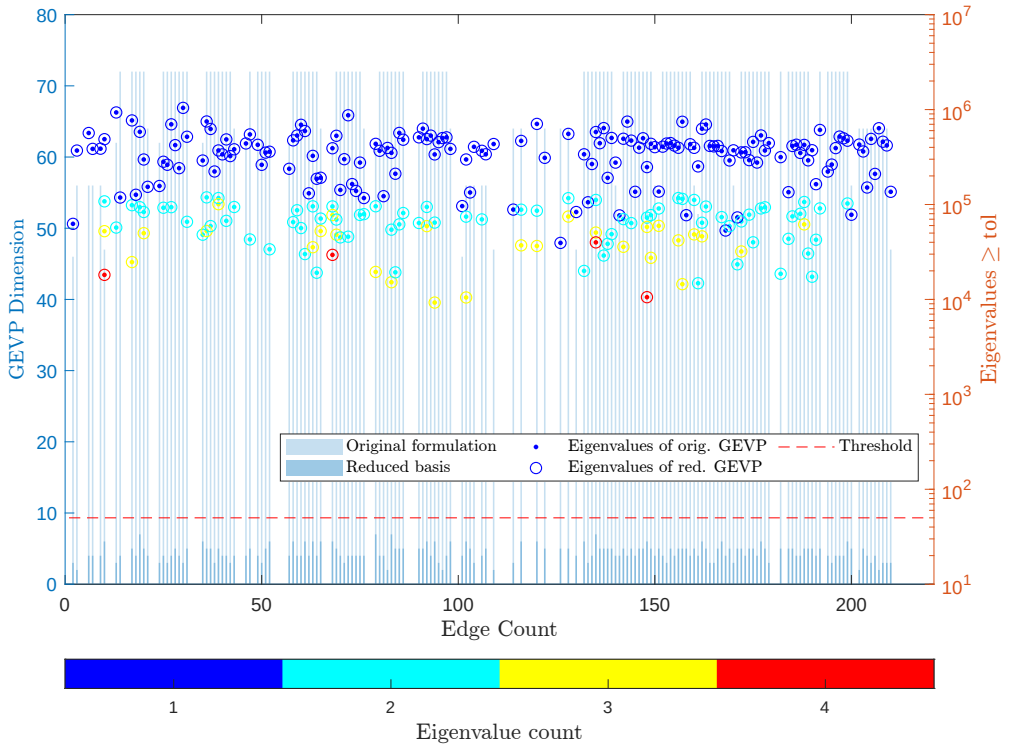


Figure 3.15: Analysis of eigenvalues of **scalar** problem 3.13a from original and reduced-dimensional GEVPs. The histograms represent the dimensions of the GEVPs solved. All eigenvalues exceeding the specified threshold are visualized, each color-coded according to their order on the interface. The close matching of eigenvalues across both methods demonstrates that the adaptive constraints generated by the original and the novel reduced-basis approach are of almost identical quality.

major advantage is also that the criteria for basis construction act as a fairly good estimator of when GEVPs can be completely discarded, i.e., neither calculated nor set up. The reduced basis approach calculated a total of 67 localized GEVPs, whereas in the full basis without any technique to exclude unnecessary eigenproblems, 170 GEVPs were calculated. Furthermore, it turns out that when the heuristic determines that a basis needs to be constructed, the reduced basis GEVP quite reliably yields eigenvalues exceeding a certain threshold, ensuring that it is not assembled unnecessarily. Specifically, this construction accurately evaluated 64 interfaces out of a total of 72, where at least one eigenvalue larger than $\mathsf{tol} = 50$ was found in the original formulation, and only 3 interfaces where the calculated GEVP did not contain any eigenvalue larger than the threshold in the original formulation. The eight cases where the reduced basis was mistakenly not constructed correspond to edges with high eigenvalues exceeding 10^3, that are clearly visible on the right side of Fig. 3.16. Clearly, there is something amiss with the bases themselves at these interfaces. Closer examination of the dominant missing mode reveals that it is a very atypical case; see this degenerative mode in Fig. 3.17. As observed, there is a relatively continuous stiff artefact along the shared edge, resulting in a minor maximum jump in coefficient profiles along this edge reaching the value of approximately 40 (recall that the threshold was set to 50). Consequently, the reduced basis is evaluated as unnecessary because the criterion for a jump in coefficients is not met. At this moment, the issue lies with the criterion for determining individual clusters, which should be set lower. As further evidence, we include another similar plot in Fig. 3.18, where the threshold for clustering is lower, in particular the threshold equals to $\mathsf{tol}/2$. With a more flexible lower threshold for the clustering criterion, the quality of the approximated eigenvalues immediately improves. Interestingly, the maximum dimensionality of localized GEVPs remained at 15, showing no increase. In this scenario, the number of solved GEVPs increased from 64 to 77. This specifically indicates that the criterion correctly evaluated 72 interfaces, where it succeeded in finding at least one desired eigenmode, and only at five interfaces where reduced-basis GEVPs

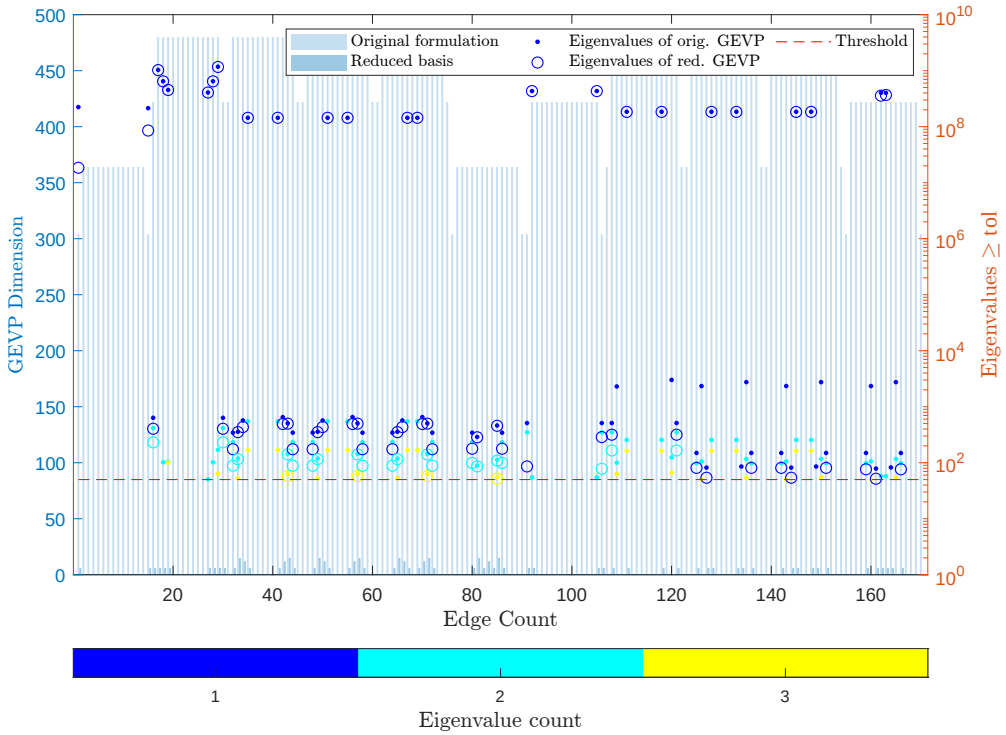


Figure 3.16: Analysis of eigenvalues of a topology optimization **elasticity** problem (it. 100) from original and reduced-dimensional GEVP. The histograms represent the dimensions of the GEVP solved. All eigenvalues exceeding the specified threshold are visualized, each color-coded according to their order on the interface. The same tolerance for selection of eigenmodes and for clustering is used.

were calculated but no eigenvalues were found in the original formulation. However, a crucial outcome is that no interfaces, at which the original formulation enforced constraints were neglected. This reduced basis identified a total of 119 out of 158 eigenmodes greater than the given tolerance. Furthermore, it predominantly ignored low eigenvalues, which is more convenient for the algorithm. The resulting condition number estimate is 197.2.

It might appear as merely tweaking a parameter, but there is no conceptual inconsistency in setting the tolerance for clustering to values lower than those used in the GEVP itself. Therefore, when dealing with varying coefficients, or when there is a risk of potential occurrence of such degenerative distributions, we suggest a careful lowering of this clustering threshold. With binary or sharply changing coefficients, we believe that the threshold for clustering can be approximately the same or slightly lower than the threshold for the GEVP itself.

Despite the lack of any theoretical analysis, we believe that the presented reduced-basis approach opens a way toward new efficient heuristics. Inspired by observed behavioral solution patterns, it primarily focuses on extreme outliers with the end goal of substantial reduction in the dimensionality of the coarse space augmentation. We believe that even these initial results are promising and they encourage further research into more cost-effective adaptive techniques. Next steps are to appropriately adjust the construction of reduced basis, either to make it computationally cheaper or to better approximate the subspace of all outlying eigenvalues.

3.5 Heuristic selection of primal nodes

Finally, we present an extremely cheap heuristic approach for the systematic selection of primal nodes in the context of problems with coefficient distributions with high contrast in material parameters. Contrary to the previous sections, we directly enrich the set of primal variables Π by evaluating information pertinent to nodes on a single edge at a time, requiring

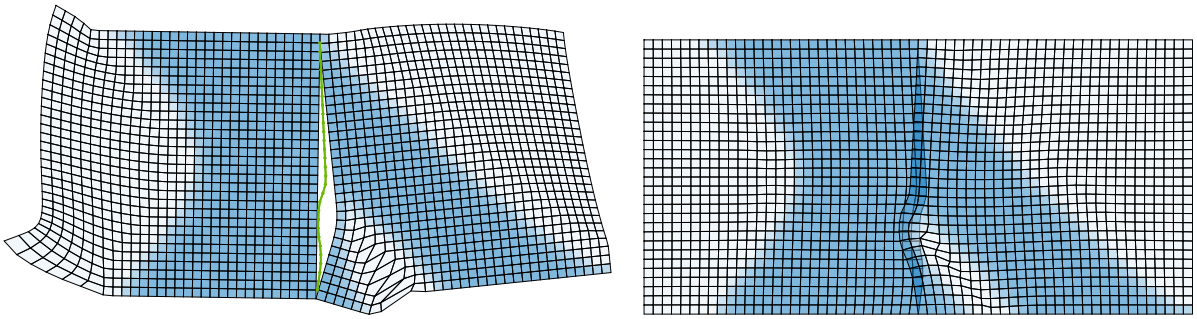


Figure 3.17: Visualization of the dominant eigenmode ($2.25 \cdot 10^{-6}$) of the localized GEVP between two subdomains, one of the having a tricky coefficient distribution. **Left:** Relatively low energetical mode (energy $7.72 \cdot 10^{-6}$) and splitting of the gap based on ρ scaling (green line). **Right:** Right hand side mode of the localized GEVP leading to energy $1.73 \cdot 10^{-2}$.

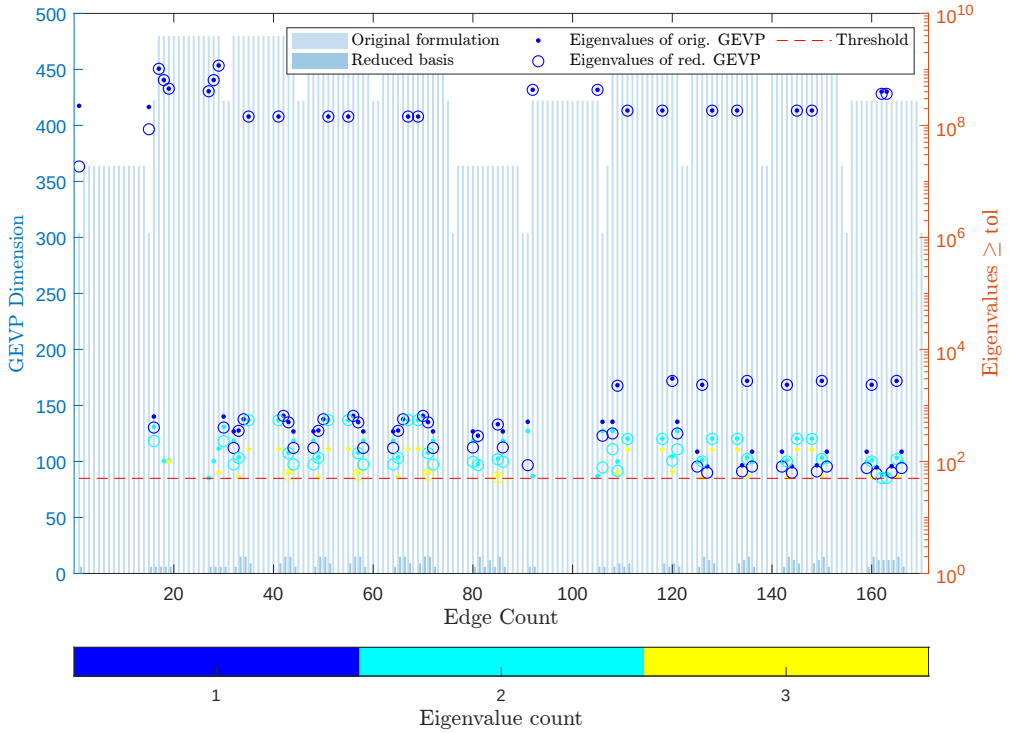


Figure 3.18: Analysis of eigenvalues of a topology optimization elasticity problem (it. 100) from original and reduced-dimensional GEVP. The histograms represent the dimensions of the GEVP solved. All eigenvalues exceeding given threshold are visualized. **Lower threshold for clustering used:** $\mathsf{tol} = 50$ for GEVP and $\mathsf{tol}_C = 25$ for clustering of elements.

only very limited information about the material coefficients. Notably, the criterion introduced here does not require any additional information beyond what is typically provided to the solver if the widely used ρ scaling is adopted.

The primary motivation behind this approach is to identify edges that are potentially problematic due to the occurrence of ill-posed modes (recall Subsection 2.5 for context) and to consequently add nodal primal continuity constraints on these edges. By focusing solely on nodal constraints, we maintain the simplicity and cost-effectiveness of the method. Our strategy is to keep the solver settings as straightforward as possible, even at the cost of more iterations, since the setup phase in the solver is kept to a minimum.

This heuristics builds on a version developed in the author's bachelor thesis [36]. By transforming carefully selected DOFs from the dual to the primal character, the original method allowed for a decrease in the condition number of the FETI-DP system by several orders of

magnitude. Given the newly acquired experience with adaptive approaches, GeNEO-type [51] and P_D-based [35, 43] in particular, we believe that the original heuristics can be improved despite delivering desired performance in many cases. With better understanding of the role of ill-posed modes and their mitigation, we present here an enhanced yet equally computationally inexpensive version for the selection of the set Π of primal variables.
Using only nodal constraints allows us to work in the initial basis without any transformations or the use of projections. The advantages of avoiding these transformations, as well as the problems associated with the introduction of scaling, are discussed in Section 2.5. Thus, we aim for a very low-dimensional nodal approximation of the necessary adaptively obtained constraints, targeting the most harmful nodes specifically.

For our heuristics, we assume that the coefficient jumps inside subdomains are at least partially reflected on the interface. Our procedure for selecting suitable primal nodes starts with processing edge characteristics and flagging the edges that can be omitted from the enrichment. For each dual interface between subdomains $\Omega^{(i)}$ and $\Omega^{(i)}$, we first construct nodal coefficient profiles $\Xi^{(l)}$, $l \in \{i, j\}$ based on the material properties pertinent to elements adjacent to edge E_{ij}. In particular, assuming that ordering of nodes matches across the subdomains, the coefficient profile $\Xi^{(l)}$ collects nodal coefficient maxima in a sense of expression (3.6), i.e. each component follow as

$$\Xi_k^{(l)} = \hat{\alpha}^{(l)}(x_k) \tag{3.31}$$

for all nodes x_k in locally ordered index set $\mathtt{E}_{\mathtt{ij}}^{(l)} := \{1 \dots n_{E_{ij}}^{(l)}\}$, $l \in \{i, j\}$ on edge E_{ij}.
Following our observation from adaptive constraints that large components in constraint modes are related to Lagrange multipliers binding elements with relatively high coefficients, we first eliminate the nodes with relatively low coefficient values as they are not expected to be suitable candidates for new, pair-wise nodal constraints. To retain maximal simplicity, we employ a user-defined relative tolerance factor σ and assess if a relevant jump in coefficients occurs on this edge simply by comparing

$$\frac{\max\Big(\max(\Xi^{(i)}), \max(\Xi^{(j)})\Big)}{\min\Big(\min(\Xi^{(i)}), \min(\Xi^{(j)})\Big)} \geq \sigma\,.$$

If the ratio in the expression above does not exceed the given tolerance σ, we do not enrich the set of primal variables Π on this edge. On the other hand, if this criterion is satisfied, we proceed with this edge start eliminating nodes with marginal values in the coefficient profiles. In particular, we keep as candidate points at edge E_{ij} only those that have profile values higher than the overall minimum. The domain-wise index sets of potential candidates $\mathtt{C}_{\mathtt{ij}}^{(l)}$ is then defined as

$$\mathtt{C}_{\mathtt{ij}}^{(l)} = \left\{k \mid \Xi_k^{(l)} \geq \sigma \cdot \min\left(\min\left(\Xi^{(i)}\right), \min\left(\Xi^{(j)}\right)\right)\right\} \qquad l \in \{i, j\}, \tag{3.32}$$

Eventually we combine the two sets of candidates, one for each subdomain, by intersecting them:

$$\mathtt{C}_{\mathtt{ij}} = \mathtt{C}_{\mathtt{ij}}^{(i)} \cap \mathtt{C}_{\mathtt{ij}}^{(j)}$$

Note that in this operation we might have completely excluded some of the high coefficient nodes on individual subdomains if the coefficient profile is low on the corresponding nodes in the second subdomain. At the same time, we might have split sequences of consecutive nodes on individual subdomains into multiple, non-consecutive sequences. This is in accordance with our motivation, we want to identify isolated floating high-coefficient segments, which are usually related to ill-posed modes and thus are often the source of high eigenmodes found by adaptive approaches. The resulting set $\mathtt{C}_{\mathtt{ij}}$ contains nodes for which a nodal-based continuity condition is considered to be meaningful, i.e., it is expected to be relevant for substantial enhancement of desired robustness of the coarse problem.

Next, we proceed to identifying clusters $\mathtt{C}_{\mathtt{ij}}$ of contiguous nodes. This concept of clustering has been already adopted in Section 3.4, where we needed a floating threshold. Here, the

clustering algorithm is less complicated, as each cluster $\mathtt{c}_{\mathtt{ij},k}$ represents a contiguous set of candidates

$$\mathtt{c}_{\mathtt{ij},k} = \left\{ m \in \mathtt{C}_{\mathtt{ij}} \mid \exists n \in \omega^{E_{ij}}(m), n \in \mathtt{C}_{\mathtt{ij}} \right\},$$

where $\omega^{E_{ij}}(m)$ denotes the congruent nodes at E_{ij}. For each identified cluster $\mathtt{c}_{\mathtt{ij},k}$, we select one representative node and include this node in the set of primal nodes Π. The representative node index p_k^* is chosen as an index, which pertain to the maximum value in coefficient profile in the cluster $\mathtt{c}_{\mathtt{ij},k}$, i.e.

$$p_k^* = \max_{m \in \mathtt{c}_{\mathtt{ij},k}} \min(\Xi_m^{(i)}, \Xi_m^{(j)}).$$

This specific choice is strongly motivated by the concepts discussed in the introduction of weighted averages and adaptive constraints. From weighted averages, we know that the focus should be directed towards high-coefficient elements. However, their formulation allows only for a predefined number of constraints; in practice, more constraints might be beneficial. Our approach reflects this, as it can to a certain extent recognize how many non-connected high-coefficient blocks meet at the interface.
Finally, we collect all $n_{\text{clusters}}^{(ij)}$ selected nodes into the final set of primal nodes

$$\mathcal{P} = \left\{ x_{p_k^*} \mid k = 1, \ldots, n_{\text{clusters}}^{(ij)} \right\}.$$

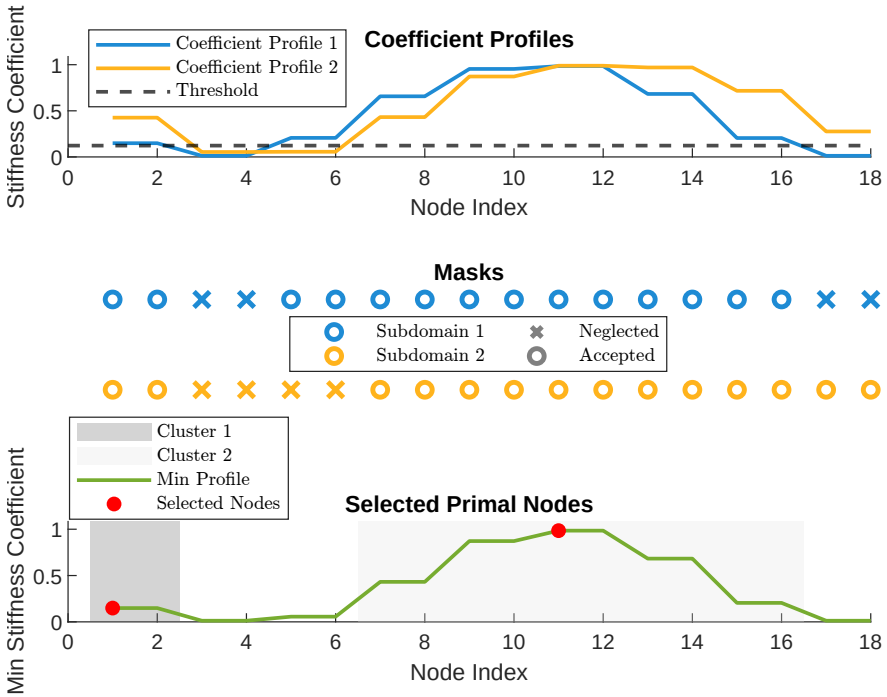


Figure 3.19: Visualization of the step-by-step process used for the identification and heuristic selection of additional primal nodes on an edge between two subdomains.

The workflow of this heuristic selection is illustrated in Fig. 3.19. The first subplot shows the coefficient profiles for two subdomains along the shared edge. Even though the profiles are generally similar, some significant jumps still occur. The maximum and minimum values in coefficient profiles is in this case equal to 0.99 and 0.012, respectively. Hence, the ratio exceeds the user-defined tolerance σ, in this case chosen to be 10. The threshold $\sigma \cdot \min(\Xi^{(i)}, \Xi^{(j)})$ for filtering parts with too low coefficients out of the edge is also visualized with dashed line. The second subplot displays the mask values of accepted and denied candidates for each subdomain. Nodes that meet the criterion are marked as accepted (circles), while those that do not meet the criterion are marked as neglected (crosses). The first subdomain is represented in blue and the second subdomain in orange. In the third subplot, the minimum

coefficient profile is depicted in green. Clusters are highlighted with gray-shaded areas, and finally, the selected primal nodes within each cluster are marked with red circles. In this case, the first node is chosen to be adjacent to the corner node, which is also primal in this problem. Generally, it is better for the conditioning of the local problem if the primal nodes are not too close to each other. Even though this might not be optimal, we keep these transformed nodes as they still have a positive effect on preventing rotational modes in cases when the high-coefficient part is connected to only one primal node.

At first glance, the final heuristics presented here might seem overengineered, but it is extremely simple from an implementation point of view. The only thing we need initially are the material coefficients at elements adjacent to the considered edge, which is standardly provided also for, e.g., ρ scaling. If this is not the case, diagonal entries of local stiffness matrices can be used instead.

This enrichment of Π with additional, heuristically identified nodes is particularly efficient for severe variations in material coefficient distribution. By ensuring that each cluster contributes one primal node, the approach maintains a balance between the posed computational overhead and the enhanced robustness. Considering a single primal node per each identified cluster might not resolve the problems with rotational modes if the aggregate of high-coefficient elements in subdomain $\Omega^{(i)}$ are not connected to the complement of edge E_{ij} with respect to $\Gamma^{(i)}$, that is if there is a high-coefficient aggregate that crosses one edge but then vanishes within a subdomain. However, as this is a relatively rare occasion, we rely on the interaction of coarse degrees of freedom: if this high-coefficient aggregate contains more heuristically adopted coarse degrees of freedom on multiple edges, these combined suffice to prevent the rotational mode as well. Hence, in practice it is not necessary to impose as many constraints to prevent all the nearly rigid body modes on each edge separately.

To conclude, the application of this heuristic approach can significantly improve the robustness and convergence of adaptive coarse spaces in domain decomposition methods, particularly in the presence of highly heterogeneous material properties, while maintaining the simplicity of nodal-valued primal DOFs.

Chapter 4

Numerical tests

Throughout this thesis, all proposed modifications were illustrated with numerical examples specifically designed to highlight the limitations of the original methods and the impact of our improvements. Most of these examples focused on problems with a binary distribution of coefficients. However, the primary motivation for our research was the application of FETI-DP to systems of equations arising in modular topology optimization. As discussed next, intermediate stages of topology optimization present more challenges than standard binary coefficient distributions. At these stages, there is already a spatial distribution of high-contrast coefficients, but the topology of high-coefficient regions is not fully established. Consequently, heuristics developed for binary problems may struggle with these more complex scenarios.

4.1 Topology optimization problems

Topology optimization is a crucial tool for designing the optimal material distribution within a provided space based on various criteria, such as minimizing compliance under given constraints. However, tasks arising in topology optimization are naturally poorly conditioned due to the high contrast in material properties. For example, in the Solid Isotropic Material with Penalization (SIMP) method [4], the material distribution is parameterized by a scalar field of relative density $\rho(x)$, where $0 \leq \rho \leq 1$. This field affects the stiffness tensors as follows:

$$\mathbf{E}(x) = \mathbf{E}_{\min} + \rho^p(x)\,(\mathbf{E}_0 - \mathbf{E}_{\min})\ , \tag{4.1}$$

where p is a penalization coefficient (typically $p \geq 3$), used to disfavour intermediate values and promote a clear "0-1" design. The constants $\mathbf{E}_0$ and $\mathbf{E}_{\min}$ are the stiffness tensors of the solid material and voids, respectively. Minimum stiffness is used for numerical purposes to prevent an indefinite Hessian matrix, balancing between substituting the voids and avoiding ill-posed problems.

The optimal design is typically sought for iteratively in a staggered approach, solving state equations with fixed densities and updating design density variables based on the objective sensitivities with respect to the design parameters. In this chapter, we use several snapshots of the optimization iterations as test cases for the investigated coarse-space enrichments. In particular, we focus on a specific modular-topology optimization (MTO) problem where the domain is divided into a 6×16 grid of subdomains. In addition to the aforementioned high contrast in coefficients (1×10^6 in our cases), the modular formulation can lead to non-perfect continuity of material distribution at interfaces due to repeated patterns. This is very sensitive for FETI-DP methods, as seen in Figure 4.1, which displays the spectral distribution of the preconditioned system $(\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{F})$ for seven topology optimization snapshots, six of which are shown in Fig. 4.2 (some were omitted for brevity as the visual difference were minimal).

Looking at the spectra, one can see that initially the problem is easily solvable with eigenvalues close to one. Toward the tenth iteration, the spectrum becomes very broad, making the problem challenging for iterative solvers. While there are some very isolated

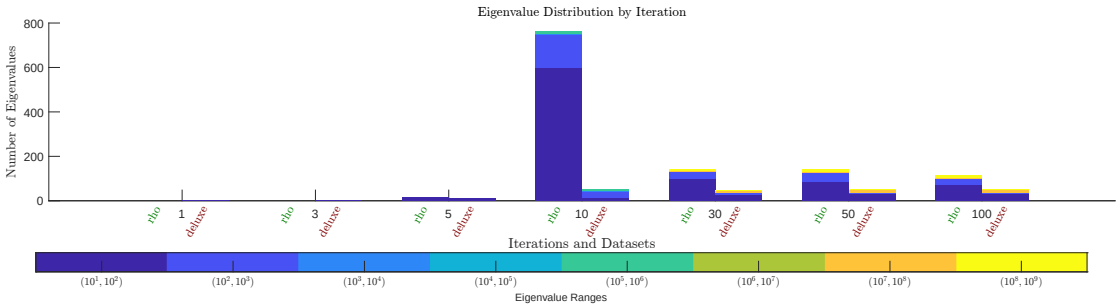


Figure 4.1: Spectral distribution of the preconditioned system ($\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{F}$) of seven topology optimization snapshots for ρ and deluxe scaling. Only eigenvalues greater than 10 are visualized.

eigenvalues in the later iterations of topology optimization, the spectrum becomes well-clustered, indicating that poor modes are predictable and thus easier to solve. We mention this to counter the common conviction that binary-valued coefficient distributions are the most challenging ones for iterative solvers. Our experience with modular topology optimization problems is different.

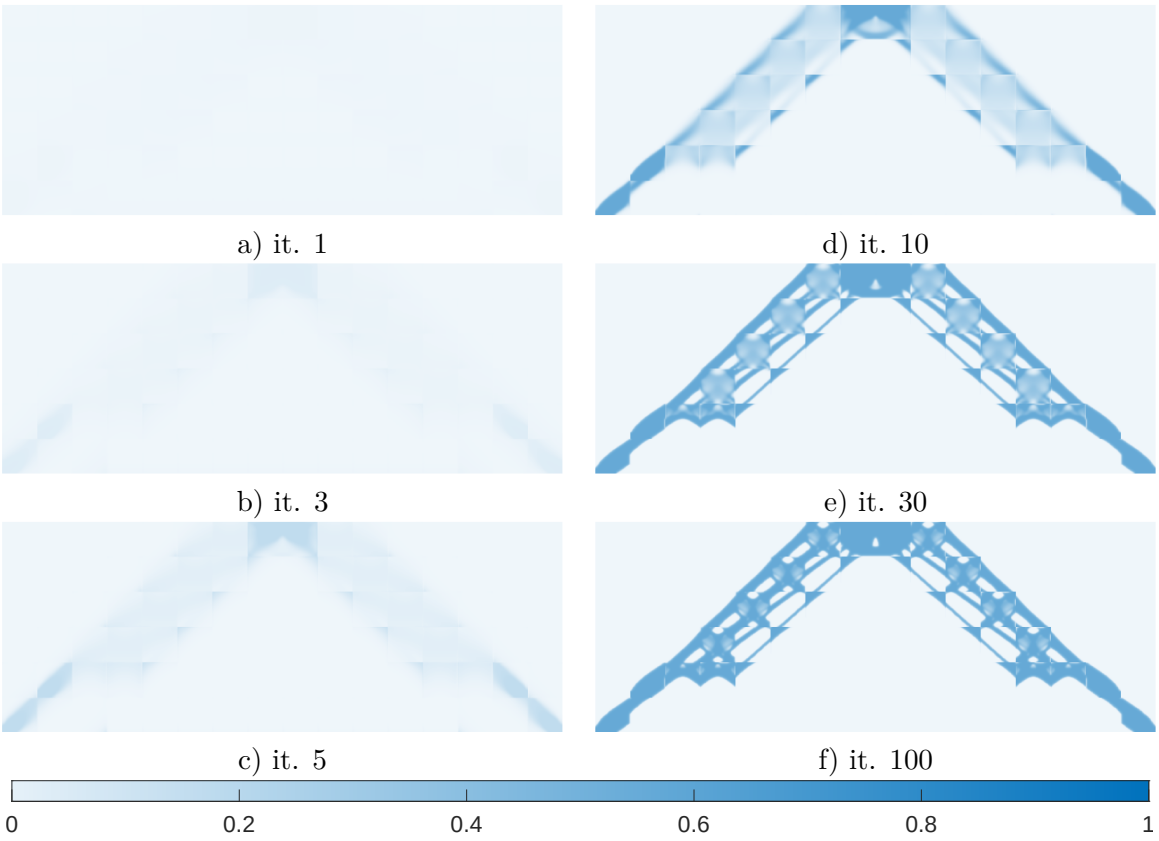


Figure 4.2: Six snapshots of iteration in a modular-topology optimization problem, in which the compliance of a simple supported beam loaded in the middle of the top part was minimized. The number of iteration each snapshot belongs is given in the left column.

scaling			MTO iteration 1	3	5	10	30	50	100
none	multiplicity	**it.**	16	24	105	1975(*)	1985(*)	1888(*)	1865(*)
		κ	6.8	$2.0 \cdot 10$	$5.9 \cdot 10^2$	$2.1 \cdot 10^7$	$7.8 \cdot 10^7$	$4.4 \cdot 10^7$	$2.9 \cdot 10^7$
		$\lvert\Pi^*\rvert$	0	0	0	0	0	0	0
	rho	**it.**	14	15	21	439	1994(*)	1997(*)	1999(*)
		κ	6.1	6.4	$1.2 \cdot 10$	$1.2 \cdot 10^5$	$8.2 \cdot 10^7$	$8.0 \cdot 10^7$	$8.1 \cdot 10^7$
		$\lvert\Pi^*\rvert$	0	0	0	0	0	0	0
	deluxe	**it.**	16	15	19	232	2000(*)	2000(*)	1997(*)
		κ	7.2	6.7	9.9	$4.0 \cdot 10^4$	$4.8 \cdot 10^7$	$4.6 \cdot 10^7$	$4.7 \cdot 10^7$
		$\lvert\Pi^*\rvert$	0	0	0	0	0	0	0
arithmetic averages	multiplicity	**it.**	6	12	68	1993(*)	1988(*)	1962(*)	1900(*)
		κ	2.1	8.9	$3.6 \cdot 10^2$	$1.4 \cdot 10^6$	$1.4 \cdot 10^8$	$1.3 \cdot 10^8$	$8.9 \cdot 10^7$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
	rho	**it.**	4	5	10	163	85	93	104
		κ	1.3	1.4	4.7	$2.5 \cdot 10^3$	$1.6 \cdot 10^4$	$2.2 \cdot 10^4$	$2.2 \cdot 10^4$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
	deluxe	**it.**	4	5	9	62	42	45	49
		κ	1.3	1.4	4.0	$8.4 \cdot 10^2$	$1.3 \cdot 10^4$	$1.5 \cdot 10^4$	$1.7 \cdot 10^4$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
weighted averages	multiplicity	**it.**	6	11	65	630	1968(*)	1994(*)	1988(*)
		κ	2.1	7.2	$3.7 \cdot 10^2$	$2.0 \cdot 10^5$	$1.4 \cdot 10^8$	$1.5 \cdot 10^8$	$1.3 \cdot 10^8$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
	rho	**it.**	4	5	9	91	56	61	62
		κ	1.3	1.5	5.4	$2.1 \cdot 10^2$	$1.7 \cdot 10^2$	$2.1 \cdot 10^2$	$2.1 \cdot 10^2$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
	deluxe	**it.**	4	5	9	19	19	24	26
		κ	1.3	1.5	4.6	$1.3 \cdot 10^2$	$1.2 \cdot 10$	$2.0 \cdot 10$	$1.9 \cdot 10$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
frugal	multiplicity	**it.**	7	13	71	571	1918(*)	1999(*)	-1799
		κ	2.1	6.4	$3.4 \cdot 10^2$	$1.2 \cdot 10^5$	$9.1 \cdot 10^7$	$1.1 \cdot 10^8$	$9.5 \cdot 10^7$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
	rho	**it.**	5	7	9	84	55	60	56
		κ	1.4	1.8	3.5	$1.7 \cdot 10^2$	$2.1 \cdot 10^2$	$2.7 \cdot 10^2$	$2.5 \cdot 10^2$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
	deluxe	**it.**	6	7	8	11	16	19	18
		κ	1.4	1.7	2.2	2.5	$1.2 \cdot 10$	$2.0 \cdot 10$	$1.9 \cdot 10$
		$\lvert\Pi^*\rvert$	510	510	510	510	510	510	510
MS GEVP (tol = 50)	multiplicity	**it.**	16	24	55	68	59	60	70
		κ	6.8	$2.0 \cdot 10$	$1.1 \cdot 10^2$	$9.2 \cdot 10$	$9.2 \cdot 10$	$6.6 \cdot 10$	$8.3 \cdot 10$
		$\lvert\Pi^*\rvert$	0	0	72	554	203	234	238
	rho	**it.**	14	15	21	83	51	56	66
		κ	6.1	6.4	$1.2 \cdot 10$	$1.3 \cdot 10^2$	$4.8 \cdot 10$	$6.0 \cdot 10$	$7.8 \cdot 10$
		$\lvert\Pi^*\rvert$	0	0	0	204	44	50	55
	deluxe	**it.**	16	15	19	21	31	27	29
		κ	7.2	6.7	9.9	$6.4 \cdot 10$	$1.9 \cdot 10$	$2.4 \cdot 10$	$2.6 \cdot 10$
		$\lvert\Pi^*\rvert$	0	0	0	69	30	38	44

Table 4.1: Comparison of iteration counts and condition cumbers across different scaling strategies in the FETI-DP method. Results from various coarse space enhancements for seven snapshots of MTO problems are provided. **Annotation: it.** - number of iterations required to meet stop criterion $\epsilon_{L_2} \leq 10^{-6}$, κ indicates the condition number, and $|\Pi^*|$ denotes the number of constraints added to augment the coarse space. Prior vertex-based coarse space contains 230 constraints. If the solver reached the maximum number of iterations, 2000, an asterisk symbol (*) indicates the iteration in which the smallest value of the error estimator was achieved. **MS GEVP** stands for adaptive constraints obtained from the solution of localized GEVPs proposed by Mandel and Sousedík.

4.2 Comparison of enforcement approaches

In this subsection, we provide a brief comparison of three strategies implemented for enforcing additional constraints. Namely, we will consider (i) projector preconditioning, (ii) the balancing approach, and (iii) generalized transformation of basis. These strategies were discussed in detail in Section 2.5. The comparison is not supposed to be exhaustive, we include it here though to illustrate the difficulties we encountered with projection-based approaches. As a test case, we select three iterations from modular topology optimization: i) iteration five, where the problem is well-conditioned even without coarse space enhancement; ii) iteration ten, which represents a poorly conditioned problem yet not reaching maximum contrast; and iii) iteration thirty, which features a significantly clustered spectrum with the highest outlier eigenvalues. Distribution of eigenvalues for all three cases is shown in Fig. 4.1. Since iteration five is well-conditioned, there is no need to solve any GEVPs and pose adaptive constraints. For the sake of comparison of enforcement strategies, we pose the additional constraints in iteration five in the form of arithmetic averages. The reason for choosing the arithmetic averages is that their enforcement can be easily verified by a visual inspection of the obtained results. For the poorly conditioned iterations 10 and 30, we use the adaptive technique described in Subsection 3.2.1. Specifically, we consider all constraints with eigenvalues exceeding threshold $\mathsf{tol} = 50$.

For comparison, we report several quantifiers:

- the standard error estimator given by the product $\mathsf{r}_k^{\mathsf{T}}\mathsf{M}_{\mathrm{D}}^{-1}\mathsf{r}_k$ (recall that r_k denotes the k-th iteration residual vector), which is commonly used as a stopping criterion in its relative form.
- an relative L_2-error norm (denoted as ϵ_{L_2}) of the solution difference from the solution obtained by the direct method, which we declare as the reference one. This indicator is the most relevant one, as it shows how quickly we are approaching the desired solution.
- Frobenius norm of the product $(\mathsf{P}^{\mathsf{T}}\mathsf{F}\mathsf{P} - \mathsf{I})$, which measures the extent to which orthogonality of search directions is preserved in the conjugate gradient algorithm. Matrix P in this context is a matrix storing individual (not explicitly) F-normalized search directions as its columns.
- the minimum and maximum estimated eigenvalue in the preconditioned system, which we use to approximate the condition number of the system at hand. For this, we use a cheap eigenvalue estimate based on coefficients appearing in CG method as introduced in [46, Sec. 6.7.3].

Ideally, we are interested in the achieved solution accuracy, which, however, cannot be measured throughout iterations in practice because of two reasons: (i) the reference solution is not known and (ii) the current solution is not explicitly constructed in each iteration.

For calculations, we have set a fixed number of iterations to each specific task to ensure a thorough comparison of the different enforcement approaches. This decision allows us to compare the convergence behavior over a consistent range of iterations among different solver's setups. We aim to show that while the methods may initially converge well, if we do not stop the solver appropriately, some observed indicators may be misleading. This can potentially lead to less accurate results with a higher number of iterations, especially when working with very ill-conditioned and/or rank-deficient systems where round-off errors significantly influence the precision achieved in computations.

Since we often encounter a significant loss of orthogonality within search directions, we include a variant of the CG method with explicitly reorthogonalized search directions in each iteration. In this case, a conjugate gradient method with full reorthogonalization (CGFO) on-the-fly employing the modified Gram-Schmidt algorithm is used, see [15] for the application of CGFO in the context of FETI-family methods. By full reorthogonalization, we mean that

the search vector in the kth iteration is orthogonalized against all the previous ones. However, we observe that in many cases the quality of projectors is poor, i.e., these projections are not accurate. These inaccuracies have significant detrimental effect on the achievable accuracy of the results. While we would like to provide results with precise orthogonal projections, to the author's knowledge, there is a paucity of literature on robust reorthogonalization schemes in non-standard inner products. This limits the use of advanced reorthogonalization techniques such as Householder's reflections, which are known for their superior robustness in ill-conditioned systems compared to standard Gram-Schmidt algorithms. Moreover, the literature using Householder's algorithm often relies on availability of some accurate F-orthonormal basis at hand [49], which we cannot provide in the reorthogonalization process. Thus, in this thesis, we provide results only for the well-known modified Gram-Schmidt algorithm, which is generally less suitable for parallel treatment [45].

Iteration 5. The results for the first test case of the fifth iteration from MTO are presented in Fig. 4.3c. Here, we can confirm that the convergence behavior is fundamentally similar among all three approaches. Each enforcement strategy provides a solution nearly identical to the reference one; around the 20th iteration, ϵ_{L_2} decreases to approximately 10^{-10} and then temporarily reaches a plateau. Slightly better results are achieved by the balancing approach compared to projector preconditioning, which lags behind by about three iterations. However, the differences are minimal, and all approaches successfully converge to the reference solution. With standard conjugate gradients, a loss of orthogonality in search directions measured by the third indicator from the list above begins to manifest after about twenty iterations, but this only occurs when the solution has nearly converged, i.e., when the norm of the residuals rapidly approaches zero. Such behavior is then not surprising.
A closer examination of the six plots shows that after a certain number of iterations, in the range of 25 to 50, an eigenvalue less than one is detected in all cases apart from balancing with standard CG. This moment causes the error estimator to start increasing in all cases, and in the cases of gToB and CGFO, even ϵ_{L_2} begins to rise, indicating that our solution is clearly losing accuracy. In practice, we always terminate the computations when a safeguard eigenvalue estimate lower than one is detected in the subspace of search iterations to prevent this behavior. We use this criterion as it is a well-known fact that the minimal eigenvalue of the preconditioned system is bounded from below by one, see the end of Subsection 2.3.1.

Iteration 10. Next, we move to the next test case pertinent to the 10^{th} in topology optimization, where we enforce adaptive constraints. The convergence plots are provided in Fig. 4.4 in the same format as in the previous test case. Here, we observe practically identical behavior across all three approaches: they all achieve a very satisfactorily accurate solution with ϵ_{L_2} approximately 10^{-7} around iteration 60. In the end, only the balancing approach without reorthogonalization of search directions did not start diverging even after 200 iterations, although the solution reached by iteration 60 did not improve further. However, at this level, the accuracy compared to the numerically obtained reference solution does not play a significant role. In this setup, the solver did not detect an eigenvalue less than one. The difference now lies mainly in the loss of orthogonality of the direction vectors: for standard CG, this quantity remains comparable among all approaches, but with CGFO it seems that the two projection-based strategies outperform the generalized transformation of basis. This is not particularly relevant in this case, as the quality of the solution and the main error estimator are similar, and the computations would have been stopped earlier. It also seems that the loss of orthogonality (given by product $\|\mathsf{P}^{\mathsf{T}}\mathsf{F}\mathsf{P} - \mathsf{I}\|_{\text{Fro}}$) in this iteration does not negatively impact the quality of the results achieved, despite the fact that orthogonality is evidently violated.

Iteration 30. In iteration thirty of topology optimization, the differences in numerical performance of the three strategies become the most prominent. This time, we split the

composed plots from previous cases into two individual figures, each for one conjugate gradient solver. In Fig. 4.5, convergence details are provided for a standard conjugate gradient method, and in Fig. 4.6 for CG with reorthogonalization. Generalized transformation of basis managed

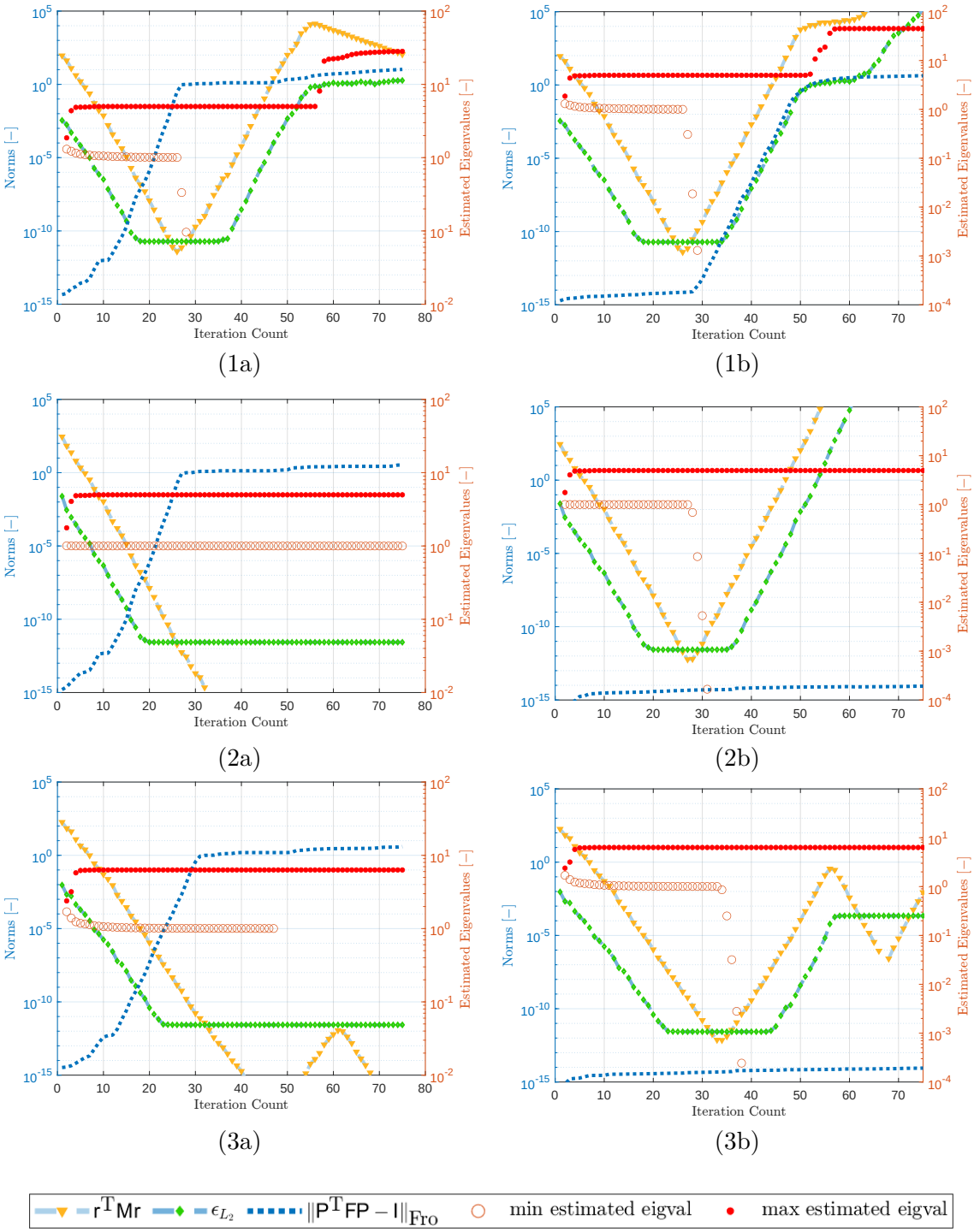


Figure 4.3: Convergence behavior of a first case of iteration 5 in MTO in terms of quantifiers introduced in the list at the beginning of Section 4.2. Each row corresponds to one enforcement approach: **(1)** generalized transformation of basis, **(2)** balancing, and **(3)** projector preconditioning. Each column provide convergence plot for one variant of the conjugate gradient (CG) iterative solver: **(a)** standard preconditioned CG algorithm and **(b)** CG with employed full reorthogonalization. For all simulations, a vertex-based prior coarse space with ρ scaling augmented with arithmetic averages is adopted.

to achieve a very accurate solution, specifically with ϵ_{L_2} reaching approximately 10^{-8} around iteration 60 with both CG and CGFO. As usual with this approach, there is a relatively good correlation between the representative indicator, given by ϵ_{L_2}, and the error estimator.

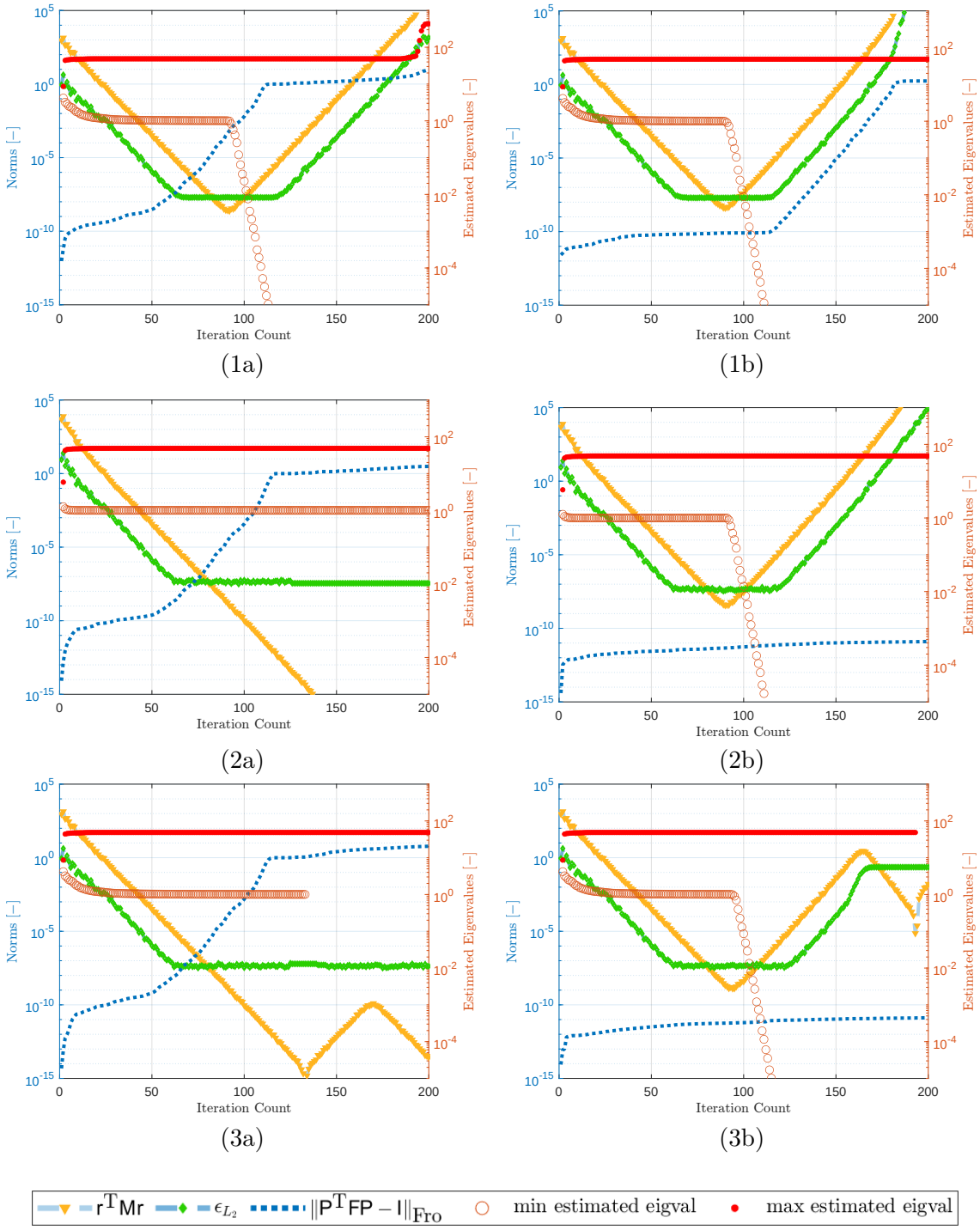


Figure 4.4: Convergence behavior of second case of iteration 10 in MTO in terms of quantifiers introduced in the list at the beginning of Section 4.2. Each row corresponds to one enforcement approach: **(1)** generalized transformation of basis, **(2)** balancing, and **(3)** projector preconditioning. Each column provide convergence plot for one variant of the conjugate gradient (CG) iterative solver: **(a)** standard preconditioned CG algorithm and **(b)** CG with employed full reorthogonalization. For all simulations, a vertex-based prior coarse space with ρ scaling with adaptive coarse space augmentation via GEVP (3.12) with $\mathsf{tol} = 50$ is adopted.

This holds until the solver reaches its capacity and the solution accuracy appears to stop approaching the reference one; the error estimator does not reflect this stagnation and continues to decrease uniformly. In both cases, eigenvalues less than one are detected in gToB, and divergence soon occurs. We do not concern ourselves with this since this approach manages to find a sufficiently accurate solution. We again note that in practice, calculations are always stopped before a relative error norm drops by almost fourteen orders of magnitude.

The orthogonality of search directions is soon severely violated in finite precision computations with the standard CG method. When we use conjugate gradients with a modified Gram-Schmidt algorithm for the full reorthogonalization of search directions, the Frobenius norm of the product $\mathsf{P}^{\mathsf{T}}\mathsf{F}\mathsf{P} - \mathsf{I}$ remains low throughout the initial iterations where the solver still converges. Hence, it is apparent that CGFO maintains the orthogonality of search directions well until around iteration 90. This is visualized in Fig. 4.51b and Fig. 4.61b, where the loss of orthogonality among the two search directions is shown by plotting $\sigma_{\mathsf{F}}(\mathsf{u}_i, \mathsf{u}_j) = \dfrac{\mathsf{u}_i^{\mathsf{T}}\mathsf{F}\mathsf{u}_j}{\sqrt{\mathsf{u}_i^{\mathsf{T}}\mathsf{F}\mathsf{u}_i} \cdot \sqrt{\mathsf{u}_j^{\mathsf{T}}\mathsf{F}\mathsf{u}_j}}$. An asymmetry in the non-assembled left-hand side matrix F is clearly noticeable; hence, the values below the diagonal are clearly lower in these plots.

The same plots are given for balancing and projector preconditioning approaches in the second and third rows in the same figures. Here, we can see that the performance of these two approaches is again comparable, and this time very poor. Although the norm of the preconditioned residual uniformly decreases, the solution does not improve throughout the iterations. This is undesirable, as the iterative solver does not warn us that its capacity has been reached. The best solution provided by these approaches is several orders of magnitude worse compared to the one obtained by the generalized transformation of basis. Interestingly, despite its inability to provide a more accurate solution, the projector preconditioning approach detects a detrimental eigenvalue lower than 1 only after 90 iterations. This suggests that the previous ninety iterations were essentially unnecessary. The balancing approach does not detect any eigenvalue lower than one and continues to produce excessively low values of the preconditioned residual. It is generally known that balancing is less prone to completely diverging [28], even though, due to the loss of orthogonality of search directions and projection matrices, it might stop providing more accurate solutions. In this case, however, the differences with the generalized transformation of basis are extreme, and therefore we always consider gToB to be the method of preferred choice.

To mitigate potential risks, we conducted several tests comparing our implementation with the results provided in the literature, e.g. synthetic tasks in [43], and we obtained satisfactory results with the same condition numbers and accurate solutions. However, for problems with (slowly) varying coefficients together with high material contrast, such as those arising in topology optimization, we were not able to enforce our adaptive constraints. Hence, we were forced to implement the generalized transformation of basis approach.

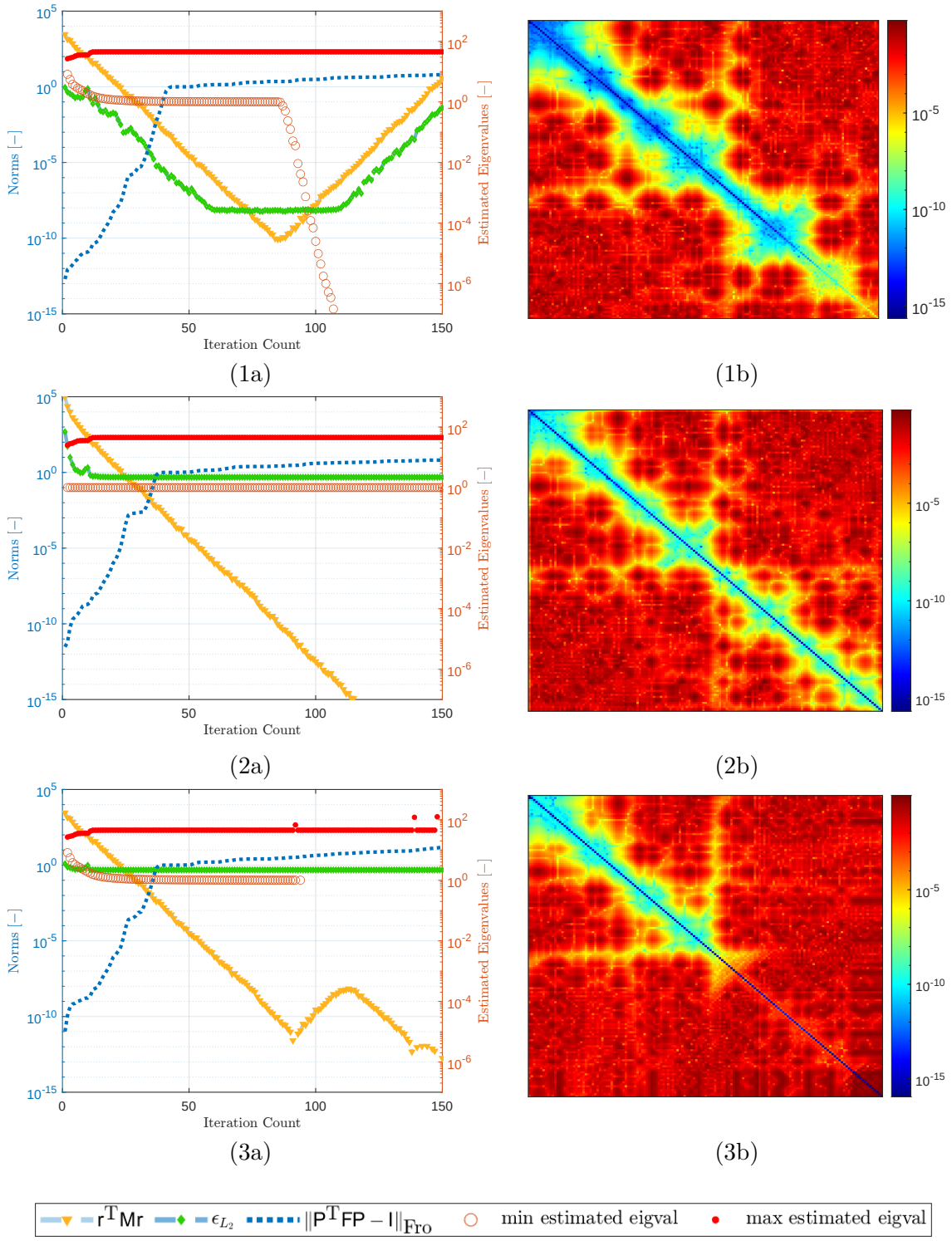


Figure 4.5: Convergence behavior of third case of iteration 30 in MTO in terms of quantifiers introduced in the list at the beginning of Section 4.2. Each row corresponds to one enforcement approach: **(1)** generalized transformation of basis, **(2)** balancing, and **(3)** projector preconditioning. First column provides convergence plots. Second column illustrates the loss of orthogonality among search directions: products $\sigma_{\mathsf{F}}(\mathsf{u}_i, \mathsf{u}_j)$ are plotted. Results for standard preconditioned CG algorithm are shown. For all simulations, a vertex-based prior coarse space with ρ scaling with adaptive coarse space augmentation via GEVP (3.12) with $\mathsf{tol} = 50$ is adopted.

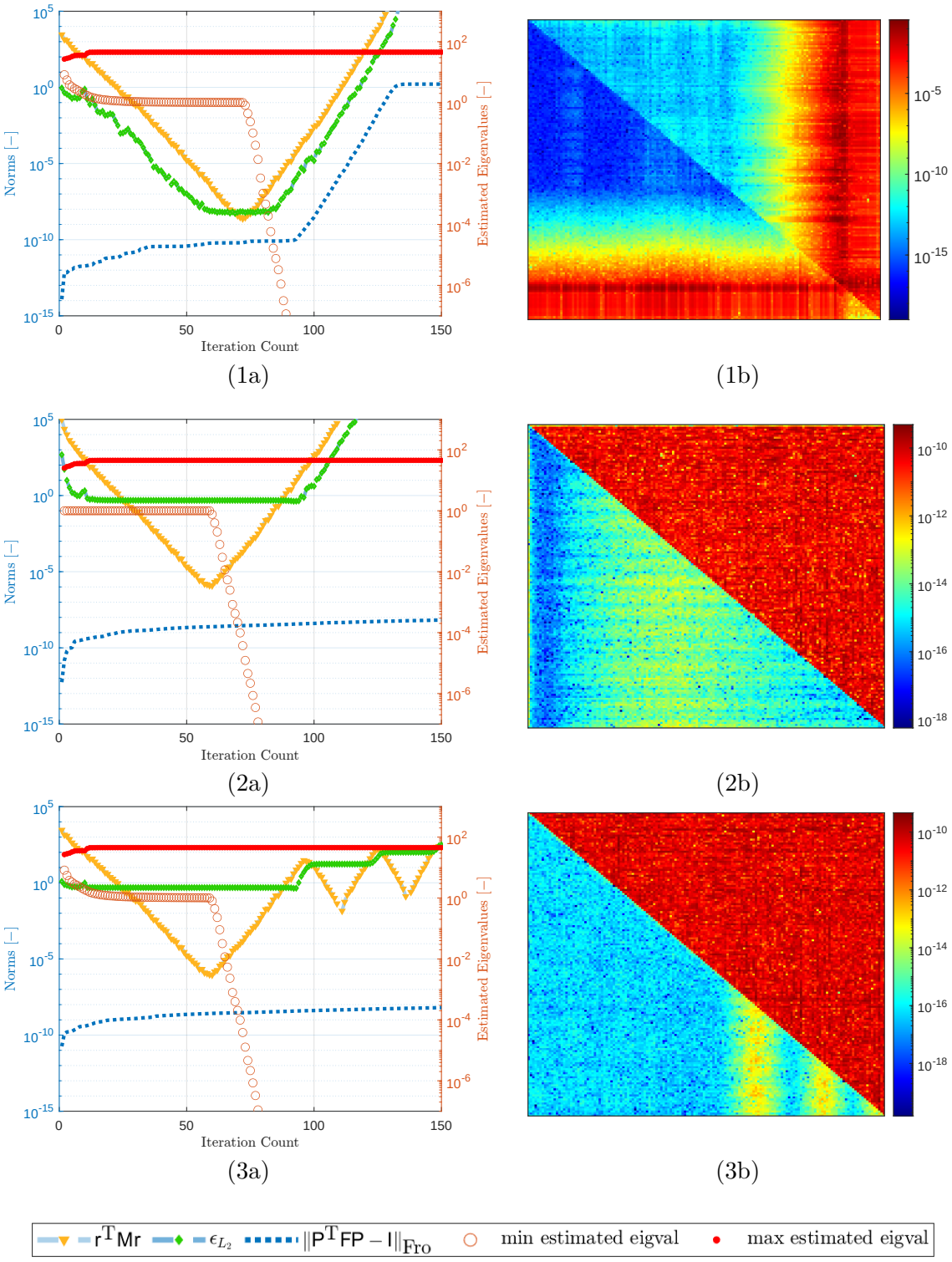


Figure 4.6: Convergence behavior of third case of iteration 30 in MTO in terms of quantifiers introduced in the list at the beginning of Section 4.2. Each row corresponds to one enforcement approach: **(1)** generalized transformation of basis, **(2)** balancing, and **(3)** projector preconditioning. First column provides convergence plots. Second column illustrates the loss of orthogonality among search directions: products $\sigma_{\mathsf{F}}(\mathsf{u}_i, \mathsf{u}_j)$ are plotted. Results for standard preconditioned CG algorithm are shown. For all simulations, a vertex-based prior coarse space with ρ scaling with adaptive coarse space augmentation via GEVP (3.12) with $\mathsf{tol} = 50$ is adopted.

4.3 Comparison of coarse space enhancements

In this section, we showcase the practical usability of our adjusted adaptive and heuristic strategies by conducting a series of numerical tests. Our goal is to demonstrate the potential efficacy and robustness of the novel strategies in adressing highly heterogeneous problems within modular topology optimization framework.

In this comparison, we include heuristic approaches in the form of arithmetic and weighted averages (denoted as "avg weighted classic") and their novel formulation (denoted as "avg weighted min max"). We also include the two variants of Frugal approach: with and without the selective criterion for discarding constraints. Next, we include the proposed heuristic with extra nodal constraints, and adaptive constraints obtained by solving (1) the original formulation of eigenproblem (3.10) (denoted as "MS GEVP full") and (2) the reduced-basis strategy (denoted as "MS GEVP red. basis" in the legends) as introduced in Subsection 3.4.

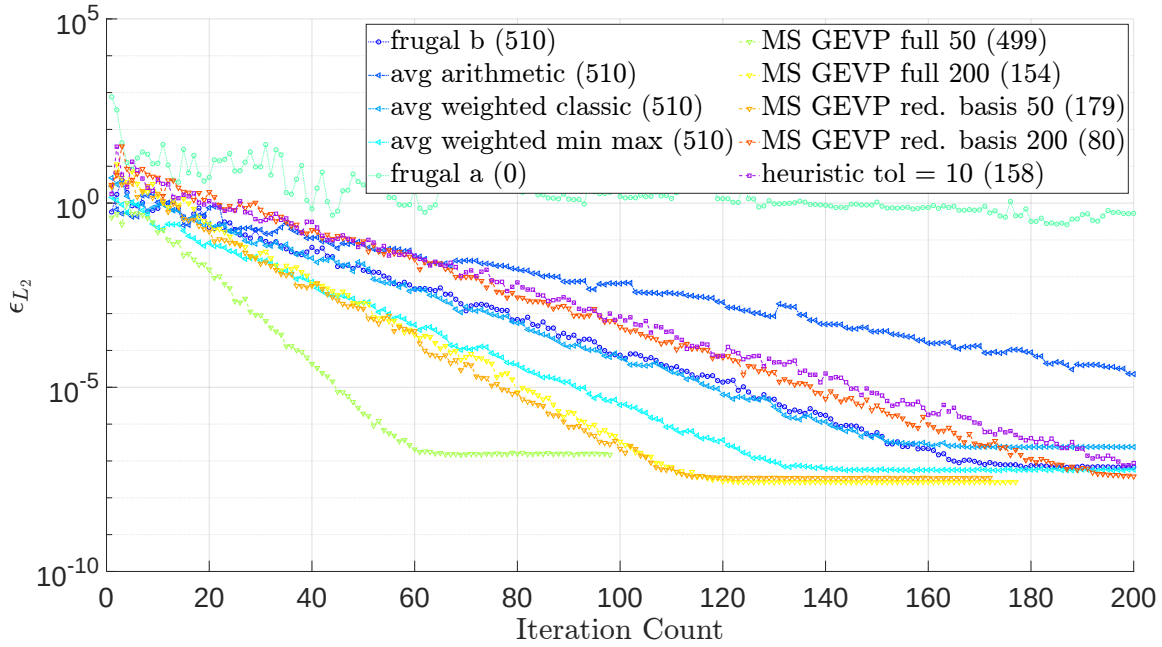


Figure 4.7: Comparison of different coarse space enhancements in terms of ϵ_{L_2} reduction with progressive iterations. Results are shown for the test case of 10^{th} iteration in MTO (recall Fig. 4.2) on a mesh with 30×30 bilinear elements per module. Numbers in parentheses in the legend state the number of constraints imposed for each method. In this case, ρ scaling is used and vertex-based prior coarse space contains 230 primal DOFs.

Fig. 4.7 presents a comprehensive set of convergence plots illustrating the performance of various coarse space enhancement strategies in tenth iteration of the topology optimization process. We consider this run as an economic one, thus we employ a relatively cheap ρ scaling in this example. To provide a fair comparison, ten different set-ups are compared in terms of ϵ_{L_2}.

Starting with the simplest approach available, arithmetic averages (dashed sky blue line) were competitive during the first approximately 60 iterations compared to our novel heuristic strategy. However, their convergence soon slowed and they were the only approach that failed to reach ϵ_{L_2} norm 10^{-5} within the first 200 iterations.

The reader may also note the poor performance of **frugal a**. We do not consider variant **frugal a** as a separate type of coarse space enhancement, but only as an adjustment of the frugal approach to potentially exclude some unnecessary constraints, particularly those for which the eigenvalue estimator μ does not exceed the given tolerance; recall (3.17) in Subsection 3.3.1. Despite the fact that we used a very low threshold $\mathsf{tol} = 5$ for this adaptive selection, **frugal a** discarded all of the available constraints. This is not surprising,

as with smoothly spatially-varying coefficients, we often prescribe very high gradients on the high-coefficient elements, thus penalizing the low-energetical side of the eigenproblem (3.12) in the expression for μ. Therefore, the azure blue line, representing **frugal a**, shows the worst convergence rate provided in the figure, despite the fact that frugal with all constraints considered (dashed royal blue line), provides significant improvement in convergence behavior. The Frugal approach itself, without an adaptive selection of constraints, showed numerical performance similar to that of the classic weighted averages (dashed turquoise line). Given that the setup of the frugal approach is still costly compared to the weighted averages, this does not reflect well on the Frugal approach, as both strategies use the same number of all three constraints on each interface $\Gamma^{(ij)}$. In contrast, our enhanced strategy for constructing weighted averages, otherwise equally computationally expensive as the original one, yielded significantly better results (light cyan dashed line), attesting to the merits of our modification.

Unsurprisingly, the best results were achieved by the adaptive approach based on the localized GEVPs. The run with a threshold equal to 50 outperformed all others, though at the cost of 499 additional constraints. Very satisfactory results were also obtained with the solution of the original GEVP with a higher threshold of 200 (yellow line), resulting in only 154 constraints. Note that the practically identical results were achieved with our reduced basis (RB-GEVP) approach (in orange), but a lower threshold of 50 was needed. We wish to emphasize that the maximum dimension of the GEVP was only 12 in this case. More specifically, 143 out of 170 possible GEVPs were computed in the reduced-basis approach, with the average dimension of the solved GEVPs only 7.85, which is almost two orders of magnitude fewer compared to the average dimension of 437.6 in the case of the full basis in the original setting. Even more significant savings in the dimensionality of the second coarse space were obtained by our reduced-dimensional approach with higher $\mathsf{tol} = 200$. Now, only 80 constraints were identified, yet the approach still seems to be competitive with heuristics that have a posteriori coarse problem six times larger. Nevertheless, our heuristic nodal approach (in purple) for enriching the set of primal constraints Π would probably be the preferred choice. Despite leading to twice as many constraints as RB-GEVP, the constraints count (158) is still three times lower than in standard heuristics such as (weighted) averages. In spite of that, this heuristic demonstrates promising performance. Note that these heuristically recognized constraints are obtained instantly and do not require any extra complexity in the solver, such as transformation of basis or projections. Hence, it might be a better choice even when compared to seemingly better-performing strategies such as classic weighted averages.

Generally speaking, the differences in the number of iterations between the various approaches using ρ scaling are quite significant. The situation changes with deluxe scaling; see Fig. 4.8. Now, the only two approaches that converge noticeably worse are the arithmetic averages and the frugal approach with the selection criterion turned on, which again discarded all the otherwise beneficial constraints. Comparing the remaining methods is difficult because of their similar performance. However, there are two observations to be made: First, our heuristic approach with addition of nodes was the third-best among all the methods in this comparison, although it added only 158 constraints (that means, 79 nodes) compared to 510 for the other heuristic approaches. Second, we can clearly see that the standard heuristic approaches lead to unnecessary large coarse space augmentation with deluxe scaling. It is evident that the deluxe scaling alone can eliminate a significant part of the ill-conditioned modes present in other scalings, and it is therefore sufficient to add only a smaller number of additional constraints, which do not need to be obtained adaptively. However, the deluxe scaling itself does not guarantee convergence of the solver within a reasonably low number of iterations, and some coarse space enhancement have to be incorporated.

To show that the previous results are not tied to a specific test problem, we provide another set of results, this time for the 30^{th} iteration of the modular topology optimization scheme. In this case, we report results for a modular problem with a finer mesh with 50×50 elements per module. Hence, the a priori coarse problem is still assembled in 230 DOFs, but now we iterate on 16,478 lambdas. Despite the used ρ scaling, convergence rates are comparable

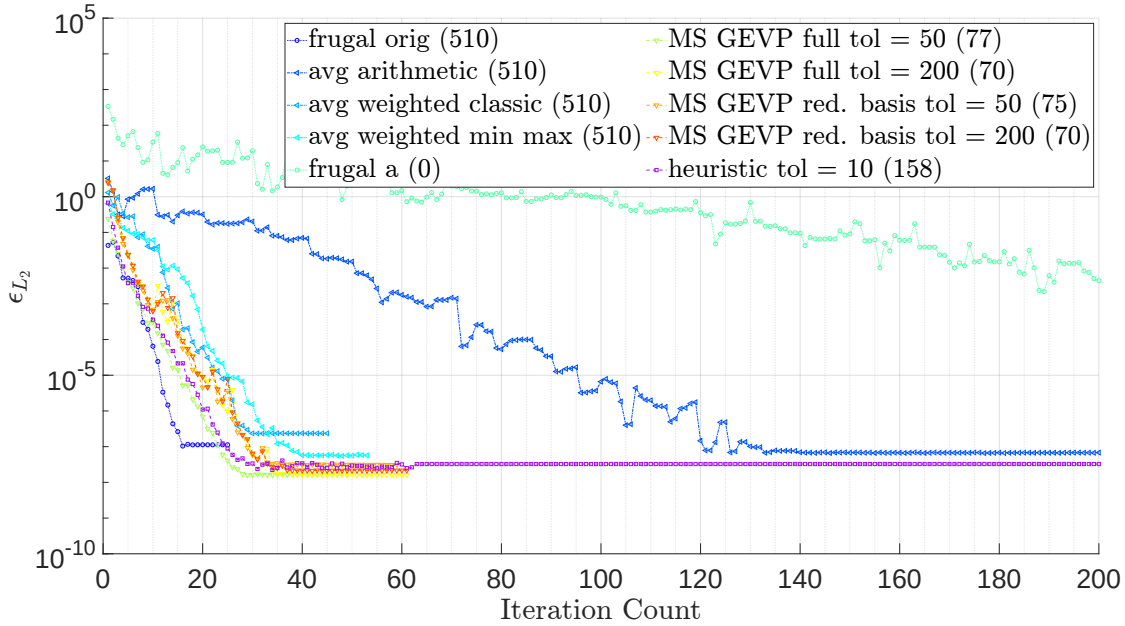


Figure 4.8: Comparison of different coarse space enhancements in terms of ϵ_{L_2} reduction with progressive iterations. Results are shown for the test case of 10^{th} iteration in MTO (recall Fig. 4.2) on a mesh with 30×30 bilinear elements per module. Numbers in parentheses in the legend state the number of constraints imposed for each method. In this case, deluxe scaling is used and vertex-based prior coarse space contains 230 primal DOFs.

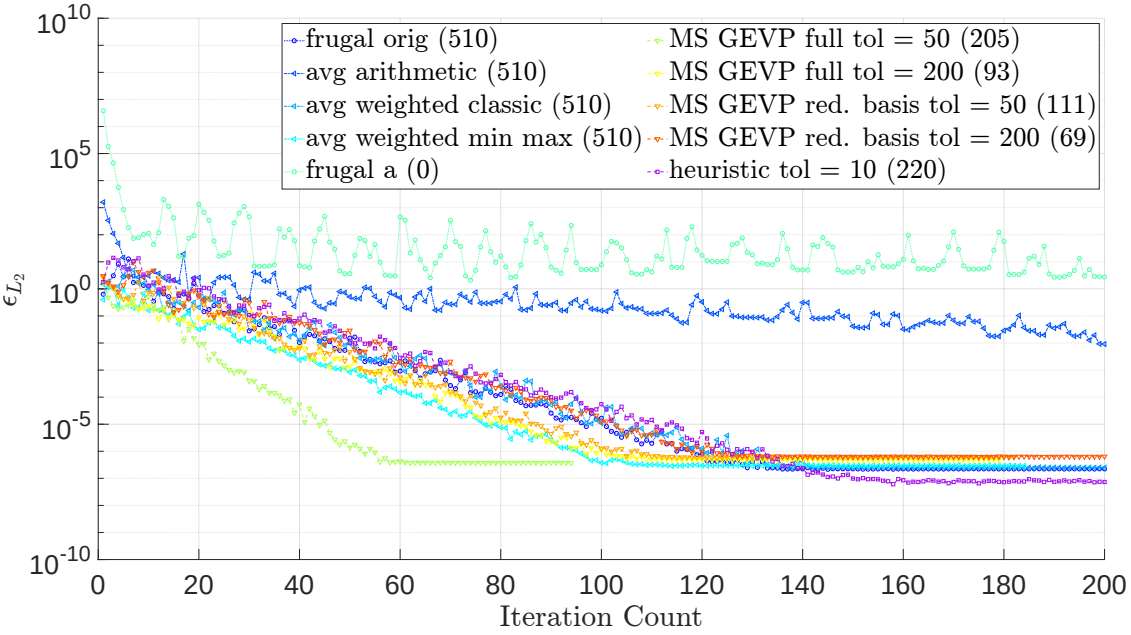


Figure 4.9: Comparison of different coarse space enhancements in terms of ϵ_{L_2} reduction with progressive iterations. Results are shown for the test case of 30^{th} iteration in MTO (recall Fig. 4.2) on a mesh with 50×50 bilinear elements per module. Numbers in parentheses in the legend state the number of constraints imposed for each method. In this case, deluxe scaling is used and vertex-based prior coarse space contains 230 primal DOFs.

among several heuristic approaches; similarly to the previous test problem in which, however, deluxe scaling was used. Interestingly, the proposed modification to the weighted averages outperformed other heuristics. Classic weighted averages and the Frugal approach, both with 510 constraints, exhibited similar performance to our cheap heuristic with only 220 constraints and to RB-GEVP with only 69 constraints. This represents a reduction in the dimensionality of the second coarse space by more than a factor of seven. The original GEVP with 205

constraints for tol of 50 and with 93 constraints for tol of 200 can be used as a baseline for comparison. RB-GEVP with 111 constraints shows similar performance to GEVP with 93 constraints, although the threshold needs to be lowered to 50.

Basis	**Solved GEVPs**				**Dimensions of GEVP**		
	Computed	Correct	Unnecessary	Neglected	min	max	avg
Reduced	85	66	19	0	6	18	9,07
Full	170	66	104	-	504	800	728,4

Table 4.2: The table describes the number of solved GEVPs for two variants (with reduced or full basis) and their dimensions. The terms mean: **Correct**: GEVPs were correctly solved, where the original formulation found at least one eigenvalue exceeding tol = 50. **Unnecessary**: This GEVP was unnecessarily solved, as the original formulation did not find any eigenvalue. **Neglected**: Interfaces where the original formulation found an eigenvalue, but the reduced-dimensional formulation neglected this interface.

Moreover, in this case, the savings in the reduction of GEVP dimensionality are maximized, as seen in the Tab. 4.2. This table provides a comparison of the two approaches for setting adaptive constraints: the original formulation of GEVP and RB-GEVP.
The table shows that the RB-GEVP approach successfully eliminated 85 GEVPs that were computed in the original formulation but were unnecessary. Specifically, the RB-GEVP strategy computed a total of 85 GEVPs. Out of these, 66 contained eigenvalues that exceeded the given tolerance tol, while 19 were identified as unnecessary because they did not contain any eigenvalue exceeding tol. Importantly, no pair of subdomains containing ill-posed modes was neglected from the computation. In contrast, the original formulation resulted in the computation of 170 GEVPs, of which 104 did not contain any eigenvalues higher than tol.
The dimensions of the GEVPs also highlight the efficacy of the reduced basis approach. For the RB-GEVP strategy, the GEVP dimensions ranged from 6 to 18 with an average of 9, whereas for the full basis, the dimensions ranged from 50 to 800 with an average of approximately 730, dimension almost two orders of magnitude higher.

These results demonstrate a promising potential of the adaptive constraints obtained by the RB-GEVP approach, which effectively reduces the dimensionality of localized GEVPs while still managing to accurately approximate the most detrimental modes. However, the novel heuristic also proves to be an effective engineering tool for efficiently solving highly heterogeneous problems. Furthermore, the novel formulation of weighted averages has shown to be highly beneficial, as it in certain cases significantly accelerates convergence.

Chapter 5

Conclusions

Motivated by linear systems arising in modular-topology optimization, this thesis focused on enhancing robustness and convergence of the Finite Element Tearing and Interconnecting Dual Primal (FETI-DP) method when applied to problems exhibiting a heterogeneous distribution of high-contrast coefficients. In particular, we implemented and tested commonly-used coarse-space enrichment strategies, ranging from the weighted averages to adaptive approaches based on generalized eigenvalue problems.

One observation we made while testing different scaling options is that the most complex deluxe scaling is particularly powerful when the severely varying coefficients appear inside individual subdomains. However, it represents a valid choice for all kinds of tasks with high coefficients, yet the related gains compared to, e.g., stiffness scaling are less pronounced than one might expect.

While not initially anticipated, we dedicated significant effort to the methods of enforcing the identified enrichments with the iterative procedure pertinent to FETI-DP, because the classical projection-based method did not perform well in the complex problems arising in modular-topology optimization. In particular, the projection-based method's performance was treacherous: it is highly sensitive to inaccuracies in calculations, possibly leading to a loss of orthogonality among search directions. This becomes particularly problematic when the iterative solver encounters small eigenvalues close to zero. In such cases, the solver might fail to converge to the accurate solution without any warning, as the monitored error norm may continue decreasing. Therefore, it was essential to adopt the generalized transformation of basis, which provided superior accuracy and stability compared to projection-based enforcement approaches, in order to compare individual enrichment strategies.

The study, implementation, and testing of existing enrichment approaches were crucial in understanding what the critical modes are, for which the coarse-space must be enhanced. Although not fully applicable out-of-the-box, we showed that the widely used heuristic approaches can be further improved with minor modifications. Specifically, we proposed a slight adjustment of the classic weighted averages approach that improves performance of the solver in additional problems, see the numerical comparison in Section 4.3. Surprisingly, the frugal approach overall did not generally provide better performance than the weighted averages, particularly the modified weighted averages. Consequently, we proposed another minor modification inspired by ρ scaling, which is capable of identifying most of the edges which should be enriched with an adaptive approach.

Motivated by these observations, we proposed a novel reduced-dimensional strategy for computing adaptive constraints. The low-dimensional basis can be constructed with information on coefficient distribution limited to the interfaces, which is the same amount of data that is needed for the widely-used ρ scaling. We demonstrated that the performance of this strategy is almost comparable to the adaptive constraints generated by the unreduced generalized eigenvalue problem, while reducing the size of the system by more than one order of magnitude.

Finally, we returned to the simple heuristic proposed in author's bachelor thesis and improved it to reflect the observed behavior of adaptive strategies. Analyzing the coefficient

profile along an interface, similarly to the reduced-dimensional strategy, we identify high-coefficient clusters and choose a characteristic degree of freedom in those clusters that are then incorporated into the primal coarse space of FETI-DP.

Admittedly, some heuristic strategies may lead to an unnecessarily large number of constraints, but, importantly, they still ensure (albeit possibly slow) relatively reliable convergence to an accurate solution. We emphasize here that without coarse space enhancements, the solver might not be able to reach a satisfactorily accurate solution at all. Even adaptive techniques, due to their locality, in some cases tend to produce an excessive number of constraints, among which some provide only redundant information. This leaves an open door for further research.

Bibliography

[1] Badia, S., Martín, A. F., and Principe, J. (2016). Multilevel Balancing Domain Decomposition at Extreme Scales. *SIAM Journal on Scientific Computing*, 38(1):C22–C52, DOI: 10.1137/15M1013511.

[2] Bakalakos, S., Georgioudakis, M., and Papadrakakis, M. (2022). Domain decomposition methods for 3D crack propagation problems using XFEM. *Computer Methods in Applied Mechanics and Engineering*, 402:115390, DOI: 10.1016/j.cma.2022.115390.

[3] Balay, S., Abhyankar, S., Adams, M., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Dener, A., Eijkhout, V., Gropp, W., Karpeyev, D., Kaushik, D., Knepley, M., May, D., McInnes, L., Mills, R., Munson, T., Rupp, K., Sanan, P., Smith, B., Zampini, S., Zhang, H., and Zhang, H. (2019). PETSc Users Manual (Rev. 3.11). Technical Report ANL–95/11 Rev 3.11, 1577437, 155920, Argonne National Laboratory, DOI: 10.2172/1577437.

[4] Bendsøe, M. P. and Sigmund, O. (2004). *Topology Optimization*. Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN: 978-3-642-07698-5 978-3-662-05086-6, DOI: 10.1007/978-3-662-05086-6.

[5] Bovet, C., Parret-Fréaud, A., Spillane, N., and Gosselet, P. (2017). Adaptive multipreconditioned FETI: Scalability results and robustness assessment. *Computers & Structures*, 193:1–20, DOI: 10.1016/j.compstruc.2017.07.010.

[6] Brenner, S. C. and Sung, L.-Y. (2007). BDDC and FETI-DP without matrices or vectors. *Computer Methods in Applied Mechanics and Engineering*, 196(8):1429–1435, DOI: 10.1016/j.cma.2006.03.012.

[7] Dohrmann, C. R. (2003). A Preconditioner for Substructuring Based on Constrained Energy Minimization. *SIAM Journal on Scientific Computing*, 25(1):246–258, DOI: 10.1137/S1064827502412887.

[8] Dohrmann, C. R. and Widlund, O. B. (2013). Some Recent Tools and a BDDC Algorithm for 3D Problems in H(curl). In Bank, R., Holst, M., Widlund, O., and Xu, J., editors, *Domain Decomposition Methods in Science and Engineering XX*, volume 91, pages 15–25. Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN: 978-3-642-35274-4 978-3-642-35275-1, DOI: 10.1007/978-3-642-35275-1_2. Series Title: Lecture Notes in Computational Science and Engineering.

[9] Dolean, V., Jolivet, P., and Nataf, F. (2015). *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, ISBN: 978-1-61197-405-8 978-1-61197-406-5, DOI: 10.1137/1.9781611974065.

[10] Dostál, Z., Horák, D., and Kučera, R. (2006). Total FETI-an easier implementable variant of the FETI method for numerical solution of elliptic PDE. *Communications in Numerical Methods in Engineering*, 22(12):1155–1162, DOI: 10.1002/cnm.881.

[11] Farhat, C., Lesoinne, M., and Pierson, K. (2000a). A scalable dual-primal domain decomposition method. *Numerical Linear Algebra with Applications*, 7(7-8):687–714, DOI: 10.1002/1099-1506(200010/12)7:7/8<687::AID-NLA219>3.0.CO;2-S.

[12] Farhat, C., Lesoinne, M., and Pierson, K. (2000b). A scalable dual-primal domain decomposition method. *Numerical linear algebra with applications*, 7:687–714, DOI: 10.1002/1099-1506(200010/12)7:7/83.0.CO;2-S.

[13] Farhat, C., Mandel, J., and Roux, F. X. (1994). Optimal convergence properties of the FETI domain decomposition method. *Computer Methods in Applied Mechanics and Engineering*, 115(3-4):365–385, DOI: 10.1016/0045-7825(94)90068-X.

[14] Farhat, C. and Roux, F.-X. (1991). A method of finite element tearing and interconnecting and its parallel solution algorithm. *International Journal for Numerical Methods in Engineering*, 32(6):1205–1227, DOI: 10.1002/nme.1620320604.

[15] Gosselet, P. and Rey, C. (2006). Non-overlapping domain decomposition methods in structural mechanics. *Archives of Computational Methods in Engineering*, 13(4):515–572, DOI: 10.1007/BF02905857.

[16] Gosselet, P., Rixen, D., Roux, F.-X., and Spillane, N. (2015). Simultaneous FETI and block FETI: Robust domain decomposition with multiple search directions. *International Journal for Numerical Methods in Engineering*, 104(10):905–927, DOI: 10.1002/nme.4946.

[17] Heinlein, A., Klawonn, A., and Kühn, M. J. (2020a). Local Spectra of Adaptive Domain Decomposition Methods. In Haynes, R., MacLachlan, S., Cai, X.-C., Halpern, L., Kim, H. H., Klawonn, A., and Widlund, O., editors, *Domain Decomposition Methods in Science and Engineering XXV*, volume 138, pages 167–175. Springer International Publishing, Cham, ISBN: 978-3-030-56749-1 978-3-030-56750-7, DOI: 10.1007/978-3-030-56750-7_18. Series Title: Lecture Notes in Computational Science and Engineering.

[18] Heinlein, A., Klawonn, A., Lanser, M., and Weber, J. (2020b). A frugal FETI-DP and BDDC coarse space for heterogeneous problems. *ETNA - Electronic Transactions on Numerical Analysis*, 53:562–591, DOI: 10.1553/etna_vol53s562.

[19] Klawonn, A., Kühn, M., and Rheinbach, O. (2016). Adaptive Coarse Spaces for FETI-DP in Three Dimensions. *SIAM Journal on Scientific Computing*, 38(5):A2880–A2911, DOI: 10.1137/15M1049610.

[20] Klawonn, A., Kühn, M., and Rheinbach, O. (2020a). Coarse spaces for FETI-DP and BDDC Methods for heterogeneous problems: connections of deflation and a generalized transformation-of-basis approach. *ETNA - Electronic Transactions on Numerical Analysis*, 52:43–76, DOI: 10.1553/etna_vol52s43.

[21] Klawonn, A., Kühn, M. J., and Rheinbach, O. (2020b). A Closer Look at Local Eigenvalue Solvers for Adaptive FETI-DP and BDDC. In Haynes, R., MacLachlan, S., Cai, X.-C., Halpern, L., Kim, H. H., Klawonn, A., and Widlund, O., editors, *Domain Decomposition Methods in Science and Engineering XXV*, volume 138, pages 235–242. Springer International Publishing, Cham, ISBN: 978-3-030-56749-1 978-3-030-56750-7, DOI: 10.1007/978-3-030-56750-7_26. Series Title: Lecture Notes in Computational Science and Engineering.

[22] Klawonn, A., Kühn, M. J., and Rheinbach, O. (2020c). Parallel adaptive FETI-DP using lightweight asynchronous dynamic load balancing. *International Journal for Numerical Methods in Engineering*, 121(4):621–643, DOI: 10.1002/nme.6237.

[23] Klawonn, A., Lanser, M., and Weber, J. (2022). Adaptive Three-level BDDC Using Frugal Constraints. Technical Report, Universität zu Köln. Volume: 2022-04.

[24] Klawonn, A., Radtke, P., and Rheinbach, O. (2015). FETI-DP Methods with an Adaptive Coarse Space. *SIAM Journal on Numerical Analysis*, 53(1):297–320, DOI: 10.1137/130939675.

[25] Klawonn, A. and Rheinbach, O. (2006). A Parallel Implementation of Dual-Primal FETI Methods for Three-Dimensional Linear Elasticity Using a Transformation of Basis. *SIAM Journal on Scientific Computing*, DOI: 10.1137/050624364. Publisher: Society for Industrial and Applied Mathematics.

[26] Klawonn, A. and Rheinbach, O. (2007a). Inexact FETI-DP methods. *International Journal for Numerical Methods in Engineering*, 69(2):284–307, DOI: 10.1002/nme.1758.

[27] Klawonn, A. and Rheinbach, O. (2007b). Robust FETI-DP methods for heterogeneous three dimensional elasticity problems. *Computer Methods in Applied Mechanics and Engineering*, 196:1400–1414, DOI: 10.1016/j.cma.2006.03.023.

[28] Klawonn, A. and Rheinbach, O. (2012). Deflation, Projector Preconditioning, and Balancing in Iterative Substructuring Methods: Connections and New Results. *SIAM Journal on Scientific Computing*, 34(1):A459–A484, DOI: 10.1137/100811118.

[29] Klawonn, A. and Widlund, O. B. (2006). Dual-primal FETI methods for linear elasticity. *Communications on Pure and Applied Mathematics*, 59(11):1523–1572, DOI: 10.1002/cpa.20156.

[30] Kozubek, T., Vondrák, V., Menšik, M., Horák, D., Dostál, Z., Hapla, V., Kabeliková, P., and Čermák, M. (2013). Total FETI domain decomposition method and its massively parallel implementation. *Advances in Engineering Software*, 60-61:14–22, DOI: 10.1016/j.advengsoft.2013.04.001.

[31] Kühn, M. J. (2018). *Adaptive FETI-DP and BDDC methods for highly heterogeneous elliptic finite element problems in three dimensions*. PhD thesis, University of Cologne.

[32] Li, J. and Widlund, O. B. (2006). FETI-DP, BDDC, and block Cholesky methods. *International Journal for Numerical Methods in Engineering*, 66(2):250–271, DOI: 10.1002/nme.1553.

[33] Mandel, J. and Dohrmann, C. R. (2003). Convergence of a balancing domain decomposition by constraints and energy minimization. *Numerical Linear Algebra with Applications*, 10(7):639–659, DOI: 10.1002/nla.341.

[34] Mandel, J., Dohrmann, C. R., and Tezaur, R. (2005). An algebraic theory for primal and dual substructuring methods by constraints. *Applied Numerical Mathematics*, 54(2):167–193, DOI: 10.1016/j.apnum.2004.09.022.

[35] Mandel, J. and Sousedík, B. (2007). Adaptive selection of face coarse degrees of freedom in the BDDC and the FETI-DP iterative substructuring methods. *Computer Methods in Applied Mechanics and Engineering*, 196(8):1389–1399, DOI: 10.1016/j.cma.2006.03.010.

[36] Medřický, T. (2022). Comparison and adaptation of dual domain decomposition methods for modular topology optimization problems. Technical report, ČVUT v Praze.

[37] Medřický, T., Doškář, M., Pultarová, I., and Zeman, J. (2022). Comparison of FETI-based domain decomposition methods for topology optimization problems. *Acta Polytechnica CTU Proceedings*, 34, DOI: 10.14311/APP.2022.34.0043.

[38] Molina, R. and Roux, F. (2019). New implementations for the Simultaneous-FETI method. *International Journal for Numerical Methods in Engineering*, 118(9):519–535, DOI: 10.1002/nme.6024.

[39] Nabben, R. and Vuik, C. (2008). A comparison of abstract versions of deflation, balancing and additive coarse grid correction preconditioners. *Numerical Linear Algebra with Applications*, 15(4):355–372, DOI: 10.1002/nla.571.

[40] Pavarino, L. F., Scacchi, S., Widlund, O. B., and Zampini, S. (2018). Isogeometric BDDC deluxe preconditioners for linear elasticity. *Mathematical Models and Methods in Applied Sciences*, 28(07):1337–1370, DOI: 10.1142/S0218202518500367.

[41] Pechstein, C. and Dohrmann, C. R. (2017). A Unified Framework for Adaptive BDDC. *Electronic Transactions on Numerical Analysis*, 46:64.

[42] Rheinbach, O. and Kühn, M. J. (2017). FETI-DP and BDDC Methods with a Generalized Transformation of Basis For Heterogeneous Problems: Connections to Deflation. Technical report, TU Bergakademie Freiberg.

[43] Rheinbach, O., Radtke, P., and Klawonn, A. (2016). A Comparison of Adaptive Coarse Spaces for Iterative Substructuring Methods in Two Dimensions. *Electronic Transactions on Numerical Analysis*, 45:75–106.

[44] Rixen, D. J. and Farhat, C. (1999). A simple and efficient extension of a class of substructure based preconditioners to heterogeneous structural mechanics problems. *International Journal for Numerical Methods in Engineering*, 44(4):489–516, DOI: 10.1002/(SICI)1097-0207(19990210)44:4<489::AID-NME514>3.0.CO;2-Z.

[45] Rozložník, M. (2011). Orthogonalization with a non-standard inner product and approximate inverse preconditioning - conference abstract.

[46] Saad, Y. (2003). *Iterative methods for sparse linear systems*. SIAM, Philadelphia, 2nd ed edition, ISBN: 978-0-89871-534-7.

[47] Saad, Y. (2011). *Numerical Methods for Large Eigenvalue Problems: Revised Edition*. Society for Industrial and Applied Mathematics, ISBN: 978-1-61197-072-2 978-1-61197-073-9, DOI: 10.1137/1.9781611970739.

[48] Saad, Y., Yeung, M., Erhel, J., and Guyomarc'h, F. (2000). A Deflated Version of the Conjugate Gradient Algorithm. *SIAM Journal on Scientific Computing*, 21(5):1909–1926, DOI: 10.1137/S1064829598339761.

[49] Shao, M. (2023). Householder Orthogonalization with a Nonstandard Inner Product. *SIAM Journal on Matrix Analysis and Applications*, 44(2):481–502, DOI: 10.1137/21M1414814.

[50] Sousedík, B. (2022). Inexact and primal multilevel FETI-DP methods: a multilevel extension and interplay with BDDC. *International Journal for Numerical Methods in Engineering*, 123(20):4844–4858, DOI: 10.1002/nme.7057.

[51] Spillane, N., Dolean, V., Hauret, P., Nataf, F., and Rixen, D. J. (2013). Solving generalized eigenvalue problems on the interfaces to build a robust two-level FETI method. *Comptes Rendus Mathematique*, 351(5-6):197–201, DOI: 10.1016/j.crma.2013.03.010.

[52] Spillane, N. and Rixen, D. (2013). Automatic spectral coarse spaces for robust finite element tearing and interconnecting and balanced domain decomposition algorithms. *International Journal for Numerical Methods in Engineering*, 95(11):953–990, DOI: 10.1002/nme.4534.

[53] Stewart, G. W. (2002). A Krylov–Schur Algorithm for Large Eigenproblems. *SIAM Journal on Matrix Analysis and Applications*, 23(3):601–614, DOI: 10.1137/S0895479800371529.

[54] Toivanen, J., Avery, P., and Farhat, C. (2018). A multilevel FETI-DP method and its performance for problems with billions of degrees of freedom. *International Journal for Numerical Methods in Engineering*, 116(10-11):661–682, DOI: 10.1002/nme.5938.

[55] Toselli, A. and Widlund, O. B. (2005). *Domain Decomposition Methods — Algorithms and Theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, DOI: 10.1007/b137868.

[56] Tyburec, M., Zeman, J., Doškář, M., Kružík, M., and Lepš, M. (2021). Modular-topology optimization with Wang tilings: an application to truss structures. *Structural and Multidisciplinary Optimization*, 63(3):1099–1117, DOI: 10.1007/s00158-020-02744-8.

[57] Weber, J. (2022). *Efficient and robust FETI-DP and BDDC methods – Approximate coarse spaces and deep learning-based adaptive coarse spaces*. PhD Thesis, Universität zu Köln.

[58] Widlund, O. B., Zampini, S., Scacchi, S., and Pavarino, L. F. (2021). Block FETI–DP/BDDC preconditioners for mixed isogeometric discretizations of three-dimensional almost incompressible elasticity. *Mathematics of Computation*, 90(330):1773–1797, DOI: 10.1090/mcom/3614.